Advances in Intelligent Systems and Computing 557

Ana Maria Madureira
Ajith Abraham
Dorabela Gamboa
Paulo Novais *Editors*

Intelligent Systems Design and Applications

16th International Conference on Intelligent Systems Design and Applications (ISDA 2016) held in Porto, Portugal, December 16–18, 2016

Springer

Advances in Intelligent Systems and Computing

Volume 557

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Advisory Board

Chairman

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India
e-mail: nikhil@isical.ac.in

Members

Rafael Bello Perez, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba
e-mail: rbellop@uclv.edu.cu

Emilio S. Corchado, University of Salamanca, Salamanca, Spain
e-mail: escorchado@usal.es

Hani Hagras, University of Essex, Colchester, UK
e-mail: hani@essex.ac.uk

László T. Kóczy, Széchenyi István University, Győr, Hungary
e-mail: koczy@sze.hu

Vladik Kreinovich, University of Texas at El Paso, El Paso, USA
e-mail: vladik@utep.edu

Chin-Teng Lin, National Chiao Tung University, Hsinchu, Taiwan
e-mail: ctlin@mail.nctu.edu.tw

Jie Lu, University of Technology, Sydney, Australia
e-mail: Jie.Lu@uts.edu.au

Patricia Melin, Tijuana Institute of Technology, Tijuana, Mexico
e-mail: epmelin@hafsamx.org

Nadia Nedjah, State University of Rio de Janeiro, Rio de Janeiro, Brazil
e-mail: nadia@eng.uerj.br

Ngoc Thanh Nguyen, Wroclaw University of Technology, Wroclaw, Poland
e-mail: Ngoc-Thanh.Nguyen@pwr.edu.pl

Jun Wang, The Chinese University of Hong Kong, Shatin, Hong Kong
e-mail: jwang@mae.cuhk.edu.hk

More information about this series at http://www.springer.com/series/11156

Ana Maria Madureira · Ajith Abraham
Dorabela Gamboa · Paulo Novais
Editors

Intelligent Systems Design and Applications

16th International Conference on Intelligent Systems Design and Applications (ISDA 2016) held in Porto, Portugal, December 16–18, 2016

Springer

Editors
Ana Maria Madureira
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Port
Porto
Portugal

Ajith Abraham
Scientific Network for Innovation
and Research Excellence
Machine Intelligence Research Labs
Auburn, WA
USA

Dorabela Gamboa
Polytechnic Institute of Porto
Felgueiras
Portugal

Paulo Novais
Campus of Gualtar
University of Minho
Braga
Portugal

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-53479-4 ISBN 978-3-319-53480-0 (eBook)
DOI 10.1007/978-3-319-53480-0

Library of Congress Control Number: 2017930789

© Springer International Publishing AG 2017
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Welcome to the 16th International Conference on Intelligent Systems Design and Applications (ISDA16), which is held in Porto, Portugal during December 16–18, 2016. ISDA 2016 is jointly organized by the Instituto Superior de Engenharia do Porto, Portugal and Machine Intelligence Research Labs (MIR Labs), USA.

ISDA 2016 brings together researchers, engineers, developers and practitioners from academia and industry working in all interdisciplinary areas of intelligent systems and system engineering to share their experiences, and to exchange and cross-fertilize their ideas. The aim of ISDA 2016 is to serve as a forum for the dissemination of state-of-the-art research and development of intelligent systems, intelligent technologies, and applications.

The themes of the contributions and scientific sessions range from theories to applications, reflecting a wide spectrum of the coverage of intelligent systems and computational intelligence areas. ISDA 2016 received submissions from over 32 countries and each paper was reviewed by at least five reviewers in a standard peer-review process. Based on the recommendation by five independent referees, finally about 105 papers were accepted for publication in the proceedings published by Springer, Verlag.

Many people have collaborated and worked hard to produce the successful ISDA 2016 conference. First, we would like to thank all the authors for submitting their papers to the conference, for their presentations and discussions during the conference. Our thanks to Program Committee members and reviewers, who carried out the most difficult work by carefully evaluating the submitted papers. Our special thanks to J.A. Tenreiro Machado, Institute of Engineering, Polytechnic of Porto and Francisco Almada Lobo, Critical Manufacturing, Portugal for the exciting plenary talks.

We express our sincere thanks to special session chairs, organizing committee chairs for helping us to formulate a rich technical program.

Ana Maria Madureira
Ajith Abraham
Dorabela Gamboa
Paulo Novais

ISDA 2016 – Organization

General Chairs

Ana Maria Madureira	Instituto Superior de Engenharia do Porto, Portugal
Ajith Abraham	Machine Intelligence Research Labs (MIR Labs), USA

Program Chairs

Dorabela Gamboa	Polytechnic Institute of Porto, Porto, Portugal
Paulo Novais	University of Minho, Braga, Portugal

Advisory Board members

Aboul Ella Hassanien	University of Cairo, Egypt
André Ponce de Leon F de Carvalho	University of Sao Paulo, Brazil
Andries Engelbrecht	University of Pretoria, South Africa
Emilio Corchado	Universidad de Salamanca, Spain
Francisco Herrera	University of Granada, Spain
Francesco Marcelloni	University of Pisa, Italy
Hideyuki Takagi	Kyushu University, Japan
Imre J. Rudas	Obuda University, Hungary
Jeng-Shyang Pan	Harbin Institute of Technology, China
Krzysztof Cios	Virginia Commonwealth University, USA
Mario Koeppen	Kyushu Institute of Technology, Japan
Janusz Kacprzyk	Polish Academy of Sciences, Poland
Sebastián Ventura	University of Cordoba, Spain

Tutorials/Workshops and Special Sessions Chairs

Amélia Loja	Polytechnic Institute of Lisbon, Portugal
Eduardo Solteiro Pires	University of Trás-os-Montes and Alto Douro, Portugal
Leonilde Varela	Minho University, Portugal

Publication Chair

Azah Kamilah Muda Universiti Teknikal Malaysia Melaka, Malaysia

Web Master

Kun Ma University of Jinan, China

Local Organizing Committee

Judite Ferreira	Polytechnic Institute of Porto, Porto, Portugal
Ivo Pereira	Polytechnic Institute of Porto, Porto, Portugal
Bruno Cunha	Polytechnic Institute of Porto, Porto, Portugal
André Gomes	Polytechnic Institute of Porto, Porto, Portugal
José Ricardo Puga	Polytechnic Institute of Porto, Porto, Portugal
Luis Coelho	Polytechnic Institute of Porto, Porto, Portugal

Technical Program Committee

Alberto Cano	Virginia Commonwealth University
Alex James	Nazarbayev University, Kazakhstan
Alexander Mendiburu	The University of the Basque Country, Spain
Alfredo Arias Montano	Instituto Politécnico Nacional, Mexico
Alma Rahat	University of Exeter, UK
Amelia Zafra Gómez	University of Córdoba, Spain
Anan Banharnsakun	King Mongkut's University of Technology Thonburi, Thailand
Anand Jayant Kulkarni	Nanyang Technological University, Singapore
Anang Hudaya Muhamad Amin	Multimedia University, Malaysia
Andrea Corradini	University of Southern Denmark, Denmark
Andrew L. Nelson	Androtics LLC, USA

Anna Kononova	Heriot-Watt University, UK
Anne Laurent	University of Montpellier II, France
Antonio Dourado	University of Coimbra, Portugal
Antonio J. Tallón Ballesteros	Universidad de Sevilla, Spain
Antonio LaTorre	Universidad Politécnica de Madrid, Spain
Antreas Nearchou	University of Patras, Greece
Arshin Rezazadeh	Shahid Chamran University of Ahvaz, Iran
Ashwani Kush	Kurukshetra University, India
Asoke Nath	St. Xaviers College (Autonomous), India
AzaBahareh Asadi	Islamic Azad University, Iran
Biju Issac	Teesside University, UK
Bing Xue	Victoria University of Wellington, New Zealand
Binod Kumar Prasad	Maharastra Academy of Engineering, India
Carlos Alberto Ochoa Ortiz	Universidad Autónoma de Ciudad Juárez, Mexico
Carlos Fernandez Llatas	Universidad Politecnica de Valencia, Spain
Carlos Pereira	Instituto Superior de Engenharia de Coimbra, Portugal
Cerasela Crisan	"Vasile Alecsandri" University of Bacau, Romania
Chi Kin Chow	University of Hong Kong, China
Chu Hsing Lin	Tunghai University, Taiwan
Chun Wei Lin	National University of Kaohsiung, Taiwan
Clara Pizzuti	Università della Calabria, Italy
Dalia Kriksciuniene	Vilnius University, Lithuania
Dalila Duraes	Technical University of Madrid, Madrid, Spain
Daniela Zaharie	West University of Timisoara, Romania
Davide Ciucci	Università di Milano Bicocca, Italy
Diaf Moussa	Université Mouloud Mammeri, Algeria
Doaa Hassan	National Telecommunication Institute, Egypt
Edgar Alfredo Portilla Flores	Instituto Politécnico Nacional, Mexico
Eduardo Solteiro Pires	University of Trás-os-Montes and Alto Douro, Portugal
Eiji Uchino	Yamaguchi University, Japan
Eliana Costa e Silva	Polytechnic of Porto, Portugal
Eliska Ochodkova	VSB – Technical University of Ostrava, Czech Republic
Elpida Tzafestas	University of Athens, Greece
El-Sayed M. El-Alfy	King Fahd University of Petroleum and Minerals, Saudi Arabia
Fernando Jimenez	University of Murcia, Spain
Francesco Moscato	Seconda Università degli Studi di Napoli, Italy

Francisco Chicano	Universidad de Málaga, Spain
Franco Frattolillo	Università degli Studi del Sannio, Italy
Gabriel Luque	Universidad de Málaga, Spain
Gai Ge Wang	Jiangsu Normal University, China
Georg Peters	Munich University of Applied Sciences, Germany
George Georgiev	University of Wisconsin Oshkosh, USA
Georgy Sofronov	Macquarie University, Australia
Ghanshyam Thakur	National Institute of Technology, India
Giancarlo Mauri	Università di Milano Bicocca, Italy
Heder Bernardino	Universidade Federal de Juiz de For a, Brazil
Ian Sillitoe	University of Wolverhampton, UK
James O'Shea	Manchester Metropolitan University, UK
José Raúl Romero	University of Córdoba, Spain
Jose Santos	Universidad de A Coruña, Spain
Kang Tai	Nanyang Technological University, Singapore
Konstantinos Parsopoulos	University of Ioannina, Greece
Korhan Karabulut	Yaşar Üniversitesi, Turkey
Kun Ma	University of Jinan, China
Laurence Amaral	Universidade Federal de Uberlândia, Brazil
Leonardo Trujillo	Instituto Tecnológico de Tijuana, Mexico
Leonilde Varela	Universidade do Minho, Portugal
Li Zhang	Soochow University, China
Luigi Benedicenti	University of Regina, Canada
Majid Bakhtiari	Universiti Teknologi Malaysia, Malaysia
Mario Koeppen	Kyushu Institute of Technology, Japan
Marjan Kuchaki Rafsanjani	Shahid Bahonar University of Kerman, Iran
Mauricio Ayala Rincon	Universidade de Brasilia, Brazil
Mengjie Zhang	Victoria University of Wellington, New Zealand
Miguel Rocha	Universidade do Minho, Portugal
Mohammad Reza Bonyadi	The University of Adelaide, Australia
Ngo Hea Choon	Universiti Teknikal Malaysia Melaka, Malaysia
Patrick Siarry	Universit de Paris, France
Paulo Moura Oliveira	University of Trás-os-Montes and Alto Douro, Portugal
Ricardo Landa	Instituto Politécnico Nacional, Mexico
Riyad Alshammari	Dalhousie University, Canada
Roberto Antonio Vázquez Espinoza De Los Monteros	Lasallistas de Corazón, Mexico
Rodica Ioana Lung	Babeş-Bolyai University, Romania
Solteiro Pires	UTAD University, Portugal
Sung Bae Cho	Yonsei University, South Korea

Tatiana Kalganova	Brunel University London, UK
Thomas Hanne	University of Applied Sciences Northwestern Switzerland, Switzerland
Tien Tsin Wong	University of Hong Kong, China
Tzung Pei Hong	National University of Kaohsiung, Taiwan
Victor Manuel Landassuri Moreno	UAEM Valle de Mexico, Mexico
Vincenzo Piuri	Università degli Studi di Milano, Italy
Wei Chiang Hong	Oriental Institute of Technology, Taiwan
Yi Mei	RMIT University, Australia
Ying Ping Chen	National Chiao Tung University, Taiwan
Yogan Jaya Kumar	Universiti Teknikal Malaysia Melaka, Malaysia
Yong Wang	Central South University, China
Yun Huoy Choo	Universiti Teknikal Malaysia Melaka, Malaysia

Contents

An Innovative Approach to Manage Heterogeneous Information Using Relational Database Systems 1
Cosmin Sabo, Petrică C. Pop, Honoriu Vălean, and Daniela Dănciulescu

Reliable Attribute Selection Based on Random Forest (RASER) 11
Aboudi Noura, Hechmi Shili, and Lotfi Ben Romdhane

Estimating the Number of Clusters as a Pre-processing Step to Unsupervised Learning 25
Paulo Rogerio Nietto and Maria do Carmo Nicoletti

Agglomerative and Divisive Approaches to Unsupervised Learning in *Gestalt* Clusters 35
Rodrigo C. Camargos, Paulo R. Nietto, and Maria do Carmo Nicoletti

Improving Imputation Accuracy in Ordinal Data Using Classification . . . 45
Shafiq Alam, Gillian Dobbie, and XiaoBin Sun

GA-PSO-FASTSLAM: A Hybrid Optimization Approach in Improving FastSLAM Performance 57
Alif Ridzuan Khairuddin, Mohamad Shukor Talib, Habibollah Haron, and Muhamad Yazid Che Abdullah

Three Case Studies Using Agglomerative Clustering 67
Rodrigo C. Camargos and Maria do Carmo Nicoletti

Historic Document Image De-noising Using Principal Component Analysis (PCA) and Local Pixel Grouping (LPG) 77
Han-Yang Tang, Azah Kamilah Muda, Yun-Huoy Choo, Noor Azilah Muda, and Mohd Sanusi Azmi

A Robust and Optimally Pruned Extreme Learning Machine 88
Ananda L. Freire and Ajalmar R. Rocha Neto

Investigating the Effect of Combining Text Clustering with Classification on Improving Spam Email Detection 99
Doaa Hassan

Radial Basis Function Neural Networks for Datasets with Missing Values 108
Diego P. Paiva Mesquita and João Paulo P. Gomes

Evaluation Method for an Adaptive Web Interface: GOMS Model 116
Rebai Rim, Maalej Mohamed Amin, Mahfoudhi Adel, and Abid Mohamed

Diversification Strategies in Differential Evolution Algorithm to Solve the Protein Structure Prediction Problem 125
Pedro Henrique Narloch and Rafael Stubs Parpinelli

Using Cluster Barycenters for the Generalized Traveling Salesman Problem 135
Mehdi El Krari, Belaïd Ahiod, and Bouazza El Benani

On Pollution Attacks in Fully Connected P2P Networks Using Trusted Peers 144
Cristóbal Medina-López, Ilshat Shakirov, L.G. Casado, and Vicente González-Ruiz

Certification Under Uncertainties of Control Methods for Multisource Elevators 154
Chloé Desdouits, Mazen Alamir, Rodolphe Giroudeau, and Claude Le Pape

Robust and Reliable Bionic Optimization of Nonlinear Control Problems 165
Lukas Haas and Rolf Steinbuch

Human Detection Using Biological Signals in Camera Images with Privacy Aware 175
Toshihiro Kitajima, Edwardo Arata Y. Murakami, Shunsuke Yoshimoto, Yoshihiro Kuroda, and Osamu Oshiro

Nuclei Malignancy Analysis Based on an Adaptive Bottom-Hat Filter 187
Tiia Ikonen, Keijo Haataja, Pekka Toivanen, Teemu Tolonen, and Jorma Isola

Self-Organizing Maps and Fuzzy C-Means Algorithms on Gait Analysis Based on Inertial Sensors Data 197
Rafael Caldas, Yabing Hu, Fernando Buarque de Lima Neto, and Bernd Markert

Age, Gender, Race and Smile Prediction Based on Social Textual and Visual Data Analyzing 206
Onsa Lazzez, Wael Ouarda, and Adel M. Alimi

Test Suite Prioritization Using Nature Inspired Meta-Heuristic Algorithms 216
Daya Gupta and Vishal Gupta

A New Trajectory Optimization Approach for Safe Mobile Robot Navigation: A Comparative Study (Khepera II Mobile Robot) 227
Walid Ellili, Abdelfetteh Lachtar, and Mounir Samet

The Improvement of an Image Compression Approach Using Weber-Fechner Law 239
Mourad Rahali, Habiba Loukil, and Mohamed Salim Bouhlel

A Minimal Rare Substructures-Based Model for Graph Database Indexing 250
Mehdi Azaouzi and Lotfi Ben Romdhane

Multibiometrics Enhancement Using Quality Measurement in Score Level Fusion 260
Saliha Artabaz, Layth Sliman, Hachemi Nabil Dellys, Karima Benatchba, and Mouloud Koudil

Effects of Random Sampling on SVM Hyper-parameter Tuning 268
Tomáš Horváth, Rafael G. Mantovani, and André C.P.L.F. de Carvalho

Bayesian Networks for Identifying Semantic Relations in a Never-Ending Learning System 279
Edimilson Batista dos Santos, Massilon Lourenço Fernandes, Estevam R. Hruschka, Jr., and Maísa Cristina Duarte

Training a Spiking Neural Network to Generate Online Handwriting Movements 289
Mahmoud Ltaief, Hala Bezine, and Adel M. Alimi

Using a Synthetic Character Database for Training Deep Learning Models Applied to Offline Handwritten Recognition 299
Jorge Sueiras, Victoria Ruiz, Ángel Sánchez, and José F. Vélez

Cognitive Conversation Language - CCL 309
Wesley Willy Oliveira de Souza, and Estevam Rafael Hruschka Júnior

A Genetic-Fuzzy Classification Approach to Improve High-Dimensional Intrusion Detection System 319
Imen Gaied, Farah Jemili, and Ouajdi Korbaa

A New Data Placement Approach for Scientific Workflows in Cloud Computing Environments 330
Hamdi Kchaou, Zied Kechaou, and Adel M. Alimi

A Novel Simulated Annealing-Based Learning Algorithm for Training Support Vector Machines 341
Madson L. Dantas Dias and Ajalmar R. Rocha Neto

Fish School Search Variations and Other Metaheuristics in the Solution of Assembly Line Balancing Problems 352
I.M.C. de Albuquerque, J.B.M. Filho, F.B.L. Neto, and A.M.O. Silva

Improved Search Mechanisms for the Fish School Search Algorithm . . . 362
João Batista Monteiro Filho, Isabela Maria Carneiro Albuquerque, Fernando Buarque Lima Neto, and Filipe Vieira Silva Ferreira

A Survey on Outlier Detection in the Context of Stream Mining: Review of Existing Approaches and Recommadations 372
Imen Souiden, Zaki Brahmi, and Hajer Toumi

Heuristic for Scheduling Intrees on *m* Machines with Non-availability Constraints 384
Khaoula Ben Abdellafou, Hatem Hadda, and Ouajdi Korbaa

Network Intrusion Detection Using Danger Theory and Genetic Algorithms 394
João Lima Santanelli and Fernando Buarque de Lima Neto

A Gamification Model for Resource Sharing in Malaysian Schools Using Cloud Computing Platform 406
M. Nordin A. Rahman, Abdullahi Nababa Saidu, M. Fadzil A. Kadir, Syadiah Nor Shamsudin, and Syarilla Iryani A. Saany

A Modified Naïve Possibilistic Classifier for Numerical Data 417
Karim Baati, Tarek M. Hamdani, Adel M. Alimi, and Ajith Abraham

Combinatorial Structural Clustering (CSC): A Novel Structural Clustering Approach for Large Scale Networks 427
Liang Chen, Hongbo Liu, Weishi Zhang, and Bo Zhang

Memetic Algorithms for the Automatic Discovery of Software Architectures 437
Aurora Ramírez, Rafael Barbudo, José Raúl Romero, and Sebastián Ventura

Multi-objective Dynamic Analysis Using Fractional Entropy 448
E.J. Solteiro Pires, J.A. Tenreiro Machado, and P.B. de Moura Oliveira

Multi-agent Based Truck Scheduling Using Ant Colony Intelligence in a Cross-Docking Platform 457
Houda Zouhaier and Lamjed Ben Said

From Traditional Data Warehouse To Real Time Data Warehouse 467
Senda Bouaziz, Ahlem Nabli, and Faiez Gargouri

Intelligent Traffic Congestion Prediction System Based on ANN and Decision Tree Using Big GPS Traces 478
Wiam Elleuch, Ali Wali, and Adel M. Alimi

Forecasting Using Elman Recurrent Neural Network 488
Emna Krichene, Youssef Masmoudi, Adel M. Alimi, Ajith Abraham, and Habib Chabchoub

Multi-objective Particle Swarm Optimisation for Robust Dynamic Scheduling in a Permutation Flow Shop 498
Mohanad Al-Behadili, Djamila Ouelhadj, and Dylan Jones

Mining Perfectly Rare Itemsets on Big Data: An Approach Based on Apriori-Inverse and MapReduce 508
F. Padillo, J.M. Luna, and S. Ventura

Patient-Specific Epilepsy Seizure Detection Using Random Forest Classification over One-Dimension Transformed EEG Data 519
Marco A. Pinto-Orellana and Fabio R. Cerqueira

A New Approach to Human Activity Recognition Using Machine Learning Techniques 529
Leandro B. Marinho, A.H. de Souza Junior, and P.P. Rebouças Filho

Lung Segmentation in Chest Computerized Tomography Images Using the Border Following Algorithm 539
Murillo B. Rodrigues, Leandro B. Marinho, Raul Victor M. Nóbrega, João Wellington M. Souza, and Pedro Pedrosa Rebouças Filho

Learner's Profile Hierarchization in an Interoperable Education System 549
Leila Ghorbel, Corinne Amel Zayani, Ikram Amous, and Florence Sèdes

ACO-PSO Optimization for Solving TSP Problem with GPU Acceleration 559
Olfa Bali, Walid Elloumi, Ajith Abraham, and Adel M. Alimi

A Parallel Adaptive PSO Algorithm with Non-iterative Electrostatic Repulsion and Social Dynamic Neighborhood 570
Daniel Soto and Wilson Soto

Fishing Monitor System Data: A Naïve Bayes Approach 582
Joao C. Ferreira, Serge Lage, Iola Pinto, and Nuno Antunes

Detection of Behavioral Patterns for Increasing Attentiveness Level 592
Dalila Durães, Sérgio Gonçalves, Davide Carneiro, Javier Bajo, and Paulo Novais

Towards a Medical Intensive Care Unit Decision Support System Based on Intuitionistic Fuzzy Logic 602
Hanen Jemal, Zied Kechaou, and Mounir Ben Ayed

M2Onto: An Approach and a Tool to Learn OWL Ontology from MongoDB Database 612
Hanen Abbes and Faiez Gargouri

Comparison of Hard and Probabilistic Evidence in Bayesian Model . . . 622
Rebai Rim, Maalej Mohamed Amin, Mahfoudhi Adel, and Abid Mohamed

Towards Activity Theory - Preliminary Report: Ambient Intelligence Applied to Smart Education 630
Frederica Gonçalves, Eduardo Fermé, Ana Lúcia Martins, and João C. Ferreira

Emerging Opportunities for Ambient Intelligence in Creativity Support Tools 640
Frederica Gonçalves, Eduardo Fermé, and João C. Ferreira

Time Series Data Mining for Energy Prices Forecasting: An Application to Real Data 649
Eliana Costa e Silva, Ana Borges, M. Filomena Teodoro, Marina A.P. Andrade, and Ricardo Covas

Efficient Parameterization for Automatic Speaker Recognition Using Support Vector Machines 659
Rania Chakroun, Mondher Frikha, and Leila Beltaïfa Zouari

Security Incident Response: Towards a Novel Decision-Making System 667
Samih Souissi, Ahmed Serhrouchni, Layth Sliman, and Benoit Charroux

A New Social Media Mashup Approach 677
Abir Troudi, Corinne Amel Zayani, Salma Jamoussi, and Ikram Amous

Enabling the Definition and Reuse of Multi-Domain Workflow-Based Data Analysis 687
Rubén Salado-Cid and José Raúl Romero

An Improved Elman Neural Network for Daily Living Activities Recognition 697
Zaineb Liouane, Tayeb Lemlouma, Philippe Roose, Fréderic Weis, and Hassani Messaoud

Towards an Approach Based on Ontology for Semantic-Temporal Modeling of Social Network Data 708
Rim Chiha and Mounir Ben Ayed

Evaluation of the Simulated Annealing and the Discrete Artificial Bee Colony in the Weight Tardiness Problem with Taguchi Experiments Parameterization 718
André S. Santos, Ana M. Madureira, and Maria R. Varela

Industrial Plant Layout Analyzing Based on SNA 728
M.L.R. Varela, V.K. Manupati, K. Manoj, G.D. Putnik, A. Araújo, and A.M. Madureira

A Genetic Neural Network Approach for Unusual Behavior Prediction in Smart Home 738
Zaineb Liouane, Tayeb Lemlouma, Philippe Roose, Fréderic Weis, and Hassani Messaoud

New Adaptive Resource Allocation Scheme in LTE-Advanced 749
Radhia Khdhir, Kais Mnif, Aymen Belghith, and Lotfi Kamoun

An Approach for Measuring Flexibility of Business Processes Based on Distances Between Models and Their Variants 760
Asma Mejri, Sonia Ayachi Ghannouchi, and Ricardo Martinho

Specification of an Architecture for Self-organizing Scheduling Systems 771
Ana Madureira, Ivo Pereira, and Bruno Cunha

Towards Better SWRL Rules Dependency Extraction 781
Abir Boujelben, Tarak Chaari, and Ikram Amous

Clustering of Maintenance Tasks for the Danish Railway System 791
Shahrzad M. Pour and Una Benlic

Evaluating the Quality of Business Process Models Based on Measures and Criteria in Higher Education: Developing a Framework for Continuous Quality Improvement 800
Fouzia Kahloun and Sonia Ayachi Ghannouchi

Age and Gender Classification from Finger Vein Patterns 811
Wafa Damak, Randa Boukhris Trabelsi, Alima Damak Masmoudi, Dorra Sellami, and Amine Nait-Ali

Coupling Event-B/ProB for the Analysis of the Software Architecture Evolution Described in PDDL 821
Farah Fourati, Mohamed Tahar Bhiri, and Riadh Robbana

Trust Intrusion Detection System Based on Location for Wireless Sensor Network 831
Hela Maddar, Wafa Kammoun, and Habib Youssef

Temporal Patterns Visualization for Knowledge Acquisition in Dynamic Decision-Making Environment 841
Jihed Elouni, Hela Ltifi, Mounir Ben Ayed, and Mohamed Masmoudi

Cloudlets Architecture for Wireless Sensor Network 852
Hela Maddar, Wafa Kammoun, and Habib Youssef

Data Fusion Classification Method Based on Multi Agents System 863
Elhoucine Ben Boussada, Mounir Ben Ayed, and Adel M. Alimi

Community Detection in Bipartite Networks Using a Noisy Extremal Optimization Algorithm 871
Noémi Gaskó, Rodica Ioana Lung, and Mihai Alexandru Suciu

An Approach for Incorporating the Usability Optimization Process into the Model Transformation 879
Marwa Hentati, Lassaad Ben Ammar, Abdelwaheb Trabelsi, and Adel Mahfoudhi

An ILS Heuristic for the Waste Collection Vehicle Routing Problem with Time Windows 889
Alba A. Campos and José Elias C. Arroyo

An NLP-Based Ontology Population for Intentional Structure 900
Noura Labidi, Tarak Chaari, and Rafik Bouaziz

Metaheuristics Parameter Tuning Using Racing and Case-Based Reasoning in Scheduling Systems 911
Ivo Pereira, Ana Madureira, and Bruno Cunha

A Branch-and-Price Algorithm for the Double Vehicle Routing Problem with Multiple Stacks and Heterogeneous Demand 921
Jonatas B.C. Chagas and André G. Santos

BigDimETL: ETL for Multidimensional Big Data 935
Hana Mallek, Faiza Ghozzi, Olivier Teste, and Faiez Gargouri

Knowledge Integration in Collaborative Environments Using Supervised Ontological Alignment 945
Leandro Pupo Natale and Nizam Omar

A General VNS Heuristic for a Three-Stage Assembly Flow Shop Scheduling Problem 955
Saulo C. Campos, José Elas C. Arroyo, and Ricardo G. Tavares

Towards NoSQL Graph Data Warehouse for Big Social Data Analysis 965
Hajer Akid and Mounir Ben Ayed

A Contribution of Dynamical Systems Theory and Epidemiological Modeling to a Viral Marketing Campaign 974
João N.C. Gonçalves, Helena Sofia Rodrigues, and M. Teresa T. Monteiro

Developing an Ambient Intelligent-Based Decision Support System for Production and Control Planning 984
Marco Gomes, Fábio Silva, Filipa Ferraz, António Silva, Cesar Analide, and Paulo Novais

CobWeb Multidimensional Model: Visualizing OLAP Query Results Using Tag-Cloud Operators 995
Omar Khrouf, Kaïs Khrouf, and Jamel Feki

A Time Delay Neural Network for Online Arabic Handwriting Recognition 1005
Ramzi Zouari, Houcine Boubaker, and Monji Kherallah

A Proposal to Model Knowledge Dimension in Sensitive Business Processes 1015
Mariam Ben Hassen, Mohamed Turki, and Faïez Gargouri

Using Intelligent Systems to Improve Case Flow in Court Systems 1031
Ana Lúcia Martins

Similarity and Trust Metrics Used in Recommender Systems: A Survey 1041
Maryam Jallouli, Sonia Lajmi, and Ikram Amous

Analysing the Performance of a Tomographic Reconstructor with Different Neural Networks Frameworks 1051
Sergio Luis Suárez Gómez, Carlos González Gutiérrez, Jesús Daniel Santos Rodríguez, María Luisa Sánchez Rodríguez, Fernando Sánchez Lasheras, and Francisco Javier de Cos Juez

Robot Swarms Theory Applicable to Seek and Rescue Operation 1061
Jose León, Gustavo A. Cardona, Andres Botello, and Juan M. Calderón

Linguistic Representation by Fuzzy Formal Concept and Interval Type-2 Feature Selection 1071
Sahar Cherif, Nesrine Baklouti, Adel M. Alimi, and Vaclav Snasel

Author Index 1083

An Innovative Approach to Manage Heterogeneous Information Using Relational Database Systems

Cosmin Sabo[1(✉)], Petrică C. Pop[1], Honoriu Vălean[1], and Daniela Dănciulescu[2]

[1] Technical University of Cluj-Napoca, Cluj-Napoca, România
cosmin_sabo@cunbm.utcluj.ro
[2] University of Craiova, Craiova, România

Abstract. In this paper, we propose a novel database design structure that can deal with all the aspects of the complexity of data that has to be managed, using the concepts of defining objects in object oriented programming (OOP). As well, we create a set of procedures in database system that allows us to manage all type of data, without knowing the structure of the database. The creation of the database structure, also the mechanism of inserting and retrieval the information is made by using a metadata set of information.

The major benefit of the proposed approach is that we can use a relational database management system (RDBMS), that can assure ACID (atomicity, consistency, isolation and durability) principles, low cost management and quick development based on metadata structure. The main advantages of our approach comparing with NoSQL database system is that we preserve ACID properties of the information and comparing with NewSQL is that the cost of the projection of database structure and management of the system is much lower.

Our proposed system is functional and can manage very large amount of data from heterogenic sources that can be managed by companies without a lot of know-how.

Keywords: Databases · Data logical models · Relational models · Object-oriented models

1 Introduction

How to manage big quantities of dataset with a high rate of changes from heterogeneous sources and preserve in the same time ACID (atomicity, consistency, isolation and durability) principles of database systems is a question that each of us has to deal with.

Most of the applications used by companies to manage their data use relational database management systems (RDMSs) and this fact will remain for the next years, even though database management systems (DMSs) are evolving very fast. On the other hand, data structure, are changing often and the number of data sources needed to be used by companies to remain competitive are rising every day. This is the reason why we create this new approach in order to retrieve and manage information datasets.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_1

The evolution of database management system capabilities is a clear indicator of the existence of necessity of finding mechanisms for structuring information. Databases are defined by Frawley [7] as collections of data integrated into one or more files, organized to facilitate the storage, change, query and retrieval of information relevant to users needs. Frawley estimated that global information generated will double every 20 months and the size and number of databases will grow even faster, see for more information [7]. Coltri believes that the database is one of the areas with the most important developments in software engineering [19].

In the last period, librarianship has developed and implemented a suite of standards for information management, and for ensuring the sustainability of stored information. We will mention few of these standards used for resource description: BIBFRAME - Bibliographic Framework Initiative [8], EAD - Encoded Archival Description [9] EDTF - Extended Date/Time Format, MARCXML - Machine-Readable Cataloging XML [10], MODS - metadata Object Description Standard, but there are standards for digital resources, such as PREMIS - Preservation metadata or METS - metadata Encoding and Transmission Standard [11]. These standards can be extended in order to be used in other areas.

As is shown by Rokitskii [5], object oriented models are data logical in their essence but, at the same time, since object oriented models are formal, they allow one to specify formal constructions of data, formal relations between them, and formal operations on them. Object-oriented modeling has been already used at the interface level of three-level architectures of database management systems (DBMSs) and also at the conceptual level of design. In contrast to object-oriented tools occupying a rather large segment of the market of creation of application programs at the present time, the market of object-oriented database management systems has a low acceptance in small and medium business software. Based on this premises, Rokitskii [5] used relational database management systems (DBMSs) in order to create a data structure model based on object-oriented principals. For more information, see [5].

Vysniauskas and Nemuraite in [6] have presented a solution for a problem that has arisen from practical needs: namely, possibilities for storing ontological information and processing this information by user applications. For this purpose, they used relational database (RDB) considering that it is a good candidate that has proven capabilities to cope with large amounts of data. Methodologies for transforming entity relationship and object-oriented conceptual models to relational database structures are well-established and implemented in their tools.

Ontology Definition Metamodel, initiated by OMG, is seeking to define transformations between OWL, UML, ER and other modelling languages, where Simple Common Logic is chosen for definition of constraints. On the base of existing methodologies, there are some possible ways to relate ontological information described by ontological language, with relational schemas. For more information, see [4].

In this paper we present a novel solution of structuring the information in a relational database system. Our approach is using concepts from library standards and object oriented programming, and is totally different from previous studies.

The aim of this paper is to describe an innovative approach that combines informational standards from librarianship field with object oriented programming techniques in order to manage financial and contact information of possible partners or clients.

2 Defining the Concepts Used for Solving this Problem

2.1 Library Standards for Defining Information

MARC (Machine Readable Cataloging) is currently the most widely used standard for storing and exchanging bibliographic records. Evolution of MARC standard has a history of more than forty years. MARC, is basically a concept for structuring and interchanging of information which evolved separately from concepts of management systems databases. Even if it is a standard with a high degree of use in libraries, it does not solve all problems arising from the rapid evolution of data structures and quantity of existing information.

The structure of the information in MARC format is a linear one, each type of information stored is defined by a set of metadata composed from a set of three numerical characters that defined field, an element represented by a character alphanumeric that represents subfield and field label to define the meaning of this field subfield tuple [16]. The information can be stored in field or in a subfield. Basically, if you want to specify the author of a work, the information that will identify metadata is filed 100, and the corresponding value will be written to the right of this metadata, in subfield a, as you can see in Fig. 1.

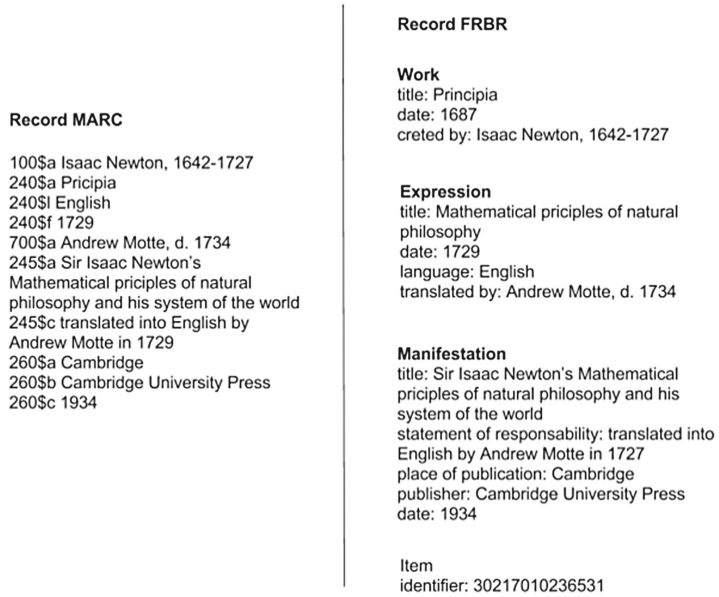


Fig. 1. A bibliographic record represented in MARC and FRBR model

If there is repetitive information for a given metadata field, we will add a new metadata representing information and the corresponding value. If a subfield is repetitive, $ separator will be used followed by subfield name to write repetitive set of values. The maximum number of characters that can be stored in a field or subfield, of MARC standard, is 999 characters.

MARC standard has proved that are some inconvenient in information management:

- The number of fields and subfields are insufficient to encapsulate all type of information needed to be represented;
- Fields and subfields containing a large amount of information are not possible to be managed;
- Special characters can generate errors in automated process.

Based on this fact regarding MARC standards, MARCXML standard was developed. This standard representation of information eliminates previously existing limitations in storing large amounts of information in a single subfield or storing special chars.

This linear way of defining information within a bibliographic record cannot solve all situations generated by evolution of information needed to be stored, ensuring sustainability of the concerned information with minimal redundancy, but allows a semantic presentation of the information stored on this structure [3]. In this sense, to solve the aforementioned issues, IFLA - International Federation of Library Associations and other institutions have proposed and developed FRBR - Functional Requirements for Bibliographic Records, which adopts a description of hierarchical information in order to increase the level of granularity and to allow better information reuse.

2.2 Representing Information Structures Using FRBR Model

FRBR [2] reduces the number of descriptive elements at record level, but significantly increases the number of records that can be described in separate facilities, as is shown in Fig. 2. The linear model description of records in MARC standard is converted into a hierarchical model in FRBR model, and this model can be mapped to object oriented information, but this aspect will be treated in the next chapter.

Clearly, in FRBR representation there are few shortcomings [12], because the aim of defining FRBR was to reduce the cost of classical description of bibliographic resources, but has left aside other aspects such as:

- Prioritizing information at field level;
- Repetitive order of subfields repetitive not standardized;
- Another important aspect is related to the description of records or parts of records in a different language.

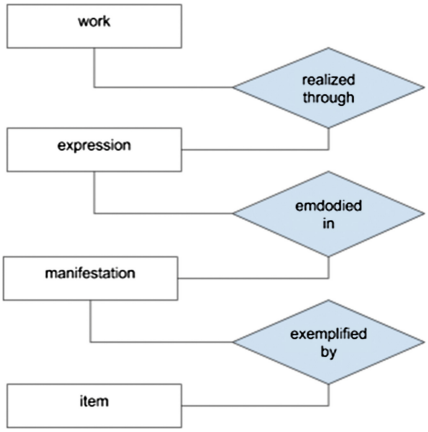


Fig. 2. Representing relations in a record FRBR

3 Problem Solution Using Metadata Definitions

Next we will define a complex informational structure using relational database management systems [18] that is easy to be managed and we will show how to combine the expertise gained by the team members in librarian bibliographic standards [1], relational database management systems and object-oriented programming [17].

Object-oriented programming (OOP) refers to a type of computer programming (software design) in which programmers define not only the data type of a data structure, but also the types of operations (functions) that can be applied to the data structure.

In this way, the data structure becomes an object that includes both data and functions. In addition, programmers can create relationships between one object and another. For example, objects can inherit characteristics from other objects [15].

OOP concepts allow the development of simple and effective solutions to many problems, enabling software programming complexity decreased, but improving the quality of the solutions obtained [14].

MARC standard with the list of fields and subfields can be associated with linear programming where we can manage, perhaps hundreds of variables and procedures, but increasing their number can generate unmanageable situations. Obviously, the continued growth of the types of records required to be described will generate situations hard to be manage in a linear structure.

Managing complex informational structures without the risk of losing information and in the same time providing tools for easy information retrieval, it is not an easy task, and this is the reason why we define a normalized information structure, using MARC definitions and concept of structuring objects specific to object oriented programming, that allow us to have an easy information management.

The MARC structure, which according to Fig. 1, is defined as linear way of defining information was mapped into a structure on three levels as we will describe in what it follows.

The first level defines the uniqueness of registration, and a set of adjacent properties. This level is equivalent in object-oriented programming with class variables defined at class level, each record in the database consist in an instance of this class.

The second level consists in the definitions MARC fields, their names having a fixed length of three characters and values between 001 and 999. Each field has clear defined meaning, and the meaning of the fields is grouped for easy identification. This level is equivalent with methods from object oriented programming.

The third level is defined by subfields, that have a fixed length that consists in one character that has values between a–z and 0–9. Most of the information is stored at subfield level. There are only a few fields that store information, the remaining fields representing the relationship between subfield and record level. This concept allows an easy way to define 1:n and n:m relations. These subfields represent in object oriented programming the variables defined inside methods.

The principle of defining the database structure for the example above is shown in Fig. 3. Since the number of tables and relations between tables is too big to be exhaustive represented here, we represent those elements that are representative for the principles shown before.

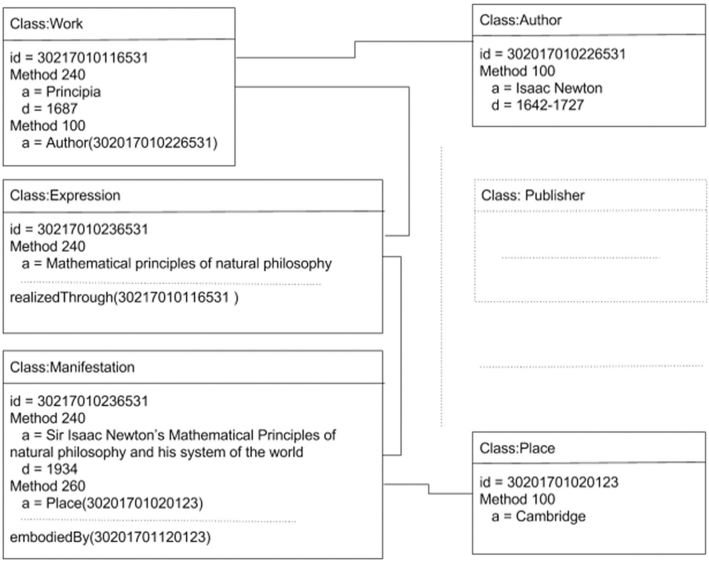


Fig. 3. OO information mapping using MARC fields

This concept has as starting point librarianship and information science, and concepts specific to object-oriented programming, which were used to define a metadata structure, which automatically generate data structures necessary through stored procedures. The result is that any user can manage all this data structure and information stored in without knowledge of database management or knowledge about standard query language, the entire process of defining and extending the database structure is based on the metadata definition.

Basically, we defined a set of tables in the database that store all metadata we need to generate: structures of tables, fields, data types, restrictions and relations between them. Based on triggers, when changes are made on these tables the appropriate operation for adding, modifying or deleting tables and relations between them are generated.

Similarly, when we interrogate the database we shall use a single stored procedure with the list of required parameters to be entered, such as the list of fields needed, the list of conditions to be fulfilled and other optional parameters.

It does not matter if you want to take information about a person, a task list, employment or any other information, you can always call the same procedure, without needing to know the database structure.

This model is implemented using MySQL database system, a system that allows storing sufficient quantities of information at the level of TB or PB, with the possibility of clustering. The concepts defined in this material may be used for other database systems. The reason for choosing this database system is the ability to achieve rapid transition from SQL to NoSQL [13] using schemas. For further studies we intend to compare our results with other NoSQL database systems and also, to map our structure to some web ontology language (OWL) [6].

4 Preliminary Results

We took in consideration datasets provided by http://data.gov.ro, a governmental website, that offers information under licensed distribution OGL-ROU-1.0, which allows the collection, processing and distribution of this information without any warranty. Now the website contains 669 sets of structured data in 16.770 files structured in various formats and informational structure.

In order to highlight the innovative concept presented in this article, we took information about Romanian companies in a set of years (2016, 2015, 2014 and 2013). Each set of data has source files consist of six separate documents structured as CSV (comma separated values) format, the content of this files is different.

Also, in order to check the entire functionality of the mentioned concepts we have taken information about companies in Romania, from other public governmental sites, to ensure the heterogeneity of information. Data source in this case was in html format, retrieving information was performed using XPath.

The implementation process has required the following steps:

- Defining the structure based on information provided by the documents published on the website data.gov.ro and information retrieved from other websites;
- Identifying other sites that provide information about public companies;
- Extension metadata structure to embed this new information;
- Taking information from files provided by data.gov.ro;
- Generate the automation to extract relevant data from HTML (Hyper Text Markup Language) files;
- Insert datasets into database structure using stored procedures specialized for this operation.

The metadata information defines 41 fields and a total of 89 subfields. The next figure shows on the left side the list of fields defined in this metadata structure and the type of information that can be stored at field level, with length limitation if it is needed and on the right side, the list of subfields defined for the filed selected (Fig. 4).

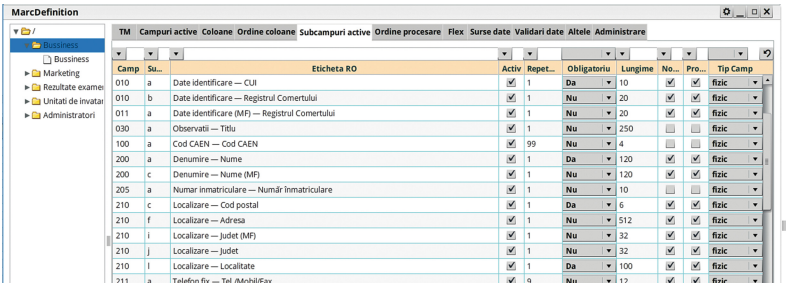


Fig. 4. Field and subfield metadata definition

Metadata set allow to define if a field or a subfield is active or not, based on this information, stored procedures will allow or not to insert information in field or subfield. We also define if a subfield is repetitive or not, how many times it can be repeated, maximum length of information stored in it, if the value stored in a field subfield tuple is entered by user, or is generated based on some automatic procedures specific to each data structure.

This metadata structure defines if a field or subfield value should be validated by a regular expression and if this validation does not pass what will be the results returned to user. Validation rule can send to user an info note that this information does not respect the rules defined, a warning, or can generate an error message and reject the insert or update operation.

Each field or subfield can have different data source information. For example, subfield CAEN code (activities classification) has as source a dictionary of values, based on fact that CAEN code is a predefined list. The owner of the company represents an authority information, which means that behind that name is a complex structure that respects the same principals. The owner company name is an authority structure because we can add another information about the owner, like contact information. The complexity of this information can be extended based on new data source information that we find. Link data source represents a connection to another data set from the same structure. For example, if a company is connected to another, we can represent this by using the link data source.

After the insertion process in database based on metadata that define the entire structure the result is 1.738.053 companies from Romania with contact information and main financial indicators from the last eight years.

To obtain the list of companies that exists in database we use the next stored procedure:

```
marc_select_advanced_fast('b2b', 'rou', '', '', '
```

The parameters set in this example show object type referred, in this case 'b2b'; also if a record it is stored in more languages it allows us to select the preferred language, values from list of field subfield tuples that will be returned, data offset is 539.973 and the number of results is 27. Parameter '*' after field subfield tuple means it will return all values if a field subfield tuple is repetitive.

If we want to find all the companies from Bucharest that have a mobile phone and email address in database, we will use:

```
CALL marc_select_advanced_fast_count('b2b', 'rou', '', '', '', '', '', '', '', '', '', '', '', '',
'', '210j', 'b', 'București', '', '212a', 'b', '07', '', '217a', 'c', '@', '', '', '', '', '', '', '', '', '', '', '', '',
'', '');
```

5 Conclusions

In this paper, we consider the problem of defining a complex database structure, based on premises that the input data are provided by heterogeneous sources and sources are changing the information provided often.

We defined a set of metadata that generates entire data structure using relational database management system and the set of stored procedures that allow us to manage the information stored in this data structure without needing to know the internal structure.

Some important features of this concept:

- Capability of generating complex data structures only using a metadata definition
- Data management using only one stored procedure to insert any type of data
- Data search and retrieve using one stored procedures
- Structure management based on metadata definition

Implementation of this concept was made using a large amount of data from heterogeneous sources.

References

1. Frâncu, V., Sabo, C.: Implementation of a UDC-based multilingual thesaurus in a library catalogue: the case of bibliophil. Knowl. Organ. **37**(3), 209–215 (2010)
2. Carlyle, A.: Understanding FRBR as a conceptual model: FRBR and the bibliographic universe. Libr. Resour. Tech. Serv. **50**(4), 264–273 (2006)
3. Giannopoulou, E., Mitrou, N., Chimos, K., Karvounidis, T., Douligeris, C.: A semantic web approach in the implementation of a linked data portal using a CMS. In: Proceedings of 10th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS, p. 164 (2014)
4. Agosti, M., Crivellari, F., Di Nunzio, G.M., Gabrielli, S.: Understanding user requirements and preferences for a digital library web portal. Intl. J. Digital Libr. **11**(4), 225–238 (2010)
5. Rokitskii, R.B.: Object-oriented databases with relational DBMSs. Cybern. Syst. Anal. **36**(6), 813–822 (2000)
6. Vysniauskas, E., Nemuraite, L.: Transforming ontology representation from OWL to relational database. Inf. Technol. Control **35**(3A), 333–343 (2006)

7. Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J.: Knowledge discovery in databases: an overview. AI Mag. **13**(3), 57–70 (1992)
8. http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf
9. Gartner, R.: An XML schema for enhancing the semantic interoperability of archival description. Arch. Sci. **15**(3), 295–313 (2015)
10. Gardner, J.R.: Information architecture planning with XML. Libr. Hi Tech **19**(3), 231–241 (2001)
11. Cheslow, S.: METS for the cultural heritage community: a literature review. Libr. Philos. Pract. **2014**(1) (2014)
12. Pacheco, K.L., Ortega, C.D.: Model FRBR in origin. Biblios **60**, 63–75 (2015)
13. Lee, C., Zheng, Y.: Automatic SQL-to-NoSQL schema transformation over the MySQL and HBase databases. In: IEEE International Conference on Consumer Electronics - Taiwan, ICCE-TW, pp. 426–427 (2015)
14. Cavaiani, T.P.: Object-oriented programming principles and the Java class library. J. Inf. Syst. Educ. **17**(4), 365 (2006)
15. SAP AG: Patent issued for systems and methods for generating a common data model for relational and object oriented databases. J. Eng. 3005 (2013)
16. Holler, J., Tsiatsis, V., Mulligan, C., Avesand, S., Karnouskos, S., Boyle, D.: From Machine-To-Machine to the Internet of Things, pp. 1–331. Academic Press, Cambridge (2014)
17. Pop, P.C., Matei, O., Sabo, C.: A new approach for solving the generalized traveling salesman problem. In: Blesa, María J., Blum, C., Raidl, G., Roli, A., Sampels, M. (eds.) HM 2010. LNCS, vol. 6373, pp. 62–72. Springer, Heidelberg (2010). doi:10.1007/978-3-642-16054-7_5
18. Cola, C., Valean, H.: Ambient recognition using BigData model and Arduino controller. In: Proceedings of 2016 IEEE International Conference on Automation, Quality and Testing, Robotics THETA 20th edition, 19–21 May, Cluj-Napoca, Romania (2016). ISBN: 978-1-4673-8691-13
19. Coltri, A.: Databases in health care. In: Lehman, H.P., Abbott, P.A., Roderer, N.K., et al. (eds.) Aspects of Electronic Health Record Systems, 2nd edn, pp. 225–251. Springer, New York (2006). (Chap. 11)

Reliable Attribute Selection Based on Random Forest (RASER)

Aboudi Noura[1(✉)], Hechmi Shili[1], and Lotfi Ben Romdhane[2]

[1] Department of Computer Science, University of Monastir, Monastir, Tunisia
aboudinoura89@gmail.com, shilihechmi@gmail.com
[2] Department of Computer Science, University of Sousse, Sousse, Tunisia
lotfi.ben.romdhane@gmail.com

Abstract. Feature selection has become one of the most active research areas in the field of data mining. It allows removing redundant and irrelevant data sets of large size. Furthermore, there are several methods in the literature for selecting attributes. In this article, a new multi-objective method is proposed to select relevant and non-redundant features. Our proposed feature selection method is divided into three stages: The first step computes the feature relevance value based on random forests. The second step, computes the dissimilarity matrix representing the dependence between the features of our training datasets, and transform it into a complete graph whose nodes represent features and edges represent the values of dissimilarities between them. The last step is for the optimization in which a multi-objective optimization algorithm is applied. The proposed method is applied on many datasets to find the most relevant and non-redundant features and the performance of the proposed method is compared with that of the popular MBEGA, mRMR (MIQ) and mRMR (MID).

Keywords: Feature selection · Feature relevance · Feature redundance · Multi-objective optimization · Random forest

1 Introduction

Feature selection is a problem that has to be addressed in many areas. The resolution of general problems is based on the treatment of features [1]. So the performance of the treatment system depends on the correct selection of features: considering many features increases the risk of considering redundant and irrelevant features. Therefore, it is necessary to use a method for reducing the data size. The feature selection identify a subset of attributes of minimal size necessary and sufficient to define the target concept [2]. Generally, a typical feature selection method illustrated in Fig. 1, consists of four components: a subset generation or search procedure, an evaluation function, a stopping criterion, and a validation procedure [3]. The selection of attributes will be based on two criteria: relevance and redundancy.

Definition 1 Relevance: In the literature, there are several definitions of the concept of relevance of a feature, the best known is that of [4, 5]. A feature fi is said relevant if its absence reduces significantly the performance of the used classification system.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_2

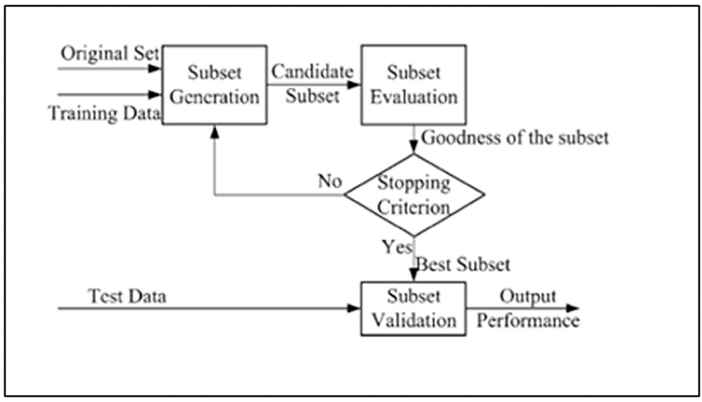


Fig. 1. General feature selection process [3]

Definition 2 Redundancy: The notion of redundancy is relatively associated with correlation function. So, we say that two attributes are redundant if their values are completely correlated [6]. The concept of redundancy will be defined with the Markov blanket [7].

If G be a set of attributes, an attribute is redundant and can be removed from G if it is not relevant and Markov blanket M in G. According to the property of Markov blanket, it is easy to see that a redundant attribute remains removed redundant when other attributes are removed.

2 Related Work

In the literature, feature selection methods are generally grouped into three categories: filter, wrapper and embedded approaches [8]. The filter approach uses statistical measures calculated to filter characteristics according to the number of criteria. This step is generally carried out before applying any classification algorithm [8]. In filter approach, there are several methods such as minimal and redundancy-maximal-relevance criterion (mRMR) [9]. The mRMR is a screening method for the selection of attributes who have the best interests of the target class, are a little redundant and most dissimilar to each other. This method is based on statistical measures such as mutual information, correlation criteria, etc. The two optimization criteria (maximum relevance (MR) and minimum redundancy (mR)) are based on mutual information [10]. There are two types of mRMR that is different depending on the combination of the two criteria(addition and : mRMR(MID) and mRMR(MIQ).

$$\mathrm{mRMR(MID)} = \max(\mathrm{Pertinence} - \mathrm{Redundancy}) \tag{1}$$

$$\mathrm{mRMR(MID)} = \max(\mathrm{Pertinence} - \mathrm{Redundancy}) \tag{2}$$

Besides, there are many other methods such as: Correlation based on feature selection (CFS), Markov blanket filter (MBF)…

The wrapper methods on the contrary, use induction algorithm to evaluate the candidate feature subsets. They generally select feature subsets more suitable for the induction algorithm than the filter methods. In wrapper approach, there are several methods such as:Markov blanket-embedded Genetic Algorithm(MBEGA) [3] provided by Zexuan and Al in the selection of genes. MBEGA used Markov blanket to narrow the search by adding some relevant attributes or removing redundant and/or irrelevant attributes in the solutions selected by genetic algorithm (GA).

The embedded feature selection method, similar to wrapper methods and the feature selection is linked to the classification stage. The embedded methods have been proposed to reduce the classification of learning. They try to combine the advantages of both previous methods. The learning algorithm takes advantage of its own variable selection algorithm. See for example: PLSRFE [11], RF-RFE [12].

Most of the traditional feature selection methods define some metrics to evaluate each individual feature, such as signal-to-noise ratio (SNR) [6] and information gain [13]. They also use many statistical hypotheses testing techniques such as parametric t-test, F-test [11]. Since evaluating 2^N subsets becomes an NP-hard problem, suboptimal subsets are found by using search algorithms that find a subset heuristically. A number of search algorithms can be used to find a subset of variables that maximizes the objective function which is the classification performance [14].

3 Reliable Attribute Selection Based on Random Forest (RASER)

The feature selection is applied to reduce the number of attributes in many applications. Most features selection methods mainly focus on finding the relevant features [15]. We show that the notion of relevance is not sufficient for effective selection of high-dimensional data because of the high correlation that may exist between different features.

In many problems, there exist different aspects of solutions which are partially or wholly in conflict. Therefore, treating these problems as single objective optimization produces an unreliable result. In multi-objective optimization problem the objectives may estimate those different aspects of solutions which are conflicting in nature. After defining the problem of multi-objective optimization, meta heuristics are designed to solve the multi-objective problems. The goal of a multi-objective optimization is to identify a set of solutions in the Pareto optimal set [6].

This section is devoted to the explanation of our contribution, which is to propose a new method for the selection of relevant non-redundant features on the basis of the random forests method. Our method is divided into three steps. The approach followed is recapitulated in Fig. 2.

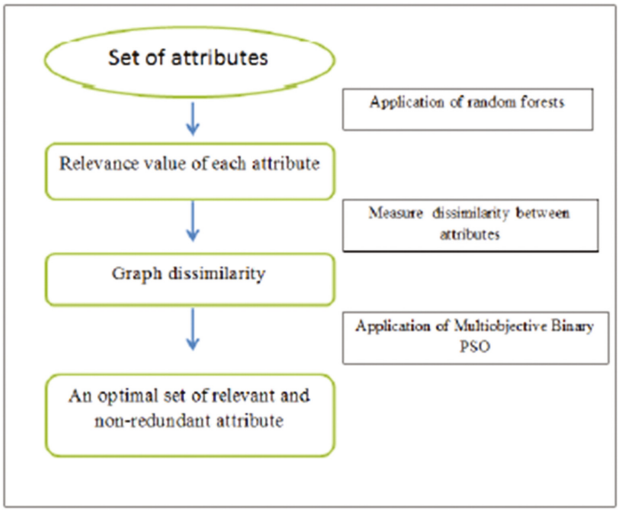


Fig. 2. The steps of our model

First, we compute the relevance values of each feature using random forests. In the second step, we compute the dissimilarity matrix from our training set by using the correlation criterion. Then, our problem is represented by a complete graph whose nodes represent the attributes and the edges represent the value of dissimilarity between them.

The last step is the optimization step in which we apply a multi-objective optimization algorithm (relevance, redundancy) on the dissimilarity graph to obtain an optimal subgraph embedding the most relevant non-redundant features.

3.1 Measuring Features Relevance

Random forests do not require the reduction of the prediction space before classification. In addition, random forests measure the relevance of features for each predictor candidate [16]. This study examines the effectiveness of variable importance measures by random forests in identifying the true indicator of a large number of candidate predictors. Random forests provide an original way for calculating the relevance value of feature. Random forest consists of a number of decision trees. Every node in the decision trees is a condition on a single feature. So, in a random forest or each node of the tree represents an atttribute. Random forests handle the problem of relevance but not the problem of redundancy. Therefore, we will use the graph dissimilarity to handle features redundancy. Algorithm 1 is used for computing of the relevance value [16].

Algorithm 1. Measure of relevance value using random forests

Input: $X_0 = [x_1, x_2, x_3, ..., x_k]^T$:training samples
$Y = [y_1, y_2, y_3, ..., y_k]^T$: classes labels
Output:list of attribute relevance values R
Begin
  initialization:$N_i = 0$,$M_i = 0$,$M_i^j = 0$,
  N_i= Number of times when the observation i appears in the OOB samples.
  M_i=Number of times when the observation i appears in the OOB samples, and misfiled.
  M_i^j= Number of times when the observation i appears in the OOB samples, and misfiled after permutation of values of the variable j in OOB.
  For $i = 1..n$ **do**
    For $j = 1..p$ **do**
      For $k = 1..K$ **do**
        IF *(observation i is in OOB_k)* **THEN**
          $N_i = N_i + 1$
        End IF
        IF *(observation i is in OOB_k and misclassified)* **THEN**
          $M_i = M_i + 1$
        End IF
        Swap the values of the random variable j in OOB_k.
        IF *(observation i is in OOB_k and misfiled after swapping the variable j values in OOB)* **THEN**
          $M_i^j = M_i^j + 1$
        End IF
      End For

    End For

  End For
  the relevance of the j variable = $\frac{1}{n}\sum_{i=1}^{n} Z_i(j)$ where $Z_i(j) = (M_i^j - M_i)/N_i$
End

3.2 Measuring Redundancy

The feature selection process is represented by the graphs, whether G = (V; E) a graph constructed by a set of nodes V and a set of edges E⊆V × V; and W is a weight matrix whose values are in the range [0:1]. Each node of the graph represents an attribute and each edge between two attributes represents the relationship between them.

For a set A of N attributes where $A = \{a_1; a_2; a_3; \ldots; a_N\}$, the arrangement of the training set may be considered as a two-dimensional matrix where the columns represent the attributes and rows represent instances attributes.

There are a variety of methods for measuring of similarity dissimilarity between the attributes such as: Euclidean distance, correlation coefficient [6], etc. By using one of these dissimilarity measures a symmetric matrix is generated called dissimilarity matrix. Therefore if we have a set of N attributes $A = \{a_1; a_2; a_3; \ldots; a_N\}$, computation of dissimilarity between attributes is handled by a S_m.

The dissimilarity matrix is computed from the data matrix by using the correlation coefficient [1] between each pair of attributes. The correlation coefficient σ between two random attributes x and y is defined by:

$$\sigma(x, y) = \frac{cov(x, y)}{\sqrt{var(x)var(y)}} \tag{3}$$

Where var(.) denotes the variance of a variable and cov(.) the covariance between two variables. If x and y are completely correlated (i.e., there is a linear dependency), then σ(x, y) is 1 or −1. If they are totally uncorrelated then σ (x; y) is 0. Hence Eq. (4) represents the dissimilarity between x and y.

$$(1 - |\sigma(x, y)|) \tag{4}$$

Subsequently, according to the dissimilarity matrix a weighted complete graph G is formed. The value at row i and column j in the dissimilarity matrix S_m, represents the weight of the edge between node a_i and a_j. As each feature has some dissimilarity value with other features (present in dissimilarity symmetric matrix S_m), hence the graph G is a complete graph.

Now our problem is how to determine the optimal number of relevant and non-redundant attributes. So, we have to select an optimal subgraph g of G that contains the most relevant attributes and non-redundant. Therefore, we will apply the algorithm Multi-objective Binary PSO based Approach [6] as explained subsequently.

3.3 Subgraph Computation

The problem of finding the most relevant and non-redundant attributes is modeled by finding the optimal subgraph from a weighted undirected graph. The attributes contained by the extracted subgraph comprise the final selected relevant and non-redundant features.

Figure 3 is an example, that the left graph is the initial graph that contains 5 nodes and value of *sij* is the dissimilarity value between attributes and each attribute has a relevance value *Ri*.

The sub-graph on the right is the returned after the application of the optimization algorithm.

The problem of finding the subgraph is an NP-hard problem [1]. To solve our problem we will use multi-objective optimization algorithm graph based MbPSO [6], which is described by the Algorithm 2. In multi-objective optimization problem the objectives may estimate those different aspects of solutions which are conflicting in nature [6].

Algorithm 2. Graph based MObPSO

Input: data matrix dt ,C=number of attributes, seuil=0.9.
Output: Archive A
Begin
  $[x_n, v_n, G_n, P_n]_{n=1}^{N}$=initialize(dt) ◁ Random locations and velocities
  $g_n[V1, E1, VW1, EW1]$=$Sm(V1_n^N \times V1_n^N)$ ◁ subgraphs g_n for N particles are formed from dissimilarity matrix S_m
  $f1 = (1 - \frac{\sum_{i=1}^{|V1|}\sum_{j=1}^{|V1|} EW1_{ij}}{\frac{|V1|(|V1|-1)}{2}})^N$ ◁ average dissimilarity value for the N subgraphs
  $f2 = (1 - \frac{\sum_{i=1}^{|V1|} VW1_i}{|V1|})^N$ ◁ average snr value for the nodes contained by N subgraphs
  IF $fitness(x_n) \not\succ fitness()u, \forall u \in A$ THEN
    A : =x_n ◁ Initialize archive A by first non-dominated x_n
  End IF
  For $n = 1..N$ do
    For $d = 1..C$ do
      v_{nd}:=$w.v_{nd}$+$r1.(P_{nd}-x_{nd})$+$r2.(G_{nd}-x_{nd})$
      $x_{nd} := x_{nd} + v_{nd}$
      IF $x_{nd} \geq seuil$ THEN
        $x_{nd} := 1$ ◁ discretize the cell value
      End IF
    End For
  End For
  For $n = 1..N$ do
    $g_n[V1, E1, VW1, EW1] = Sm(V1_n \times V1_n)$ ◁ new subgraph produced by the evaluated particles
    $f1_n = \frac{\sum_{i=1}^{|V1|}\sum_{j=1}^{|V1|} EW1_{ij}}{\frac{|V1|(|V1|-1)}{2}}$ ◁ average dissimilarity value for the new subgraph
    $f2_n = \frac{\sum_{i=1}^{|V1|} VW1_i}{|V1|}$ ◁ average snr value the nodes contained by the new subgraph
    $A := A \cup x_n$
    For $k = 1..N$ do
      IF $fitness(x_n k) \not\succ fitness(P_n)$ THEN
        $P_n := x_n$ ◁ Update personal best
      IF *Non-dominated fitness* THEN
        $P_n := Random - choice[x_n, P_n]$
      End IF
      End IF
      $G_n := random - select(A)$
    End For
  End For
  $A := x_n(Si(fitness(x_n k) \not\succ fitness(P_n), \forall u \in A)$ ◁ Non-dominated sorting is applied to the updated archive
  $Crowdingsort(A)$ ◁ crowding distance sorting for archive From step-6 to step-33 are repeated according to number of iteration
End

The Algorithm 3 describes the different steps of our proposed new method cited above.

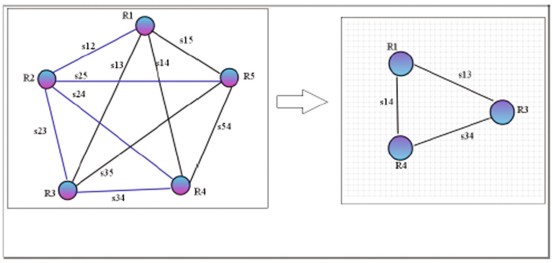


Fig. 3. Example of subgraph extration

Algorithm 3. Reliable Attribute Slection based on Random forest(RASER)

Input: $B = \{X_1, X_2, X_3, ..., X_n\}$: datasets
$Y = \{y_1, y_2, y_3, ..., y_k\}$: classes labels
seuil=0.9
Output: S:set of features
Begin
  $R \leftarrow$ Random Forest(AT, B) ◁ for each attribute , the degree of relevance is calculated using the Alogithme 1.
  For $i = 1..n$ **do**
    For $j = 1..n$ **do**
      $M_d(a_i, a_j) < -\sigma(a_i, a_j)$ ◁ computes the dissimilarity matrix M_d using the proximity measure of the equation (4).
    End For
  End For
  $G \leftarrow generationGraph(R, M_d)$ ◁ generation Graph This is a function that generates a weighted graph.
  $g \leftarrow$ MbPSO$(B, X_t, Y, seuil, G)$
End

4 Experimental Results

In this section, we first describe the datasets, their preprocessing procedure and the performance metrics followed by the results of different algorithms.

In this article nine datasets are used which are publicly available from [1]. Table 1 summarizes these datasets. The first column reports the datasets name, the second column the number of features and the third column the number of instances. The final column reports the number of classes as well as the size of each class. For example, the WBC dataset is composed of 699 instances described each by 10 features. It is composed of two classes (458 Benign, 241 Malignant).

Performance is evaluated using sensitivity, specificity, accuracy and error-rate.

Table 1. Description of the used datasets

Dataset name	# features	# instances	# classes
WBC	10	699	2 (458 Benign, 241 Malignant)
WPBC	34	198	2 (151 non-récurrent, 47 récurrent)
WDBC	32	569	2 (357 benign, 212 malignant)
Spambase	57	4601	2 (spam, not spam)
Ionosphere	34	351	2 (good, bad)
Madelon	500	4400	2
Dermatology	33	366	6 (112 psoriasis, 61 seboreic dermatitis, 72 lichen planus, 49 pityriasis rosea, 52 cronic dermatitis, 20 pityriasis rubra pilaris)

$$\text{Sensitivity} = \frac{TP}{TP + \text{F}N} \tag{5}$$

$$\text{Specificity} = \frac{T\text{N}}{T\text{N} + \text{FP}} \tag{6}$$

$$\text{Errorrate} = \frac{\text{FP} + \text{FN}}{T\text{N} + \text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{T\text{P} + \text{TN} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

Where

TP: True Positive = correctly identified
TN: True Negative = correctly rejected
FP: False Positive = incorrectly identified
FN: False Negative = incorrectly rejected

As this evaluate our method, it is necessary to use a datasets that were not used for learning: this is what is called the test datasets.

It contains also labeled examples to compare the predictions of a hypothesis with the actual value of the class. This test is generally achieved by reserving part of the initial supervised examples and that will not be used in the learning phase. Cross-validation is a method of estimating reliability of a model based on a sampling technique.

It is based on estimated performance from examples that were not used in the design of the model. The cross-validation algorithm k-fold cut the initial set of D examples in k blocks. The cross-validation algorithm is described in [1].

The performance of our proposed method RASER is compared with: MBEGA [3], mRMR(MID) [9], mRMR(MIQ) [9] methods. The datasets are arbitrarily divided into two groups: all together for learning and for testing. Each algorithm is executed for each training set and evaluated by the appropriate test portion.

The performance analysis is extended using 10-fold cross validation. The all algorithms are executed on the total sample versus datasets and the output features are validated using 10-fold cross-validation using Support Vector Machine (SVM).

4.1 Value of Relevance

To compute the features relevance value, we applied the random forest method that contains *ntree* trees. The Fig. 4 shows the influence on the choice of ntree value relevance. We can see in Fig. 4 that the number of trees *(ntree)* has practically no effect on the relevance.

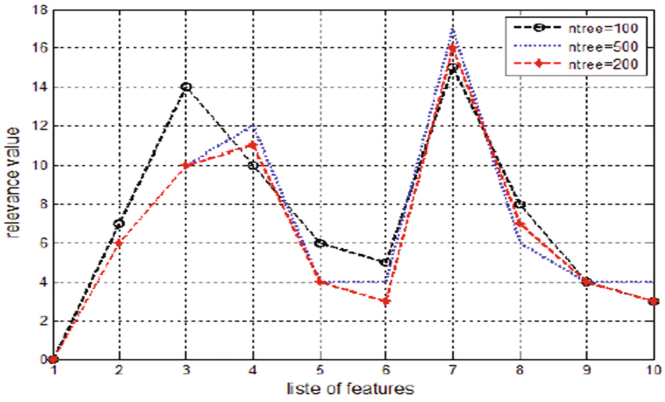


Fig. 4. Influence of *ntree* on value of relevance for WBC datasets

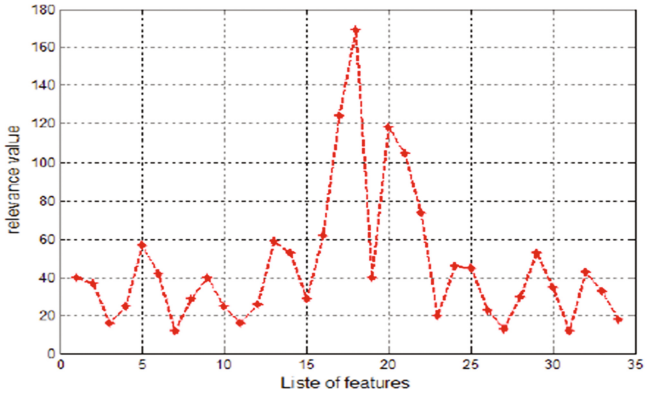


Fig. 5. Attributes relevance value of WPBC datasets

So for our work, we choose 200 as *ntree* value, that is, each built forest contains 200 trees. Figure 5 show the relevance of attribute values for the WPBC dataset.

4.2 Performance

In Table 2 we compare the classification results obtained with the original dataset that contains all the attributes and the dataset that contains the attributes selected by our method. We note that our algorithm, in general returns a classification rate higher than that returned by the classification of the original dataset.

Table 2. Comparisons of the classifications results

Dataset	Methods	Specificity (%)	**Sensitivity (%)**	% accuracy	% error rate
WBC	RASER	98	97.3	**97.6**	2.4
	Original classification	97	96.1	96.4	3.6
BC	RASER	88.1	96.1	**93.4**	6.5
	Original classification	55.3	78.1	76	24
WPBC	RASER	97.6	95.5	96.2	3.7
	Original classification	95.7	96.9	**96.5**	3.4
WDBC	RASER	95.5	98.3	**97.2**	2.7
	Original classification	93.3	96.5	95.3	4.7
Spambase	RASER	34	98	**74**	26
	Original classification	1	100	61	34
Ionosphere	RASER	96	97	**96**	3
	Original classification	52	66	64	36
Dermatology	RASER	99	100	**95**	5
	Original classification	100	98	89	10
Madelon	RASER	51	46	**50**	50
	Original classification	50	47	48	52

Table 3. Comparisons of the classifications results with other feature selection algorithms

Dataset	Methods	Specificity (%)	Sensitivity (%)	% accuracy	% error rate
WBC	RASER	98	97.3	**97.6**	2.4
	MBEGA	95.9	97.6	96.9	3.1
	mRMR(MID)	92	91.3	91.5	8.4
	mRMR(MIQ)	90.8	96.7	94.6	5.3
BC	RASER	88.1	96.1	**93.4**	6.5
	MBEGA	45.5	82.3	71.3	28.7
	mRMR(MID)	57.3	73.8	69.8	30.1
	mRMR(MIQ)	65.5	70.1	70.2	29.7
WPBC	RASER	97.6	95.5	**96.2**	3.7
	MBEGA	68.2	73.2	72.7	27.2
	mRMR(MID)	60.8	72.9	71.1	28.9
	mRMR(MIQ)	66.9	75.1	73.8	26.1
WDBC	RASER	95.5	98.3	**97.2**	2.7
	MBEGA	90.9	98.8	95.7	4.3
	mRMR(MID)	92.7	95.1	94.1	5.8
	mRMR(MIQ)	92.2	96.3	94.7	5.2
Spambase	RASER	34	98	**74**	26
	MBEGA	10	99	64	35
	mRMR(MID)	65	35	61	38
	mRMR(MIQ)	66	38	62	37
Ionosphere	RASER	96	97	**96**	3
	MBEGA	81	21	73	27
	mRMR(MID)	65	38	61	38
	mRMR(MIQ)	47	24	75	84
Dermatology	RASER	99	100	**95**	5
	MBEGA	100	97	90	10
	mRMR(MID)	97	96	81	18
	mRMR(MIQ)	98	96	93	6
Madelon	RASER	51	46	50	50
	MBEGA	38	57	**54**	46
	mRMR(MID)	5	54	0.53	47
	mRMR(MIQ)	49	51	50	50

The cross-validation scores of different algorithms are reported in Table 3. It is clear from this table for the most of datasets the sensitivity, specificity, accuracy, and errorrate values of our model are better than mRMR(MID) mRMR(MIQ) and MBEGA. The average Accuracy for our method is very much upper than other methods i.e. proposed method results are more non-redundant and relevant features than other comparative methods.

5 Conclusion

In this paper, we have proposed a novel feature selection algorithm based on random forest to select the most relevant and non-redundant one. In that purpose, we have used the random forest method to measure the relevance value attributes and the correlation coefficient to calculate the value of redundancy. The proposed method is performed in three basic steps three stages: the measure of relevance of attributes, modeling and optimization. Finally, we tested our approach for the selection of attributes on datasets and we have approved its effectiveness. These encouraging results should stimulate future research. Our immediate concern is to use another measures of dissimilarity criterion.

References

1. Samb, M.L., Camara, F., Ndiaye, S., Slimani, Y., Esseghir, M.A.: Approche de sélection d'attributs pour la classification basée sur l'algorithme rfe-svm
2. Chouaib, H.: Sélection de caractéristiques:méthodes et applications (2011). http://www.math-info.univ-paris5.fr/∼vincent/siten/Publications/theses/pdf/chouaib.pdf
3. Zhu, Z., Ong, Y.-S., Dash, M.: Markov blanket-embedded genetic algorithm for gene selection. Pattern Recogn. **40**, 3236–3248 (2007). http://www.sciencedirect.com/science/article/pii/S0031320307000945
4. John, G.H.: Enhancements to the data mining process. Doctoral dissertation, Ph.D. thesis of Stanford University (1997)
5. Kohavi, R., Pfleger, K., John, G.H.: Irrelevant features and the subset selection problem, pp. 121–129 (1994)
6. Mandal, M., Mukhopadhyay, A.: A graph-theoretic approach for identifying non-redundant and relevant gene markers from microarray data using multiobjective binary PSO. PLoS ONE **9**(3), e90949 (2014)
7. Koller, D., Sahami, M.: Toward Optimal Feature Selection. pp. 284–292. Stanford InfoLab, Stanford (1996)
8. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19), 2507–2517 (2007)
9. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of maxdependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
10. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Netw. **5**, 537–550 (1994)
11. You, W., Yang, Z., Ji, G.: PLS-based recursive feature elimination for high-dimensional small sample. Knowl.-Based Syst. **55**, 15–28 (2014)
12. Zhou, Q., Zhou, H., Zhou, Q., Yang, F., Luo, L.: Structure damage detection based on random forest recursive feature elimination. Mech. Syst. Sig. Process. **46**(1), 82–90 (2014)
13. Azhagusundari, B., Thanamani, A.S.: Feature selection based on information gain. Int. J. Innov. Technol. Explor. Eng. (IJITEE) ISSN 2278–3075 (2013)
14. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**(1), 16–28 (2014)
15. Yu, L., Liu, H., Feature selection for high-dimensional data: a fast correlation-based filter solution. In: ICML, vol. 3, pp. 856–863 (2003)

16. Ghattas, B., Ishak, A.B.: Sélection de variables pour la classification binaire en grande dimension: comparaisons et application aux données de biopuces. J. de la société française de statistique **149**(3), 43–66 (2008)
17. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genomewide expression patterns. Proc. Natl. Acad. Sci. **95**(25), 14863–14868 (1998)
18. Crescenzi, P., Kann, V., Halldórsson, M.: A compendium of NP optimization problems (1995)
19. https://archive.ics.uci.edu/ml/datasets.html

Estimating the Number of Clusters as a Pre-processing Step to Unsupervised Learning

Paulo Rogerio Nietto[1(✉)] and Maria do Carmo Nicoletti[1,2]

[1] Faculdade Campo Limpo Paulista (FACCAMP), Campo Limpo Paulista, SP, Brazil
{pnietto,carmo}@cc.faccamp.br
[2] Universidade Federal de São Carlos (UFSCar), São Carlos, SP, Brazil

Abstract. A great challenge in machine learning, as far as unsupervised algorithms are concerned, is to devise methods for pre-estimating the number of clusters associated to a given set of patterns to be clustered. By doing so and by using the number of clusters as input to clustering algorithms that require the information, the chances of getting better results increase substantially. The work described in this paper investigates the performance of an algorithm, based on the sequential clustering BSAS (*Basic Sequential Algorithmic Scheme*), to produce an ordered list (by frequency of occurrences), containing good estimates for the number of clusters in a given set of patterns. The BSAS is a convenient choice since the order in which patterns are presented to the algorithm can impact the induced clustering. The results of the experiments in eight sets of patterns can be considered empirical evidence that the procedure can be a practical and reliable option, as a pre-processing step, to using clustering algorithms that require the number of clusters.

Keywords: Clustering · Estimation of the number of clusters · Sequential clustering algorithms

1 Introduction

Machine learning (ML) algorithms that deal with data that do not have an associated class are known as unsupervised learning algorithms. Clustering algorithms constitute a group of unsupervised algorithms which, currently, can be considered the most popular among the unsupervised learning algorithms. In a simplistic way it can be said that the main goal of a clustering algorithm is to partition a set of patterns into groups (clusters) of patterns, so those that share the same cluster are similar to each other and those belonging to different clusters are not that similar. It can be found in the literature several clustering algorithms based on a variety of mathematical and statistical formalisms, such as in [1–4]; several taxonomies [5–8], associated to these algorithms, can also be found. Given a set of patterns, a reasonable number of clustering algorithms require, also, as an input parameter, the number (K) of clusters the induced clustering should have. This, somehow, makes the quality of the resulting clustering be heavily dependent on the value of K. In some applications the value of K can be estimated by human experts who have

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_3

experience and deep knowledge about the domain of the data; however, in most cases, a suitable value for K is unknown and must be estimated based only on the patterns themselves. A short description suggesting a procedure to estimate the number of clusters (K) in a set of patterns, based on the patterns themselves, was proposed in [6]. The procedure involves the use of clustering algorithms which do not require the information about the number of clusters. One of such algorithms is the *Basic Sequential Algorithmic Scheme* (BSAS) [6], which was adopted in the work described in this paper, to support the proposal of an algorithm, the *ClusterEstimate*, for determining the number K of clusters prior to a clustering task, aiming at directing it. The paper is organized as follows. Section 2 presents the *ClusterEstimate* algorithm and its main procedures. Section 3 gives some insights of a customized version of the algorithm BSAS, to provide technical material for the understanding of its use, by the *ClusterEstimate* algorithm. The eight sets of patterns used in the experiments aimed at evaluating the proposal are presented in Sect. 4 and the experiments using the *ClusterEstimate* and results are presented and discussed. Section 6 resumes the work done.

2 Estimating the Number of Clusters of a Clustering by Using a Sequential Clustering Algorithm

This section focuses on a brief textual description, suggested in [6], for determining the number of clusters associated with a given set of patterns, which has been translated as the *ClusterEstimate* algorithm (pseudocode in Fig. 1).

```
procedure ClusterEstimate(X,S,P)
Input: X = {P1, P2,...,PN} % dataset with N patterns
       S % number of times to run the BSAS to a specific Θ
       P % iterative process step
begin
  min ← minimum d(Pi,Pj)
  max ← maximum d(Pi,Pj)
  NC ← {}   % vector of number of clusters
  for Θ = min to max step P
    begin
      NS ← {}  % number of clusters for the current Θ
      for run =1 to S do
        begin
          NS ← NS ∪{BSAS_NC(X,Θ,CQ)}
         Shake(X,X)
        end
      NC ← NC ∪ mostFrequentValue(NS)
    end
  plot(NC)
end
```

Fig. 1. High level pseudocode of the *ClusterEstimate* algorithm.

The *ClusterEstimate* algorithm is based on the BSAS algorithm and employs the Euclidean distance as dissimilarity measure. Notice that the goal of the algorithm is to estimate the number of clusters that a clustering of a given set of patterns should have. With respect to the *ClusterEstimate* procedure, and considering X to be the set of patterns to be clustered, (1) variables *min* and *max* contain the lowest and highest dissimilarity values, respectively; (2) the value of the iterative process step should be informed by the user; (3) the value of the *S* parameter relates to the precision of the desired results; the greater the *S* value is, the greater is the precision of results; (4) procedure *BSAS_NC*(X, Θ, CQ) returns the number of clusters generated by the *BSAS_NC* algorithm, described in Fig. 2, using a dissimilarity threshold value of Θ; (5) the *Shake*(X,X) procedure randomly changes the order of patterns in the set X; (6) *mostFrequentValue*(NS) returns the numbers of clusters that are the most frequent, considering the *S* clusterings generated by *BSAS_NC,* with a determined threshold and (7) *plot*(NC) simply creates a plotting graph of the number of clusters *versus* threshold values – usually the plotting has a certain number of wide flat regions.

```
procedure BSAS_NC(X,Θ,CQ)
Input: X = {P1, P2,...,PN} % dataset with N patterns
       Θ                   % dissimilarity threshold
Output: CQ  % Number of clusters
begin
CQ ← 1
CCQ ← {P1}
for i=2 to N do
 find Gk: d(Pi,Gk) = min1≤j≤CQ d(Pi,Gj)
 if d(Pi,Gk) > Θ then
    CQ ← CQ +1
    GCQ ← {Pi}
  else
     Gk ← Gk ∪ {Pi}
    updateCentroid(Gk) %update the centroid of Gk
end
return CQ
end
```

Fig. 2. High level pseudocode of the *BSAS_NC* algorithm.

3 The BSAS Algorithm

BSAS is a clustering algorithm characterized as sequential, where data patterns are presented to algorithm only once or twice; the algorithm does not require a value for K. The BSAS has been employed in a few works such as [9, 10] and its customized version, *BSAS_NC* (*Basic Sequential Algorithmic Scheme for Number of Clusters)* is shown in Fig. 2.

Each new pattern from a given set of patterns is processed by *BSAS_NC* either by assigning it to an existing cluster or, then, by defining a new cluster containing it (a singleton

so far), depending on the dissimilarity measure between the pattern and the centroids of the clusters already constructed. The order in which patterns are presented has an important role when the algorithm is constructing the clustering. Different presentation orders may result in totally different clusterings, in terms of the number of clusters as well as in the patterns that belong to each of them. The original *BSAS* has two parameters: (1) the dissimilarity threshold (Θ) and (2) the maximum number of clusters allowed (q). As *ClusterEstimate* uses BSAS to help determining the number of clusters in a given set of patterns, the original algorithm was modified and turned into the *BSAS_NC*, by removing the need for the q parameter. Unlike proper clustering algorithms, the main role of the *BSAS_NC* version is only to output the number of clusters induced by the algorithm. An important issue that affects the results of the algorithm is the choice of the dissimilarity threshold value (Θ). If the value is too low, unnecessary clusters will be created and, if the value is too high, only a small number of suitable clusters will be created.

4 Data Domains for the Experiments

To investigate the performance of *ClusterEstimate*, a collection of sets of two-dimensional synthetic patterns, with different shapes, densities and quantity of patterns was used. All the eight sets of patterns presented in Fig. 3 were generated based on some typical clustering situations. Particularly, the sets of patterns in Fig. 3(a)–(d) were created based on those used in the research described in [11], that had its focus on the so called *gestalt* clusters. The clustering detection approaches described in the research were motivated by the human perception of two-dimensional sets of patterns as separate groupings or *gestalts*, where the principle of grouping is proximity, as described in [12]. Based on the human perception, each set of patterns exhibited in Fig. 3(a) and (b), can be approached as two distinct clusters of points. Although the set of patterns in Fig. 3(a) and (b) show similarities, the two clusters in each of them have completely different shapes. The set of points in Fig. 3(c) involves

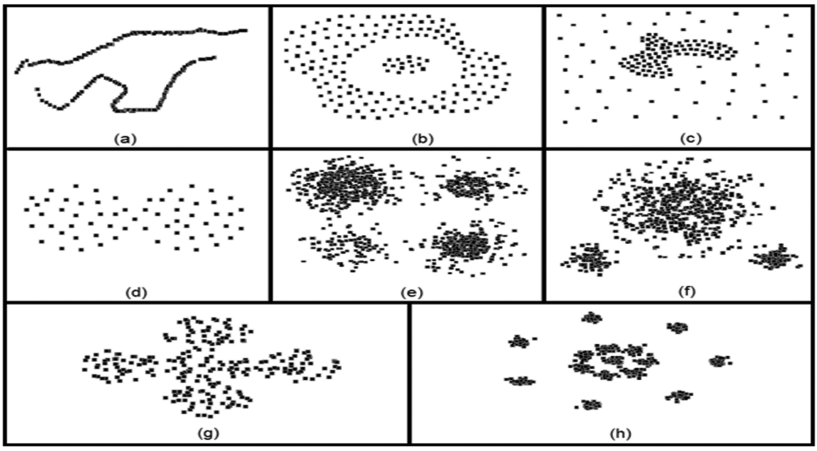


Fig. 3. Collection of eight sets of patterns used in the experiments, having as number of patterns: (a) 200, (b) 174; (c) 142; (d) 63; (e) 1,000; (f) 700; (g) 250 and (h) 600.

sharp gradient detection. Figure 3(d) shows one cluster with a small narrow section (*neck*), whose removal divides it into two distinct clusters. The set of patterns shown in Fig. 3(e), which consists of four clusters with different densities and slightly separated, was created inspired by the set "*Differentdensity*", used in the research work described in [13]. The set "*Skewdistribuition*", also described in [13], inspired the creation of the set of patterns in Fig. 3(f), which can be visually described as consisting of three clusters of patterns, where two of them are small and the third is large.

The set of patterns in Fig. 3(g) was created based on the "*AD_5_2*", used in the clustering experiments described in [14]; sets of patterns similar to (g) are typically used to evaluate algorithms with respect to the overlap areas that clusters may have. The patterns in the central area of (g) can be approached under two different perspectives (1) as a partial continuation of each of the well distinctive four clusters (each petal) or, then (2) as a cluster on its own. If (1) is assumed, the set of patterns has four clusters and if (2) is assumed, the set of patterns has 5 clusters. The set of patterns in Fig. 3(h) was created based on pattern set "*R15*", used in the research work described in [15]; its clusters are spherical and positioned so to compose two rings of clusters, that share a central cluster. The purpose of such disposition of patterns is to verify if clustering algorithms can identify the two clusters, each formed by a ring of small groups of patterns and, also, check how they behave towards the central group of patterns, central to both rings.

5 Experiments, Results and Analysis

The experiments presented in this section were conducted by visual inspection on the plotting graph which is one of the outputs of *ClusterEstimate*. Through visual inspection, the wider flat lines (parallel to the x-axis) in the plotting are compared; each flat line is associated with the number of runs (of *BSAS_NC*) where the same number of clusters has been maintained. The iterative step value (P) used in the experiments is calculated so to correspond to 100 threshold values; each threshold value is used 10 times (*i.e.*, $S = 10$) which, on average, corresponds to 1,000 executions of the *BSAS_NC* procedure. The methodology for conducting each experiment was defined by the sequence of four steps, namely: (1) establishing the iterative process step parameter using Eq. (1); (2) execute the *ClusterEstimate* with the two parameters (2.1) the iterative process step as defined by Eq. (1) and (2.2) ten executions for each threshold value; (3) visually select the five widest flat lines of the resulting plotting created by *ClusterEstimate*, ignoring those associated with one cluster and (4) visually checking if any of the five widest flat lines of the resulting plotting shows the number of clusters clearly identifiable. Consider X be a set of patterns to be clustered. Let MaxDiss and MinDiss be the largest and the smallest dissimilarity value between any two patterns of set X, respectively. Equation (1) gives the value of the iterative process step (P).

$$P = (\text{MaxDiss} + \text{MinDiss})/100 \tag{1}$$

ClusterEstimate has the bias of returning, as result, a single cluster (considering that is the most frequent). This is because (1) the higher the threshold value, the lower will be the number of clusters that will be created and (2) the order in which the patterns are

presented to *BSAS_NC* influences the clustering obtained by the algorithm. When the first pattern presented to *BSAS_NC* has dissimilarity values in relation to other patterns lower than the threshold value, only one cluster is returned. Due to that, this work ignores, in the plotting created by *ClusterEstimate*, those flat lines associated with a single cluster. For some sets of patterns, the *ClusterEstimate* output may result in several flat lines and, some of them may deserve to be inspected further, as other possibilities for alternative numbers of clusters. Therefore, as an empirical methodological decision, the five widest flat lines have been considered throughout all the experiments. In Fig. 4(a) the complete plotting graph of the overall results of *ClusterEstimate,* in the set of patterns of Fig. 3(a), using iterative process step = 0.022, is shown. Figure 4(b) shows an enlargement of the plotting shown in (a), focusing on the wider linear segments parallel to the x-axis.

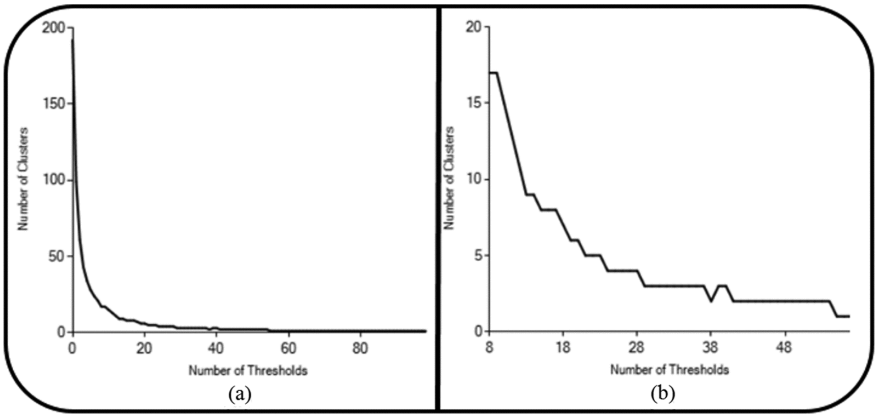


Fig. 4. Plotting graph generated by *ClusterEstimate*, having as input the set of patterns of Fig. 3(a). (a) the complete plotting. (b) enlargement of a sector of (a), focusing on its wider flat lines.

ClusterEstimate returned eight possible numbers of clusters, associated to the set of patterns in Fig. 3(a), as can be identified in Fig. 4(b), by the small linear segments parallel to the x-axis. The five first numbers of clusters, ordered by decreasing frequency are: (1) 2 clusters, (2) 3 clusters, (3) 4 clusters, (4) 5 clusters and (5) 8 clusters. *ClusterEstimate* found, as first option for the number of clusters, 2, which coincides with the number of clusters visually perceived in Fig. 3(a). Due to space restriction, all figures that follow will only show the enlarged plotting graph of the region of interest (wider linear segments parallel to the x-axis) extracted from the full plotting graph result produced by *ClusterEstimate*. Figure 5(a) shows only the enlarged plotting graph of the region of interest when using the set of patterns given in Fig. 3(b), having the iterative process step = 0.127. As can be checked in Fig. 5(a), *ClusterEstimate* found six possible numbers of clusters for the set of patterns in Fig. 3(b). The five first numbers of clusters, (concerning the length of their associated segments), ordered by decreasing frequency are: (1) 2 clusters, (2) 4 clusters, (3) 5 clusters, (4) 3 clusters and (5) 12 clusters. As can

be visually verified in Fig. 3(b), *ClusterEstimate* found, as first element of its output list (*i.e.*, 2), the right number of clusters. Figure 5(b) shows only the enlarged plotting graph of the region of interest when using the set of patterns given in Fig. 3(c), considering the iterative process step = 0.212. As can be checked in Fig. 5(b), the *ClusterEstimate* algorithm also found six possible numbers of clusters for the set of patterns in Fig. 3(c). The five first numbers of clusters (concerning the length of their associated segments), ordered by decreasing frequency are: (1) 5 clusters, (2) 3 clusters, (3) 2 clusters, (4) 9 clusters and (5) 4 clusters. Although the number of clusters that can be visually detected in the given set of patterns is 3, this number appears in the third position in the list of the 5 most suitable numbers of clusters found by *ClusterEstimate*.

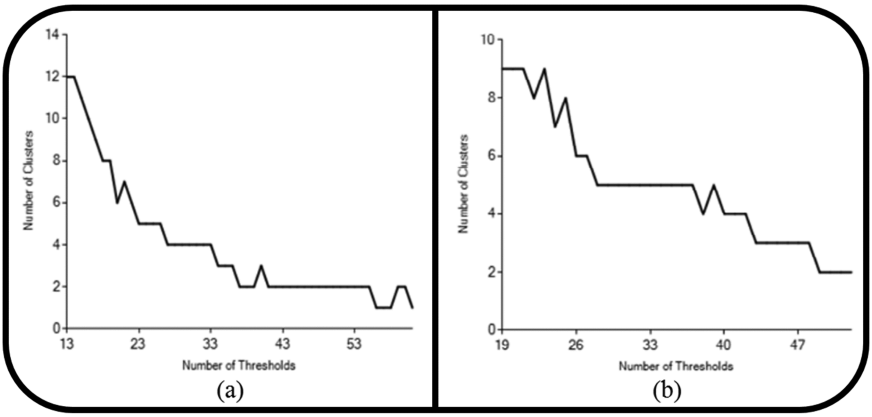


Fig. 5. (a) enlargement of the region of interest of the plotting graph obtained using Fig. 3(b) and (b) enlargement of the region of interest of the plotting graph obtained using Fig. 3(c).

Figure 6(a) shows only the enlarged plotting graph of the region of interest when using the set of patterns given in Fig. 3(d) and the iterative process step = 0.047. *ClusterEstimate* found six possible numbers of clusters associated to the set of patterns of Fig. 3(d), as can be identified in Fig. 6(a). The five first numbers of clusters ordered by decreasing frequency are: (1) 2 clusters, (2) 4 clusters, (3) 5 clusters, (4) 3 clusters and (5) 8 clusters. The first option in the list of numbers of clusters produced by *ClusterEstimate* is 2, which coincides with the number of clusters that can be visually identified in Fig. 3(d). Figure 6(b) shows only the enlarged plotting graph of the region of interest, when using the set of patterns given in Fig. 3(e) and iterative process step = 0.24. The *ClusterEstimate* procedure identified four possible numbers of clusters for the set of patterns shown in Fig. 3(e), which ordered by decreasing frequency are: (1) 4 clusters, (2) 2 clusters, (3) 3 clusters and (4) 5 clusters. As can be visually seen in Fig. 3(e), the *ClusterEstimate* procedure has delivered, as first option, the right number of clusters, i.e., 4.

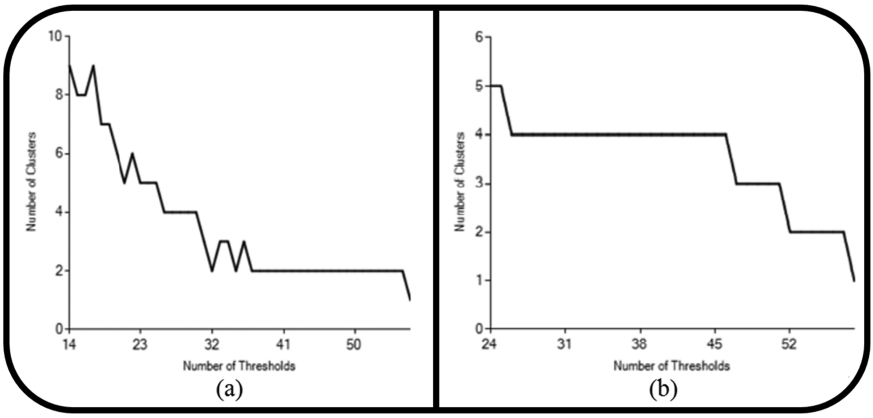


Fig. 6. (a) enlargement of the region of interest of the plotting graph obtained using Fig. 3(d) and (b) enlargement of the region of interest of the plotting graph obtained using Fig. 3(e).

Figure 7(a) shows only the plotting graph of the region of interest when using the set of patterns given in Fig. 3(f) and iterative process step = 0.08. *ClusterEstimate* found eleven possible numbers of clusters for the set of patterns in Fig. 3(f) and the five first numbers of clusters, ordered by decreasing frequency are: (1) 3 clusters, (2) 2 clusters, (3) 5 clusters, (4) 6 clusters and (5) 7 clusters. As can be confirmed via visual inspection of the set of patterns in Fig. 3(f), *ClusterEstimate* found, as its first option, the right number of clusters. Figure 7(b) shows only the enlarged plotting graph of the region of interest when using the set of patterns given in Fig. 3(g), having the iterative process step = 0.12. *ClusterEstimate* found eighteen possible numbers of clusters associated to the set of patterns in Fig. 3(g); the five first numbers of clusters ordered by decreasing frequency are: (1) 4 clusters, (2) 3 clusters, (3) 5 clusters, (4) 2 clusters and (5) 6 clusters. As discussed before, in Sect. 4, the set of patterns in Fig. 3(g) can be visually clustered into 4 or 5 clusters, depending on the approach used concerning the set of patterns 'shared' by the well-defined 4 clusters. Both options, though, are part of the 5-best options identified by *ClusterEstimate,* placed in the first and third positions, respectively, in such list. Figure 8 shows the enlarged plotting graph of the region of interest when using the set of patterns given in Fig. 3(h) and iterative process step = 0.14. *ClusterEstimate* found ten possible numbers of clusters associated with the set of patterns in Fig. 3(h); the five first numbers of clusters ordered by decreasing frequency are: (1) 8 clusters, (2) 3 clusters, (3) 2 clusters, (4) 10 clusters and (5) 15 clusters. This experiment is the one (out of 8) where the *ClusterEstimate* was not that entirely successful, because it placed the right number of clusters (15) in the fifth position (last) of its list. In five out of the eight experiments conducted *ClusterEstimate* obtained, as the first option in the list of possible numbers of clusters, the right number of visually identifiable clusters. *ClusterEstimate* has not returned, though, as the first option in its output list, the right number of visually identifiable clusters in the sets of patterns in Fig. 3(c), (g) and (h). All the three sets of patterns have clusters that "meet" each other, suggesting that the *ClusterEstimate* has its performance decreased when that happens. However, for the

three sets of patterns, *ClusterEstimate* managed to deliver the right number of clusters among the first five of its list. As pointed out before, the results obtained with the set of patterns shown in Fig. 3(g), in particular, can be argued to have either four or five clusters, depending on the criteria used for visually identifying them.

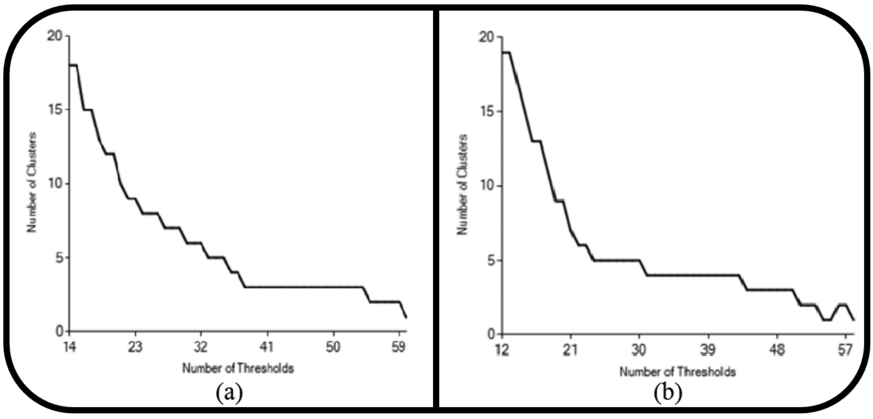


Fig. 7. (a) enlargement of the region of interest of the plotting graph obtained using Fig. 3(f) and (b) enlargement of the region of interest of the plotting graph obtained using Fig. 3(g).

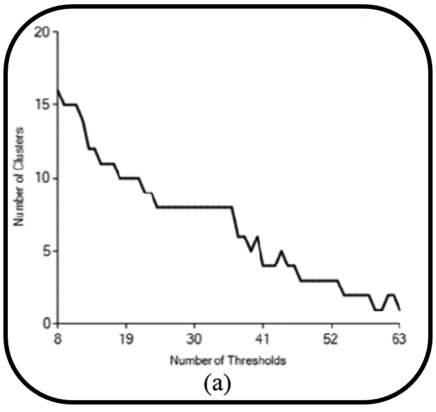


Fig. 8. Enlargement of the region of interest of the plotting graph obtained using Fig. 3(h).

6 Conclusions

This paper discusses and empirically evaluates the performance of an algorithm for estimating the number of clusters in a given set of patterns. The algorithm employs a variation of a sequential clustering algorithm, known as BSAS, which is highly dependent of the order the patterns are processed; in a way, this dependency was explored for detecting a possible number of clusters when, after various runs, the

algorithm always produces a clustering with the same number of clusters. Experiments were conducted using eight synthetics set of patterns, several of them inspired by others available in the literature. The *ClusterEstimate* procedure has detected, as its first option, the same number of clusters as a human inspection would, in five out the eight sets. In three sets of patterns, however, although not listed as first option, the procedure still had the right number among its five-best results.

Acknowledgments. Authors thank CAPES, CNPq and FACCAMP.

References

1. Asano, T., Bhattacharya, B., Keil, M., Yao, F.: Clustering algorithms based on minimum and maximum spanning trees. In: Proceedings. of the Fourth Annual Symposium on Computational Geometry (SCG 1988), pp. 252–257 (1988)
2. Hartuv, E., Shamir, R.: A clustering algorithm based on graph connectivity. Inf. Process. Lett. **76**, 175–181 (2000)
3. Päivinen, N.: Clustering with minimum spanning tree of scale-free structure. Pattern Recogn. **26**, 921–930 (2005)
4. Luxburg, U.: A tutorial on spectral clustering. J. Stat. Comput. **17**, 395–416 (2007)
5. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. **31**(3), 264–323 (1999)
6. Theodorides, S., Kotroumbas, K.: Pattern Recognition, 4th edn. Elsevier, USA (2009)
7. Berkhin, P.: A survey of clustering data mining techniques. In: Kogan, J., Nicholas, C., Teboulle, M. (eds.) Grouping Multidimensional Data, pp. 25–71. Springer, Heidleberg (2006)
8. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**, 645–678 (2005)
9. Nicoletti, M.C., Real E.M., Oliveira, O.L.: The impact of refinement strategies on sequential clustering algorithms. In: Proceedings of the 13th International Conference on Intelligent Systems Design and Applications (ISDA 2013), pp. 47–52 (2013)
10. Real, E.M., Nicoletti, M.C., Oliveira, O.L.: A closer look into sequential clustering algorithms and associated post-processing refinement strategies. Int. J. Innov. Comput. Appl. **6**, 1–12 (2014)
11. Zahn, C.T.: Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Trans. Comput. **C-20**, 68–86 (1971)
12. Wertheimer, M.: Principles of perceptual organization. In: Beardsley, D., Wertheimer, M. (eds.) Readings in Perception. Van Nostrand, Princeton (1958)
13. Liu, Y., Li, Z., Xiong, H., Gao X., Wu, J.: Understanding of internal clustering validation measures. In: Proceedings of the 10th International IEEE Conference on Data Mining (ICMD), pp. 911–916 (2010)
14. Bandyopadhyay, S., Saha, S.: Unsupervised Classification. Springer, Heidelberg (2013)
15. Veenman, C.J., Reinders, M.J.T., Backer, E.: A maximum variance cluster algorithm. IEEE Trans. Pattern Anal. Mach. Learn. **24**, 1273–1280 (2002)

Agglomerative and Divisive Approaches to Unsupervised Learning in *Gestalt* Clusters

Rodrigo C. Camargos[1(✉)], Paulo R. Nietto[1], and Maria do Carmo Nicoletti[1,2]

[1] Faculdade Campo Limpo Paulista (FACCAMP), C. L. Paulista, SP, Brazil
{rodrigocamargos,pnietto,carmo}@cc.faccamp.br
[2] Universidade Federal de São Carlos (UFSCar), S. Carlos, SP, Brazil

Abstract. Hierarchical clustering algorithms can be agglomerative or divisive, depending on how partitions are formed. Such algorithms have advantages mainly related to the desired level of granularity the partition should have. The work described in this paper approaches two hierarchical algorithms, one agglomerative (and three of its variants) and the other divisive, focusing on their performance in unsupervised learning tasks related to *gestalt* clusters. Taking into account that the point sets considered are representative of gestalt clusters, the experiments show that the best results have been obtained when the agglomerative approach was used.

Keywords: Unsupervised machine learning · Agglomerative and divisive approaches · Clustering · *Gestalt* clusters

1 Introduction

Over the past few decades many ideas of how to enable automatic learning have been presented, discussed and implemented. The success and popularity of automatic learning methods are due, mainly, to the so-called inductive learning algorithms. Numerous references address machine learning algorithm revisions and, particularly, inductive machine learning, such as [1–5]. The so-called supervised learning algorithms generalize their training data, generally described as vectors of attribute values and an associated class, into some higher-level type of representation. In many real-world situations, however, the class each training data instance belongs to is unknown. Inductive machine learning algorithms that deal with such training points are known as *unsupervised*; one such group of algorithms is known as *clustering*. As can be confirmed in the literature related to clustering, the number of available algorithms is quite large and have been organized into a diverse range of categories (see, for example, [5–7]). Usually clustering algorithms and related techniques are used when one wants to know about the structure underneath the data and there is no class to be predicted; the classical use of clustering is in situations where the original data set must be partitioned into what is conventionally called 'natural' groups (or *clusters*). It is assumed that data points in one of these clusters reflect an inherent tendency of the data domain, which causes them to be more similar to each other than similar to those in another cluster of the same clustering. Some clustering algorithms instead of inducing a single

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_4

clustering, induce a hierarchical structure of clusters. Given a set X with N data points, hierarchical algorithms construct a series of partitions (*clusterings*) of X, in the same run. So, it can be said that hierarchical algorithms induces nested hierarchic clusterings, usually using one out of two strategies: (1) bottom-up (*agglomerative*): initially each data point defines a singleton cluster and then, pairs of clusters are merged until all points belong to a single unique cluster. So, in an agglomerative approach, the first partition consists of N clusters and the last consists of a single cluster containing all the N data points; (2) top-down (*divisive*): initially all the original data points belong to a unique cluster and splits are performed up to the point that each data point belongs to a unitary cluster. Therefore, the first partition has only one cluster and the last has N single clusters. Both strategies have advantages and disadvantages. The main focus of the work described in this paper was to conduct an empirical research related to clustering algorithms characterized as hierarchical, taking into account several typical data points, referred to as *gestalt* clusters, as described in [8]. The paper focuses on the empirical evaluation of two hierarchical algorithms, one based on the generalized agglomerative scheme described in [5] and, the other, based on the divisive analysis scheme, proposed and described in [4]. For the experiments a collection of data points referred to as *gestalt* clusters, described in [8] was used for evaluating the performance of both schemes, using the K-means algorithm as baseline. The remainder of this paper is organized as follows. Section 2 presents the main characteristics of the agglomerative approach and the pseudocode of the Generalized Agglomerative Scheme (GAS) [5], which has been implemented and used in the experiments. In Sect. 3 the main characteristics of the divisive approach are discussed and the pseudocode of a variant of the divisive algorithm named DIANA [4] is given. Section 4 presents the motivation for using a particular collection of point sets (*gestalt* clusters) in the experiments, commenting about their main characteristics. Section 5 describes the methodology and the experiments conducted, discussing their results and presenting some comparative analysis. Section 6 resumes the work done and presents its main conclusions.

2 The Agglomerative Approach

In a simplistic way, given a set containing N data points to be clustered, agglomerative hierarchical clustering algorithms start with N clusters (each having a single point) and go on by merging them into larger clusters, until a single cluster, containing all the N data points, is obtained. Figure 1 presents a high-level general pseudocode description of a procedure that implements the *generalized agglomerative scheme* (*GAS*) considering, as criteria for joining two clusters, the Euclidean distance between them and the number of cluster (K) as stop criteria. At each step the algorithm merges two clusters which are the closest ones among all. Generally the differences between agglomerative algorithms found in the literature are related to their different ways of defining inter-clusters proximity. Among the several possible ways, three are particularly popular: (1) *Single Linkage*, where the distance between clusters is defined as that of the closest pair of points, where only pairs consisting of one point from each cluster are considered; (2) *Complete Linkage*, where the distance between clusters is defined as that of the most distant pair of points, where only pairs consisting of one point from each cluster are considered and (3) *Average Linkage*,

where the distance between two clusters is the average distance between each point in one cluster to every point in the other cluster. The *Single Linkage* approach is sensitive to the presence of outliers as well as dealing with clusters having substantial differences in density. The *Complete Linkage* is not strongly affected by outliers as much as the *Single Linkage*, but have the tendency of breaking large clusters and do not perform well with convex clusters. The *Average Linkage* approach is sensitive to both, the shape and the size of clusters and tends to join clusters with small variances; therefore, it can easily fail when clusters have sophisticated forms, not conforming to hyper spherical shapes (see [7]).

```
procedure GAS (X,K,AC_t)
Input: X = {P_1, P_2,...,P_N} % dataset with N patterns
       K % number of clusters to be created
Output: AC_t
begin
t ← 0
AC_t ← {{P1}, {P2},...,{PN}} % initial clustering
Nro_C ← N                    %initial number of clusters
CI ← N+1        % cluster index for new created clusters
repeat
  t ← t + 1
  among all possible pairs of clusters {C_r,C_s} in AC_t-1,
  find one {C_i,C_j} such that g(C_i,C_j) = min_r,s g(C_r,C_s),
  where g is a dissimilarity function (distance)
  New_C ← C_i ∪ C_j % create a new cluster
  Nro_C  ← Nro_C - 1
  CCI  ← New_G
  CI ← CI +1
  AC_t  ← (AC_t-1 - {C_i,C_j}) ∪ {C_CI}
until Nro_C ≤ K
end
return AC_t
end
```

Fig. 1. The generalized agglomerative scheme (GAS) procedure, using as dissimilarity function the euclidean distance, loosely based on the one described in [5].

3 The Divisive Approach

As pointed out in [4], divisive algorithms are hierarchical in nature. Given a dataset with N points, divisive hierarchical algorithms start with a single cluster (containing the N points) and go on splitting clusters; typically at each iteration one cluster is divided into two subclusters (when the bisection strategy is used), a process that constructs a binary tree of clusters. However, as pointed out in [7], they can be computationally demanding if all $2^{N-1}-1$ possible divisions into two subclusters of a cluster of N points are considered at each step. In this work, a slightly modified DIANA algorithm [4] was used in the experiments presented in Sect. 5. The original DIANA is based on the proposal described in [9], that assumes to be possible to construct divisive methods that do not consider all

divisions, based on the assumption that most of them are inappropriate. The DIANA algorithm also assumes that some of functions used can be customized, depending on the application at hand. The different 'instantiations' obtained, as a result of the different customization processes implemented, are often considered new algorithms. The algorithm, by being divisive, assumes that the initial point set forms a single cluster. The DIANA algorithm, described in Fig. 2, use the number K of clusters as stop criteria.

```
procedure DIANA (X,K,ACt)
Input: X = {P1, P2,...,PN}     % dataset with N patterns
       K = % number of clusters to be created
Output: ACt
begin
t ← 1
AC_t ← {X}                      %  initial clustering
C ← X
repeat
  Maxd ← maxdis(C)    % finds the most dissimilar pattern
  C ← C - {Maxd}
  TempC ←{Maxd}
  repeat
   for each pattern  P ∈ C find:
     DP = (averagedis(P,y), P,y ∈ C and P≠y) -
               (averagedis (P,z), z ∈ TempC)
     Find pattern P* with the highest value for D_P(D_P*)
     if D_P* > 0 then TempC ← TempC ∪ {P*}
  until all D_P ≤ 0
  t ← t+1
  C_t ← TempC
  AC_t ←  (AC_t-1 - {C}) ∪ {C,C_t}
  C ←  find_cluster_with_highest_diameter(AC_t)
until t = K
return AC_t
end
```

Fig. 2. A high level pseudocode of the DIANA divisive clustering algorithm.

The cluster splitting process goes as follows. Consider that C is a cluster already formed and the goal is to split C into two clusters such that the resulting two clusters have the largest possible dissimilarity value between them than all the other two clusters that could be obtained. To do that, initially it is necessary to identify the data point that is most dissimilar to the other data points in C; the algorithm does that by identifying a data point in C that has the maximum average dissimilarity value with the remaining data points in C. The identified data point is removed from C and starts a new cluster C*. For each data point in C two average dissimilarity values are computed: its average dissimilarity with the remaining data points in C and its average dissimilarity with data points in C*. Depending on the comparison of both results, the data point remains in C or, then, is transferred to C*.

4 Data Domain for the Experiments

Zahn [8] proposes a family of algorithms based on minimal spanning trees, for detecting various types of cluster structures in point sets characterized as *gestalt* clusters. Figure 3 exhibit 8 *gestalt* point sets, seven of them based on those described in [8], used in the experiments in Sect. 5.

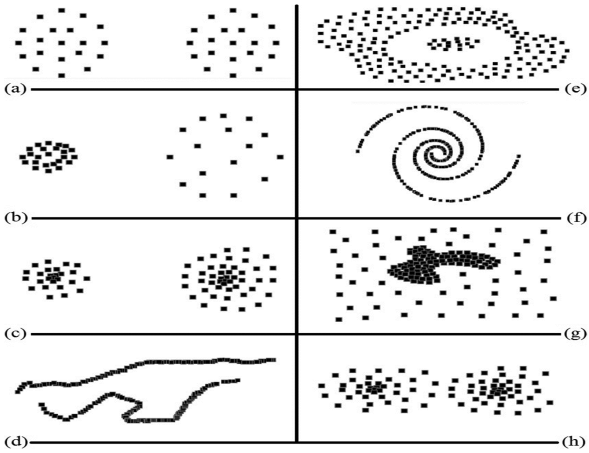


Fig. 3. Point sets used in the experiments, having (a) 40 points (20–20); (b) 40 points (15–24); (c) 72 points (26–46); (d) 200 (111–89); (e) 174 points (158–16); (f) 500 (250–250); (g) 142 (50–92); (h) 83 (39–44).

The development of such family of algorithms was based on a collection of sets containing two-dimensional data points, in view of the purpose of the author being to mimic the human perception of *gestalt* or point grouping. His research mainly focused on the problem of detecting inherent separations between clusters of a given set of points S, in a metric space governed by a distance function $\rho(x, y)$. According to the author, the expression *inherent separations* aimed to emphasize that any detected separation would only depend on the distances between the points in set S and resumes the goal of his research as a question to be answered: "what does the set S look like in terms of the structure of the space in which it is embedded?". The clustering detection method in his research was motivated by the human perception of two-dimensional point sets as separate groupings or *gestalts*, where the principle of grouping is proximity.

According to Zhan, regarding each of the point sets in Fig. 3(a), (d) and (e), could naturally be seen as two distinct clusters of points. Figure 3(a) shows a pair of well separated clusters, both with similar and homogeneous point density. The situation shown in Fig. 3(b) is quite similar to that of Fig. 3(a), except with respect to the densities of both clusters, which are different. In Fig. 3(c), although the volume of data points in each cluster is different, the density of each cluster varies in a similar way, being higher at its center; both, however, are well separated. Figure 3(a) and (d) show similarities; however, the two clusters in (a) and in (d) have completely different shapes. Figure 3(f)

is not listed in [8] but represents the well-known hard clustering task, involving recognizing each one of the two spirals as a cluster, in a two-cluster clustering task. Figure 3(h) is similar to Fig. 3(c) except that clusters touch each other and are not really separated; the separation can be detected by perceiving that the point density is at a local minimum in the region where the clusters touch each other.

5 Experiments, Results and Analysis

The comparative experiments described in this section were conducted in the eight point sets from Fig. 3. The two hierarchical approaches from Sects. 2 and 3 were employed; both have been implemented in C# and run under a Microsoft Windows platform. Results from the algorithm K-means [10] using the same input data are also presented, as baseline. To make a fair comparison, for all three algorithms (agglomerative, divisive and K-means) the number of clusters was the same, supplied by the user. In the following tables the best results are bold faced, except for the first three, considering that the algorithms performed similarly, as the values of the associated four indexes were the same.

For running the experiments the eight point sets, artificially created, as described in Sect. 4, shown in Fig. 3 and seven of them based on those proposed in [8] have been used. It is important to stress that the eight point sets try to embed the human perception of two-dimensional point sets as separate groupings or, as referred by Zahn, *gestalts*. The same procedure adopted for the hierarchical approaches under investigation was also used with the K-means, with K given as the number of actual perceptual groups in the data (*i.e.*, $K = 2$ for all eight point sets).

The results obtained were evaluated using three internal validation indexes: the Dunn's index (D) [11], the Davies-Bouldin index (DB) [12] and the Silhouette index (S) [13]. In order to quantify the number of data points incorrectly assigned (using the original two classes, which have been assigned based on the human perception of point sets as separate groupings or *gestalts*, part of the description of each data point in the synthetic datasets), the Rand index (R) external validation index was also used [14].

When using the agglomerative scheme and considering that a total of four linkage functions were employed namely: *Single Linkage*, *Complete Linkage*, UPGMA (*Unweighted Pair Group Method with Averaging*) and WPGMA (*Weighted Pair Group Method with Averaging*), the four versions of the scheme have been used in the experiments and are referred to by the linkage function each of them employs. In relation to the divisive approach, only the scheme described in Sect. 3 was used.

Results from the three validation indexes and the Rand index, applied to the clustering results from experiments with point sets from Fig. 3(a), (b) and (c) can be seen in Tables 1, 2 and 3, respectively. It can be seen in the three tables that all algorithms achieved the same results and successfully detected the two clusters. It is important to mention that the Rand index for the clusterings induced by the six algorithms in the three point sets was 1.00.

Table 1. Clustering results – point set in Fig. 3(a).

Clustering strategy	D	DB	S	R
Single Linkage	0.99	0.38	0.72	1.00
Complete Linkage	0.99	0.38	0.72	1.00
Average Linkage (UPGMA)	0.99	0.38	0.72	1.00
Average Linkage (WPGMA)	0.99	0.38	0.72	1.00
DIANA	0.99	0.38	0.72	1.00
K-means	0.99	0.38	0.72	1.00

Table 2. Clustering results – point set in Fig. 3(b).

Clustering strategy	D	DB	S	R
Single Linkage	0.86	0.31	0.78	1.00
Complete Linkage	0.86	0.31	0.78	1.00
Average Linkage (UPGMA)	0.86	0.31	0.78	1.00
Average Linkage (WPGMA)	0.86	0.31	0.78	1.00
DIANA	0.86	0.31	0.78	1.00
K-means	0.86	0.31	0.78	1.00

Table 3. Clustering results – point set in Fig. 3(c).

Clustering strategy	D	DB	S	R
Single Linkage	0.99	0.31	0.78	1.00
Complete Linkage	0.99	0.31	0.78	1.00
Average Linkage (UPGMA)	0.99	0.31	0.78	1.00
Average Linkage (WPGMA)	0.99	0.31	0.78	1.00
DIANA	0.99	0.31	0.78	1.00
K-means	0.99	0.31	0.78	1.00

The numbers obtained from experiments with the point set in Fig. 3(d) (Table 4) are evidence that the *Single Linkage* clustering strategy achieved the best result when the evaluation of results was done using indexes D and R, while the *Complete Linkage* achieved the best result when the DB and S indexes were used. Additionally, the DIANA and the K-means results had the same values for the four indices. Notice also that the value of the Rand index was 1.00 only for the clustering induced by the *Single Linkage*.

The next set of experiments had as input the point set in Fig. 3(e) and the values of the four indexes are presented in Table 5. Again, the *Single Linkage* clustering strategy, as far as the values of indexes D and R are concerned, had the best performance. However, as far as the DB index is concerned, the *Average Linkage* (WPGMA) produced the best clustering. When S index values are considered, clusterings induced by the *Complete Linkage*, the *Average Linkage* (WPGMA), the DIANA and the K-means are equally good; however, the Rand index value for the four clusterings was 0.50.

Table 4. Clustering results – point set in Fig. 3(d)

Clustering strategy	D	DB	S	R
Single Linkage	**0.10**	1.95	0.26	**1.00**
Complete Linkage	0.02	**0.76**	**0.49**	0.53
Average Linkage (UPGMA)	0.02	0.87	0.44	0.56
Average Linkage (WPGMA)	0.02	0.89	0.41	0.51
DIANA	0.02	0.84	0.45	0.50
K-means	0.02	0.84	0.45	0.50

Table 5. Clustering results – point set in Fig. 3(e)

Clustering strategy	D	DB	S	R
Single Linkage	**0.13**	7.88	0.26	**1.00**
Complete Linkage	0.07	0.92	**0.43**	0.50
Average Linkage (UPGMA)	0.06	0.99	0.40	0.50
Average Linkage (WPGMA)	0.07	**0.91**	**0.43**	0.50
DIANA	0.05	0.92	**0.43**	0.50
K-means	0.06	0.92	**0.43**	0.50

Taking into account the experiments with the point set in Fig. 3(f) (representing the two-spiral problem) and its results shown in Table 6, it can be noticed that no clustering strategy successfully detected the two clusters even though the *Single Linkage* would be the best alternative for clustering elongated clusters. Nevertheless, the *Single Linkage* clustering strategy was the one with the best performance in this experiment, although not a satisfactory one, considering the value 0.52 of its Rand index and the value 0.50 or 0.51 for all the others.

Table 6. Clustering results – point set in Fig. 3(f).

Clustering strategy	D	DB	S	R
Single Linkage	**0.08**	0.83	0.48	**0.52**
Complete Linkage	0.01	1.36	0.25	0.50
Average Linkage (UPGMA)	0.06	**0.60**	**0.59**	0.51
Average Linkage (WPGMA)	0.01	1.31	0.26	0.50
DIANA	0.02	1.03	0.38	0.50
K-means	0.01	1.17	0.31	0.51

Considering the results in Table 7, from the experiments with the point set in Fig. 3(g), the values of the R index show that the *Average Linkage* WPGMA clustering strategy induced the best clustering, even though failed to correctly identifying one of the two clusters. However, based on the values of the D, DB and S indexes the *Single Linkage* clustering strategy had a better performance. It is important to mention that the *singleton* cluster (*i.e.*, a cluster with only one data point), which was one of the two clusters resulted in the experiment with the *Single Linkage*, incorrectly influenced the

D, DB and S index computations Table 8 shows the index results with the clusterings obtained using the point set in Fig. 3(h). This is the first point set (among the last five) where the DIANA and the K-means algorithms produced a clustering with value 1.00 for the Rand index. Moreover, the values for the D, DB and S indexes also confirm and validate the clustering induced by both algorithms. Considering that the point set shown in Fig. 3(h) reflects two groups close together, this has introduced an extra difficulty for the agglomerative scheme to deal with, which increased the difficulty level of identifying them separately.

Table 7. Clustering results – point set in Fig. 3(g).

Clustering strategy	D	DB	S	R
Single Linkage	**0.12**	**0.52**	**0.63**	0.54
Complete Linkage	0.11	1.20	0.33	0.68
Average Linkage (UPGMA)	0.11	0.70	0.50	0.56
Average Linkage (WPGMA)	0.08	1.69	0.23	**0.73**
DIANA	0.03	1.08	0.35	0.51
K-means	0.04	1.21	0.34	0.68

Table 8. Clustering results – point set in Fig. 3(h).

Clustering strategy	D	DB	S	R
Single Linkage	0.11	0.59	0.57	0.49
Complete Linkage	0.06	0.68	0.50	0.80
Average Linkage (UPGMA)	0.15	0.57	0.58	0.93
Average Linkage (WPGMA)	0.15	0.57	0.58	0.93
DIANA	**0.17**	**0.54**	**0.60**	**1.00**
K-means	**0.17**	**0.54**	**0.60**	**1.00**

6 Conclusions

This paper addresses the use of hierarchical algorithms in unsupervised tasks in the domain of *gestalt* clusters (*i.e.*, point sets perceived as separate groupings). A representative of both, the agglomerative and the divisive families of hierarchical algorithms, were chosen for the study.

Except for the first three tables showing experiment results, were the results were the same for the six algorithms considered, the numbers displayed in Tables 4, 5, 6, 7, 8 suggest that, in most cases, the best results were obtained with the agglomerative approach. The *Single Linkage* strategy used by the agglomerative hierarchical approach could be considered as the winner in three out of five experiments. When specifically taking into account only results from experiments with the point sets in Fig. 3(d) and (e), the *Single Linkage* strategy presented its best performance as confirmed by the Rand index values. It is important to mention that both DIANA and K-means presented similar results and performed better than the agglomerative in experiments with the point set in

Fig. 3(h), due to the way the two clusters are positioned close to each other as well as the way the algorithms perform.

Acknowledgments. The authors would like to thank CAPES, CNPq and FACCAMP.

References

1. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, USA (2001)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, UK (2006)
3. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann-Elsevier, Amsterdan (2012)
4. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, USA (1990)
5. Theodorides, S., Kotroumbas, K.: Pattern Recognition, 4th edn. Elsevier, USA (2009)
6. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. **31**(3), 264–323 (1999)
7. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis. Wiley, London (2011)
8. Zahn, C.T.: Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Trans. Comput. **C-20**, 68–86 (1971)
9. Macnaughton-Smith, P., Williams, W.T., Dale, M.B., Mockett, L.G.: Dissimilarity analysis: a new technique of hierarchical sub-division. Nature **202**, 1034–1035 (1964)
10. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press (1967)
11. Dunn, J.: Well separated clusters and optimal fuzzy partitions. J. Cybern. **4**, 95–104 (1974)
12. Davies, D.L., Boudin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-1**, 224–227 (1979)
13. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
14. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**, 846–850 (1971)

Improving Imputation Accuracy in Ordinal Data Using Classification

Shafiq Alam[1(✉)], Gillian Dobbie[2], and XiaoBin Sun[2]

[1] Faculty of Business and IT, Whitireia Community Polytechnic, Auckland, New Zealand
shafiq.alam@whitireia.ac.nz
[2] Department of Computer Science, University of Auckland, 38 Princes Street, 1010 Auckland, New Zealand
gill@cs.auckland.ac.nz, allensun421@gmail.com

Abstract. Tackling missing data is one of the fundamental data preprocessing steps. Data analysis and pattern extraction are affected due to the underlying differences between instances with and without missing data. This is a particular problem with ordinal data, where for example a sample of a population may have all failed to answer a specific question in a questionnaire. The existing methods such as listwise deletion, mean attribute substitution, and regression substitution, naively impute data. They do not impute plausible values as they fail to take into account the relationships between the attributes, but instead consider the distribution of the attribute with missing values only. In this paper we introduce the use of Classification Based Imputation (CNI) to replace missing values with plausible values in ordinal data. The results show that not only does the CNI based technique outperform the existing approaches for imputing missing values in ordinal data but it also helps to improve the classification accuracy of machine learning algorithms.

1 Introduction

Statisticians and data analysts often distinguish quantities that can be represented in an attribute using different levels of measurement; nominal, ordinal, interval and ratio quantities [1]. A set of data is said to be ordinal if the contained values can be ranked or have a rating scale attached. The categories for an ordinal data set have a natural order, but the distinction between neighbouring values is not always the same. Examples of such data are often found in questionnaires; for the well quoted example of biscuits, suppose a group of people were asked for their likings about a biscuit on a rating scale of 1 to 5. A rating of 5 indicates more enjoyment than a rating of 4. However, the difference in enjoyment expressed by giving a rating of 2 rather than 1 might be much less than the difference in enjoyment expressed by giving a rating of 4 rather than 3.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_5

Mining ordinal data has many applications. For example, in empirical studies, surveys and questionnaires are one of the primary sources of data collection. Most of the data collected in such a way is ordinal and includes human participants. Analysis of such ordinal data is the main tool for understanding the psychological behaviour of participants. However, humans tend to omit responding to some of the questions in the questionnaires and this leads to missing values in the ordinal data. There can be various reasons for the existence of missing values in a dataset. The respondents may not want to answer certain questions with a sensitive topic, there may be data transmission problems, human or machine faults when the data is being recorded. The mechanism concerning the randomness of the missing data can be divided into three classes [2]. Missing Completely at Random (MCAR) occurs when the probability of an instance having a missing value for an attribute is unrelated to both the value of the variable itself, and also to the values of any other variables in the dataset. Missing at Random (MAR) occurs when the probability of an instance having a missing value for an attribute may depend on the known values, but not on the value of the missing data itself. Not Missing at Random (NMAR) occurs when the probability of an instance having a missing value for an attribute is related to the value of the attribute itself [3].

De Leeuw et al. [4] have investigated the problems caused by missing data in statistical analyses, which include biased parameter estimates and inflated standard errors. These possible issues may be particularly serious when the cause of a missing response is directly related to the value itself. Also, when there are large amounts of missing data, the power of the statistical analysis can be severely compromised. In Data Mining and Knowledge Discovery from Databases (KDD), data quality is a major concern [5]. Despite the fact that there are frequent occurrences of missing data in real world data sets, most machine learning algorithms naively handle the problem of missing data, which introduces bias into analysis.

The attributes of most datasets are not independent of each other. Therefore, missing values could be determined from the identification of relationships among attributes. Imputation is the procedure of replacing the missing values in a data set by some other value. Given the potential problems that arise from the presence of missing data, a variety of methods have been suggested for dealing with such problems. The most common and simplest approach to handle missing data is to simply omit the cases with missing data and to run the analysis on the remaining dataset; this approach is usually called list-wise deletion. A second approach is called mean substitution, it simply substitutes the average value of a particular attribute in the missing one. There are some problems with this approach; in the first place it adds no new information. In addition, such approaches lead to an underestimate of error, and the value being substituted may not be a plausible value, so it is not appropriate for ordinal data. A third approach is regression substitution, which uses existing variables to make a prediction and then substitute that predicted value as if it were an actual obtained value. With this method, the imputed value is in some way conditional on other information in the data set, however, the problem of error variance remains.

Su et al. [6] proposed Classifier-based Nominal Imputation (CNI), a imputation technique for nominal data that views imputation as a classification problem. The result of the experiments have shown that it improves the accuracy of imputation on nominal data, and also improves the classification performance of some machine learning algorithms.

The objective of this paper is to apply the CNI technique proposed by Su et al. to ordinal features instead of nominal data, and then compare and analyse experimental results to understand the performance on ordinal datasets. We report two parameters, firstly, estimation accuracy, which shows how accurate the imputation is for ordinal datasets as compared with a baseline imputation technique: most common value imputation (MCI), and secondly, classification accuracy, which shows to what extent this CNI technique could help to improve the classification performance of some machine learning algorithms on the ordinal datasets with missing data.

Section 2 of the paper describes some of the related work. Section 3 gives the background of the proposed technique. Section 4 explains the experimental procedure, the experiments, the experimental result and discussion around the obtained results, and Sect. 5 draws conclusions and proposes future directions for research.

2 Related Work

Most of the machine learning methods for classification assume the datasets are unordered, but in many real world applications these data do exhibit a natural order. In [1] the authors have proposed to use the ordering information of the datasets by converting the ordinal class problem into binary class problems for the learning of machine learning based classification algorithms. The reported results compared with naive approaches have shown better classification accuracy. Also when compared to other algorithms that were specifically designed for the ordinal classification problem, their method could be applied to ordinal datasets without any changing the underlying learning scheme. Huhn and Hullermeier [7] has shown that the method proposed by Frank and Hall [1] for ordinal classification is indeed able to exploit order information about classes.

Imputation for ordinal datasets such as questionnaires is common because ordinal data is often faced with the problem of missing data. In [8] the authors suggested methods to deal with the problem of missing data. In [9] the authors evaluated different methods for the imputation of missing values in likert data; a kind of ordinal data that is collected for understanding human behaviour. Their results concluded that two methods, item mean based imputation, and person-mean substitution performs well for smaller ration of missing data (less than 20%). The authors of [10] emphasized on the significance of imputation and proposed relative mean based imputation for likert data. The comparison results show that the substitution performance for the give datasets that they have tested their approach on, outperforms the simple mean based imputation.

Su et al. [11] proposed classifier-helped imputation which uses accurate imputation techniques, such as Bayesian multiple imputation (BMI), predictive mean

Algorithm 1. CNI(ImputationLearner L)

```
for t = 1...n do          % over each feature (attribute column of dataset D)
    % View feature f_i as the class and the other columns as features
    Divide instances D = D_i^+ + D_i^-
      where training set D_i^+ contains the instances with observed values of attribute
      f_i, and test set D_i^- contains the instances that miss values of attribute f_i
    Let c_i(...) = L(D_i^+) be classifier trained using learner L on D_i^+
      using L's default approach to handle missing data, as necessary
    for each instance d ∈ D_i^- do
        Let d_-i be the n - 1 "non-f_i" values in d
        Impute the "f_i" value of d as c_i(d_-i)
    % Move on to the next attribute
% return the resulting completed dataset D'
```

Fig. 1. Classifier Based Nominal Imputation (CNI) [6]

matching (PMM), and Expectation Maximization (EM) to preprocess missing data for a machine learning algorithm. Their results show that EM and BMI effectively imputed data in case of "missing completely at random (MCAR)". The problem with this method is that it has been evaluated for nominal data only and its effect is not known on ordinal data. Their proposed nominal imputation technique called Classifier-based Nominal Imputation (CNI) learns a classifier for the nominal feature and then applies the trained classifier for predicting missing attribute values. They investigated different classification algorithms for imputation, and their results show that SVM and decision tree have the best performance. Also, these learners have much higher imputation accuracies compared to other commonly used nominal imputation techniques, such as k-nearest neighbour imputation (KNNI) and most common imputation (MCI). Their results also show an improvement in classification accuracy after such imputation has been performed to treat the missing data (Fig. 1).

In the research area, the comparison of different imputation methods for numerical and nominal data is very common. Batista and Monard [12] investigated the use of the k-nearest neighbour algorithm as an imputation method for nominal datasets; the experimental results shown that it outperform the internal methods used by C4.5 and CN2 to treat missing data, and also it outperforms the mean or mode imputation method, which is a method used widely to treat missing values.

The area of ordinal data imputation and its effects on learning are still under explored. This paper intends to fill that gap and proposes the use of CNI based imputation for treating missing values in ordinal data. This paper also explores the effect of this imputation on learning for data mining.

3 Classifier Based Nominal Imputation Algorithm (CNI)

Classifier Based Nominal Imputation (CNI) was proposed for treating missing data for nominal attributes. The basic idea of CNI is to use classification for imputation. For each attribute f_i of the a tuple that contains missing values,

classifier $c_i(...)$ is trained which takes the remaining $n-1$ attributes $\{f_j|j \neq i\}$ of that tuples as an input, and returns the value for f_i of the this tuple based on the training of the learner. The learning algorithm L is used by $CNI(L)$ for the training of the classifier $c_i(...)$ based on the training data with non missing values on f_i, CNI then uses this classifier $c_i(...)$ to predict the missing values of this attribute for the remaining tuples. Algorithm 1 describes the functionality of the CNI imputation algorithm. After the imputation of missing data using CNI(L), the imputed dataset is passed to the base learner B that trains a classifier to classify the testing set of ordinal data with the missing values. This classifier is able to predict the missing values as a class for the tuples that contains incomplete ordinal data.

4 Experimental Setup

This section includes a description of the datasets, the type of experiments and the experimental results. We have performed two types of experiments, one to assess the imputation accuracy and the second to assess the classification accuracy of the algorithms on the imputed ordinal data.

Table 1. Datasets from UCI machine learning repository and WEKA

(a) Dataset details

Dataset	#Instances	#Attributes(ordinal/all)	#Class
CAR	1728	5/7	1
ERA	1000	5	1
ESL	488	5	1
LSV	1000	5	1
SWD	1000	11	1

(b) Dataset CAR

ATTRIBUTES	VALUES
buying	vhigh, high, med, low
doors	2, 3, 4, 5more
lug_boot	small, med, big
maint	vhigh, high, med, low
persons	2, 4, more
safety	low, med, high
class	unacc, acc, good, vgood

Five ordinal real-world datasets, which were drawn from the UCI machine learning repository [9] and the WEKA dataset repository [10] have been explored. (See dataset descriptions in Table 1a.) The dataset CAR contains 5 ordinal attributes and 2 non ordinal attributes and all other datasets contain ordinal attributes only. Except for the dataset CAR, all other ordinal values in each dataset are represented by numbers and their values indicate rank information. Table 1 is the summary of each attribute for dataset CAR. In order to analyse the behaviour of the CNI technique and to measure the performance of both the estimation accuracy and classification accuracy on datasets for different missing ratio levels, four datasets which contain missing data have been extracted from the original datasets in a random fashion by deleting 5%, 10%, 20%, 30% of the observed values, with the probability of 0.05, 0.1, 0.2, 0.3, using 10-folds cross validation.

4.1 Experimental Results

Estimation Accuracy: Since datasets with missing values were generated from the original complete datasets, the truth for each missing value is clear, thus, the CNI algorithm can be evaluated against the ground truth for the estimation of accuracy by comparing each imputed value against the true values accordingly. Estimation accuracy is calculated using the following equation.

$$Estimation\ Accuracy\,(in\%) = \frac{1}{N}\sum_{j=1}^{N} I(X_j = Y_j) \tag{1}$$

where function I returns 1 for $X_j = Y_j$, and 0 otherwise, X_j is the imputed value of j^{th} instance produced by the imputer, Y_j is the true value of the j^{th} instance from the original complete dataset, and N is the total number of imputed values.

To measure the performance of the estimation accuracy of the CNI technique, five different imputation learners L has been used for each dataset: most common imputation (MCI), decision tree (C4.5), support vector machine (SVM), random forest (RF) and logistic regression (LR). By applying these learners for each attribute of the dataset in WEKA, the complete imputed dataset can be obtained. We get the estimation accuracy of the CNI technique for each different ratio by comparing the imputed datasets with original dataset.

The results show that the CNI algorithm has a significant improvement in estimation accuracy when compared to the baseline most common imputation (MCI) technique on the ordinal data. The imputed values are much closer to the original datasets than the results using the MCI method. Also, the performance of estimation accuracies decreases as the missing data ratio increases. Table 2 shows the imputation accuracies. For the missing ratios of 5%, the average estimation accuracy of baseline imputer MCI is 28.94% and the average estimation accuracy for all CNI imputers is 61%, which indicates a big improvement over MCI. The lowest value is using support vector machine (SVM) as the imputer, which is 42.92% for dataset LSV and the highest estimation accuracy is using random forest (RF) as the imputer for dataset SWD, which is 94.37%. For the missing ratios of 10%, the average estimation accuracy of baseline imputer MCI is 29.08% and the average estimation accuracy for all CNI imputers is 57.66%. The lowest value is 38.61% using support vector machine (SVM) as the imputer for dataset CAR, while the highest estimation accuracy is from using random forest (RF) as the imputer for dataset SWD, which is 92.09%. For the missing ratios of 20%, the average estimation accuracy of baseline imputer MCI is 29.34% and the average estimation accuracy for all CNI imputers is 51.45%. The lowest value is 37.33% from using support vector machine (SVM) as the imputer for dataset LSV, while the highest estimation accuracy is using random forest (RF) as the imputer for dataset SWD, which is 84.64%. For the missing ratios of 30%, the average estimation accuracy of baseline imputer MCI is 30.08% and the average estimation accuracy for all CNI imputers is 47.08%. The lowest value is 33.33% using decision tree (J48) as the imputer for dataset LSV while the highest estimation accuracy is using random forest (RF) as the imputer for dataset

Table 2. Estimation accuracy for different missing ratios

(a) Estimation accuracy for missing ratio at 5%

5%	MCI	J48	LR	RF	SVM
CAR	33.37%	**46.63%**	44.19%	43.93%	44.89%
LSV	27.20%	62.88%	49.16%	**72.68%**	42.92%
ESL	32.08%	52.00%	55.00%	50.17%	**54.83%**
ERA	12.24%	76.40%	77.60%	**81.47%**	74.20%
SWD	39.82%	84.73%	58.09%	**94.37%**	53.82%

(b) Estimation accuracy for missing ratio at 10%

10%	MCI	J48	LR	RF	SVM
CAR	34.11%	40.12%	**43.82%**	41.39%	38.61%
LSV	30.31%	55.64%	46.06%	**68.08%**	41.10%
ESL	29.14%	49.25%	51.29%	47.62%	**53.74%**
ERA	12.44%	72.93%	73.53%	**79.73%**	67.67%
SWD	39.41%	77.55%	57.36%	**92.09%**	55.68%

(c) Estimation accuracy for missing ratio at 20%

20%	MCI	J48	LR	RF	SVM
CAR	35.75%	39.63%	**41.77%**	40.55%	38.99%
LSV	28.90%	39.46%	42.95%	**59.79%**	37.33%
ESL	30.10%	45.03%	46.60%	47.21%	**48.57%**
ERA	12.80%	53.90%	65.87%	**72.83%**	57.63%
SWD	39.14%	62.09%	53.50%	**84.64%**	50.73%

(d) Estimation accuracy for missing ratio at 30%

30%	MCI	J48	LR	RF	SVM
CAR	35.78%	38.81%	**39.34%**	39.27%	38.92%
LSV	29.18%	33.33%	41.51%	**52.44%**	34.75%
ESL	31.36%	42.87%	44.93%	45.25%	**46.30%**
ERA	12.76%	37.07%	57.71%	**67.31%**	50.27%
SWD	41.33%	52.70%	53.12%	**75.85%**	49.81%

SWD, which is 75.85%. All CNI based algorithms outperform the baseline MCI algorithm and show a significantly higher estimation accuracy.

4.2 Classification Accuracy

For the measurement of classification accuracy, it can be used as a way to determine how much this CNI technique could improve the classification performance when compared with the classification on missing only datasets and the baseline technique, most common value imputation. Classification accuracy is calculated using the following equation.

$$Classification\,Accuracy\,(in\%) = \frac{1}{N}\sum_{j=1}^{N} I(M_j = K_j) \tag{2}$$

where N is the total number of class values in the testing set, M_j is the classified class from the result of classification on the testing set and K_j is the true value in the original non-missing testing set, and function I returns 1 for $M_j = K_j$ and 0 otherwise.

We used the imputed datasets and randomly separate each dataset into a training set and testing set (50–50% and 20–80%). We investigated the classification performance based on imputed data for three machine learners, K Nearest Neighbour Algorithm (KNN-5), Naive Bayes algorithm (NB), multilayer perception neural network algorithm (MLP) on incomplete ordinal data. We measured the classification accuracy on both the original datasets which contain no missing values and the datasets that contain missing values with different ratios, then compare these results.

4.3 K Nearest Neighbour Algorithm (KNN-5)

In our first experiment, we evaluated the results using the K Nearest Neighbour algorithm. For the dataset CAR, the original classification accuracy is 89.58%. For the same dataset when it contains missing values, the classification accuracy is quite low (78%). With imputation using MCI, the classification accuracy goes down further to 74%. But when different CNI imputers are used as pre-processors, the classification accuracy improves and moves closer to the original classification on different missing ratios from 5% to 30%. The performance varies from the smallest lift of 1.06% (which occurs using the CNI algorithm with J48 as the imputer at 5% missing ratio level) to the biggest lift of 11.24% (which occurs using the CNI algorithm with J48 as the imputer at 20% missing ratio level). Figure 2 shows the results of the experiment. For the same dataset CAR with 20% as the training set, with the remaining 80% as the testing set, the classification accuracy is 82.49% for the original dataset. For the dataset that contains missing values, the classification accuracy decreases to 77%. Similar to the 50–50 split, the baseline imputation technique MCI performs worse on this split as well. For different CNI imputers as pre-processors, the classification accuracies improve on different missing ratios from 5% to 30% when compared to the dataset that contains missing values. The performance varies from the smallest lift of 0.18% to biggest lift of 8.86%.

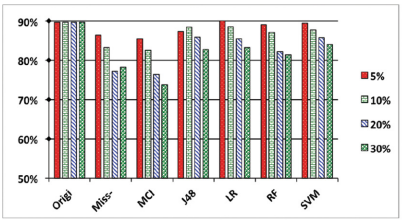
For the dataset ESL, the classification accuracy is 57.79% for the original data set with no missing values. MCI performs worse among all the imputers while for the CNI based algorithms the performance varies from the smallest lift of 2.22% (using random forest (RF) as the imputer at 10% missing ratio level) to the biggest lift of 27.81% (using logistic regression (LR) as the imputer at 30% missing ratio level).

For the dataset LSV, the classification accuracy is 60% for the original data set. For the dataset with missing values, the classification accuracy goes down to 50%. However, using different CNI imputers as pre-processors, the classification accuracy improves from the smallest lift of 1.36% (using random forest (RF) as the imputer at 5% missing ratio level) to the biggest lift of 23.58% (using support vector machine (SVM) as the imputer at 30% missing ratio level).

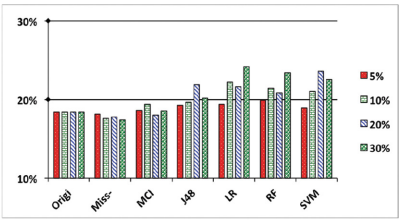
For the dataset ERA, the classification accuracy is 22.80% for the original data set. For the dataset with missing values, MCI performs worst whileusing different CNI imputers as pre-processors, the results have shown an increase in classification accuracy from the smallest lift of 4.67% (using support vector machine (SVM) as the imputer at 5% missing ratio level) to the biggest lift of 57.89% (using logistic regression (LR) as the imputer at 20% missing ratio level).

For the dataset SWD, the classification accuracy is 56.80% for the original data. Using different CNI imputers as pre-processors, the results overall have shown a better performance on classification accuracy where the smallest lift of 0.71% occurs using decision tree (J48) as the imputer at 5% missing ratio level, and the biggest lift of 13.11% occurs with support vector machine (SVM) as the imputer at 20% missing ratio level. Besides KNN, we have also used Naive Bayes algorithm and multilayer perception neural network algorithm to assess

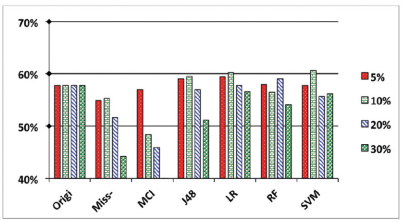
(a) CAR dataset (50,50 Split)



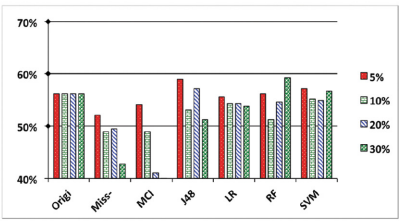
(b) CAR dataset (20,80 Split)



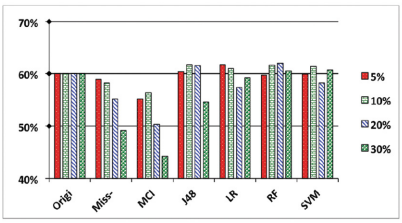
(c) ESL dataset (50,50 Split)



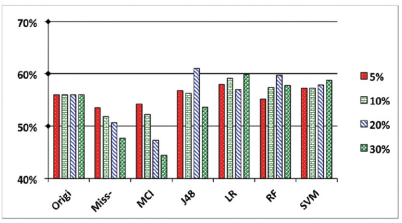
(d) ESL dataset (20,80 Split)



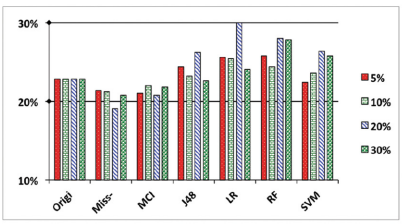
(e) LSV dataset (50,50 Split)



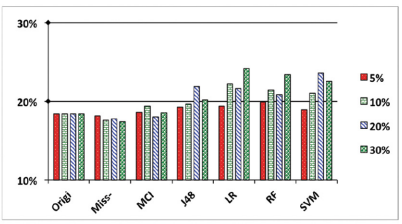
(f) LSV dataset (20,80 Split)



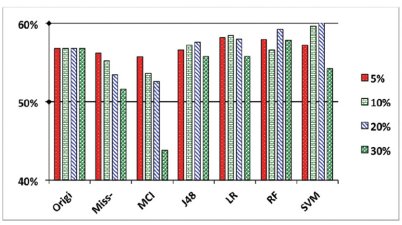
(g) ERA dataset (50,50 Split)



(h) ERA dataset (20,80 Split)



(i) SWD dataset (50,50 Split)



(j) SWD dataset (20,80 Split)

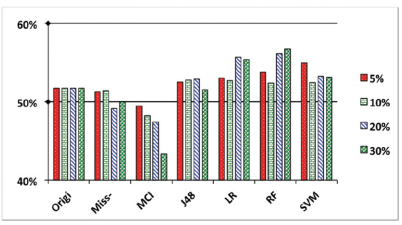


Fig. 2. Classification accuracy for the five datasets

Table 3. Classification accuracy for Naive Bayes and multilayer perception neural network (MLP)

(a) Dataset CAR using Naive Bayes algorithm

	Miss-Only	MCI	J48	LR	RF	SVM
5.00%	83.10%	81.60%	86.23%	85.30%	84.26%	**85.76%**
10.00%	82.06%	81.53%	86.46%	84.61%	84.95%	**86.81%**
20.00%	79.86%	76.12%	85.30%	85.76%	83.22%	**86.46%**
30.00%	80.32%	73.50%	81.48%	84.95%	81.02%	**85.53%**

(b) Dataset CAR using MLP algorithm

Miss-Only	MCI	J48	LR	RF	SVM
93.40%	88.08%	95.97%	95.45%	**97.11%**	94.44%
90.51%	87.32%	**94.33%**	93.64%	92.71%	92.36%
84.03%	83.87%	**88.54%**	87.38%	86.57%	87.85%
80.44%	75.81%	**85.88%**	85.65%	84.72%	84.49%

the improvement in classification accuracy of the imputed datasets. Table 3 shows these results. The accuracy for the original data without missing values is 85.07% for Naive Bayes, and 97.34% for MLP. For the sake of brevity, the detailed results have been omitted.

The estimation accuracy of different imputers varies on different missing ratios of each dataset, but on average, the random forest and logistic regression are among the top performers since these two appear in the top two more frequently, compared with the baseline MCI technique. For all these CNI imputers and MCI imputers that we have investigated here, their ranking on average estimation accuracies is: RF, LR, J48, SVM, and MCI. Note that these results are based on our 10-folds cross validation. The results also show that the estimation accuracy decreases when the missing ratio increases, which concludes that the correctness of the CNI imputation would suffer on higher ratios of missing data in the dataset.

For the classification accuracy, the aim is to evaluate how much these imputers could improve the classification performance for different kinds of machine learning algorithms. K-nearest neighbour (KNN), Naive Bayes (NB) and Multilayer Perceptron Neural Network (MLP) have been used as the base learning algorithms. The results have shown that for KNN the classification accuracy decrease dramatically with an increase in the missing ratio level. It happens even if the baseline imputation technique MCI has been applied to handle the missing data problem and in most cases this result is worse than the original incomplete dataset. But with the help of the CNI algorithm, K nearest neighbour algorithm can achieve a significant improvement over the classification accuracy when compared with the results from the original incomplete datasets. This truth prevails for all the missing ratios (5% to 30%) on both datasets with different splitting percentages (50%–50% and 20%–80% splits). All CNI based imputations have shown that the dataset imputed using these techniques significantly improves the learning of machine learning algorithms.

Overall the results show that using high quality imputation techniques such as the CNI algorithm to pre-process the incomplete ordinal data, improves the classification performance of machine learning algorithms.

5 Conclusions

Incomplete ordinal data is very common and is problematic in many fields of research, biasing data analysis because of underlying differences between instances with and without missing data. Many methods have been suggested including listwise deletion, in which incomplete cases are discarded from the dataset, mean substitution, and regression substitution, which replace the missing values in a data set by some plausible values. However, these techniques ignore the relationship of different attributes and instead impute an ordinal value based on single attribute distribution. In this work we investigated two research question, firstly, how good CNI based algorithms are for the imputation of ordinal data and secondly, does such imputation improve classification results.

To tackle the first problem we introduce the use of Classifier-based nominal imputation (CNI); an imputation technique for nominal data that views the imputation as a classification task, for tackling the problem of missing ordinal data. We applied the CNI technique on five different real world ordinal datasets and the experimental results show that the CNI imputers have better imputation accuracy as compared with the baseline imputation method: most common value imputation.

We also assessed how these imputations improve the classification accuracy of machine learning algorithms on the ordinal datasets with missing data. We evaluated three classification algorithms including k-Nearest Neighbour, Naive Bayes and Multilayer Perceptron Neural Network. The results show that the classification accuracy improves with the CNI based imputation as compared to classification accuracy of a dataset with missing ordinal data. In some cases the classification accuracy of CNI imputers even surpassed the classification accuracy of the original data with no missing values.

For the future work, it would be interesting to see what the result would be if applying this CNI algorithm on a mixed data type. It would also be beneficial to investigate possible ways to improve this algorithm, for example, using the previously imputed values for the later iterations of imputations, based on the distribution of the attributes or how dense or informative these attributes are.

References

1. Frank, E., Hall, M.: A simple approach to ordinal classification. In: Raedt, L., Flach, P. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 145–156. Springer, Heidelberg (2001). doi:10.1007/3-540-44795-4_13
2. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (1987)
3. Schafer, J.L., Graham, J.W.: Missing data: our view of the state of the art. Psychol. Methods **7**(2), 147 (2002)
4. De Leeuw, E.D., Hox, J., Huisman, M.: Prevention and treatment of item nonresponse. J. Official Stat.-Stockh. **19**(2), 153–176 (2003)

5. Finch, W.H.: Imputation methods for missing categorical questionnaire data: a comparison of approaches. J. Data Sci. **8**(3), 361–378 (2010)
6. Su, X., Greiner, R., Khoshgoftaar, T.M., Napolitano, A.: Using classifier-based nominal imputation to improve machine learning. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011. LNCS (LNAI), vol. 6634, pp. 124–135. Springer, Heidelberg (2011). doi:10.1007/978-3-642-20841-6_11
7. Huhn, J.C., Hullermeier, E.: Is an ordinal class structure useful in classifier learning? Int. J. Data Min. Modell. Manag. **1**(1), 45–67 (2008)
8. De Leeuw, E.D.: Reducing missing data in surveys: an overview of methods. Qual. Quan. **35**(2), 147–160 (2001)
9. Downey, R.G., King, C.V.: Missing data in likert ratings: a comparison of replacement methods. J. Gen. Psychol. **125**(2), 175–191 (1998)
10. Raaijmakers, Q.A.: Effectiveness of different missing data treatments in surveys with likert-type data: introducing the relative mean substitution approach. Educ. Psychol. Measur. **59**(5), 725–748 (1999)
11. Su, X., Khoshgoftaar, T.M., Greiner, R.: Using imputation techniques to help learn accurate classifiers. In: 20th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2008, vol. 1, pp. 437–444. IEEE (2008)
12. Batista, G.E., Monard, M.C.: An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. **17**(5–6), 519–533 (2003)

GA-PSO-FASTSLAM: A Hybrid Optimization Approach in Improving FastSLAM Performance

Alif Ridzuan Khairuddin, Mohamad Shukor Talib, Habibollah Haron(✉), and Muhamad Yazid Che Abdullah

Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
alifjaml991@gmail.com, yazidyt94@gmail.com, {shukor,habib}@utm.my

Abstract. FastSLAM algorithm is one of the introduced Simultaneous Localization and Mapping (SLAM) algorithms for autonomous mobile robot. It decomposes the SLAM problem into one distinct localization problem and a collection of landmarks estimation problems. In recent discovery, FastSLAM suffers particle depletion problem which causes it to degenerate over time in terms of accuracy. In this work, a new hybrid approach is proposed by integrating two soft computing techniques that are genetic algorithm (GA) and particle swarm optimization (PSO) into FastSLAM. It is developed to overcome the particle depletion problem occur by improving the FastSLAM accuracy in terms of robot and landmark set position estimation. The experiment is conducted in simulation where the result is evaluated using root mean square error (RMSE) analysis. The experiment result shows that the proposed hybrid approach able to minimize the FastSLAM problem by reducing the degree of error occurs (RMSE value) during robot and landmark set position estimation.

Keywords: Autonomous mobile robot · SLAM · FastSLAM · Soft computing · Genetic algorithm · Particle swarm optimization

1 Introduction

Simultaneous Localization and Mapping (SLAM) is an algorithm technique used for mobile robot to build and create a relative map from unknown environment. FastSLAM algorithm is one of SLAM algorithm which introduced by [1] that capable to speed up the convergence of robot path planning and map environment estimation. FastSLAM algorithm structure consists of four processes to be performed [2]. First, for each particle that incorporates the latest control draws a new robot pose. For each of the drawn pose, it is added into the appropriate robot path estimate. Second, updates the landmark using Extended Kalman Filters (EKFs) which correspond to the observed landmark with the new observation. Third, calculate the importance weight by assigning the weight value to each particle. This is to differentiate each of the particles since the robot path particle are not drawn from the true path posterior. Fourth, drawn a new particles set from the weighted particle done earlier in calculating importance weight process using importance resampling process. The importance resampling

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_6

process is important as this process ensures that the particles are distributed according to the true posterior within the limit of infinite particles. Despite that, FastSLAM algorithm suffers inconsistency results due to particle depletion problem over time.

FastSLAM main problem is it cannot produce a consistent estimation in a long-term [3]. This is because each particle that carries robot pose estimation information are implicitly records a pose history in the statistics of its associated. During resampling process in FastSLAM, only a high weighted particle is selected and replicated while the low weighted particles are eliminated. So, each time particle is lost, a whole map hypothesis is lost and this causes the depletion in historical information. Hence, the overall map statistics will suffer degradation.

In recent study, several researchers integrated the soft computing technique into SLAM algorithms [4–10]. Soft computing is technique used to solve the computational complexity problems and mathematically intractable problem. It provides flexible processing method and has been widely emerged in solving SLAM problem throughout the recent years [11]. In this work, an integration of two soft computing techniques that are genetic algorithm (GA) and particle swarm optimization (PSO) into FastSLAM algorithm for hybrid optimization approach is presented. This is to improve the performance of FastSLAM algorithm by reducing the degree of error occurs (RMSE value) made by the robot during estimating its own position and set of landmarks within the environment. Thus, it can prevent the particle depletion problem occur in FastSLAM. We adapts the works from [1, 4, 5] and the hybrid approach is called as GA-PSO-FastSLAM.

This paper is organized as follows. Section 2 explained about GA-FastSLAM and PSO-FastSLAM approaches, Sect. 3 discussed about this work that is GA-PSO-FastSLAM approach, Sect. 4 discussed about experiment setup and result and finally, Sect. 5 provides a conclusion to this work.

2 Background

2.1 GA-FastSLAM Approach

In [4], they are implementing a genetic algorithm into FastSLAM called as GA-FastSLAM. The genetic algorithm implementation is done after sampling a new robot pose for each particle. Evolution operators in genetic algorithm that is selection, crossover and mutation are used to optimize the system state represented by each particle. Metropolis algorithm is used to ensure that the evolved particles accurately approximate the posterior and the weight of particle represents it as a fitness function.

This work is trying to deal with particle depletion problem occur in FastSLAM which improved the solution quality by selecting solutions with higher fitness than average through iterated calculation. Also, by making samples moves to larger posterior probability density area, it able to minimizes the sample number which later can improve the samples diversity. Hence, the sample degradation problem occur in FastSLAM can be reduced.

2.2 PSO-FastSLAM Approach

The implementation of particle swarm optimization into FastSLAM called as PSO-FastSLAM done by [5] is introduced. In the work, particle swarm optimization is used as particle cooperation technique for optimizing the estimation of robot and landmark position accuracy. The concept of particle cooperation is to improve the performance of FastSLAM by introducing swarm intelligence to effectively tracking a target that is robot position and landmark locations.

It attracts the particles toward an actual robot position while maintaining the particle diversity. The particle cooperation is done after resampling process in FastSLAM where a target of global attractor is computed using mean of position of particles. Using the particles position means, particle swarm optimization is performed for particle cooperation.

This work is trying to deal with particle depletion problem occur in FastSLAM where by incorporating particle swarm optimization, the number of highly probable particles that estimate actual robot pose becomes larger and it can avoid the improper rejection of particles in FastSLAM.

3 The Proposed GA-PSO-FastSLAM Approach

This section explains the proposed GA-PSO-FastSLAM approach. The proposed approach is an improvement and combination of the GA-FastSLAM and PSO-FastSLAM approaches proposed by [4, 5] respectively. The idea of this work is that it adapts double evaluations (GA and PSO) into FastSLAM compared to single evaluation done by [4, 5]. The details of the proposed approach will be explained in later paragraph. The discussion will be started with the similarities and differences between proposed approach with FastSLAM, GA-FastSLAM and PSO-FastSLAM and followed by the details of the proposed approach in the later paragraph.

Three similarity of the proposed approach with previous work are discussed. First, general similarity of [1, 4, 5] is the fitness function used for both GA and PSO in the proposed approach are same which is an equation to evaluate the particle weight. Second, the original FastSLAM algorithm by [1] is adopted as a basis of the proposed approaches. Third, the PSO mechanism used by [5] is adopted in this proposed GA-PSO-FastSLAM approach.

The differences of the proposed approach with [1, 4, 5] is the step of GA where this work were optimizing the particle weight after calculating particle importance weight process while [4] were optimizing the state space of observation model during sampling robot pose process in FastSLSAM. The reason this work is trying to optimize the particle weight value is because it is considered as the most critical element in FastSLAM as it determines the correspond particles to be maintained or eliminated during resampling process. Also, the mechanism of optimization proposed by [4] might suffer an error when assigning particle weight due to miscalculation of improper data selection. Figures 1 and 2 show the flowchart and details explanation of the proposed GA-PSO-FastSLAM approach respectively.

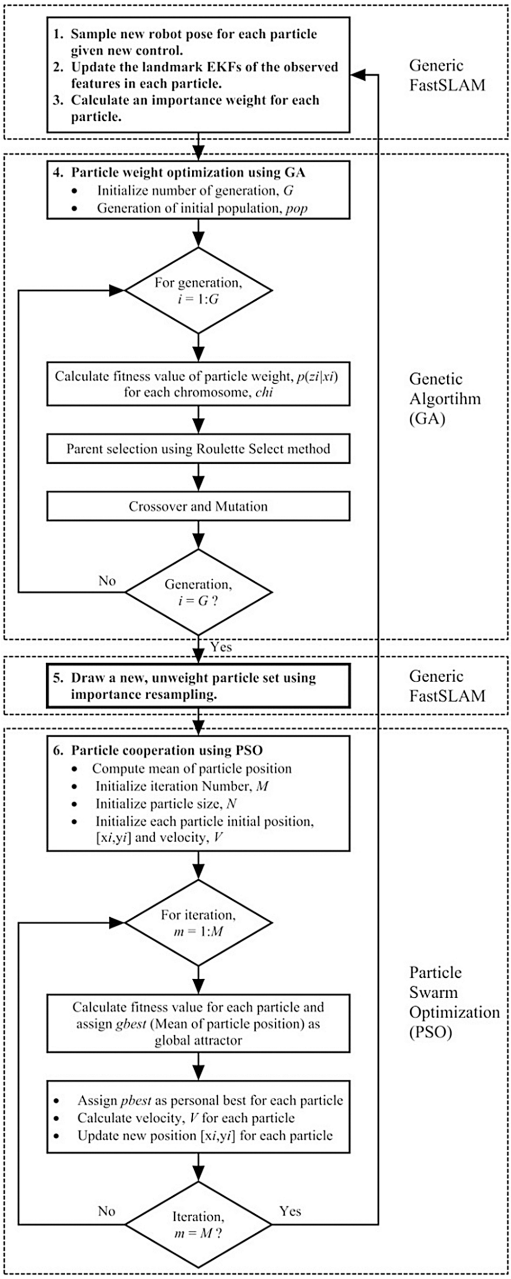


Fig. 1. GA-PSO-FastSLAM flowchart

Start Algorithm

1) Draw new robot pose sample for each particle *Generic*

2) Update landmark EKFs of observed features for each particle *FastSLAM*

3) Calculate important weight for each particle

---

4) Particle weight optimization using genetic algorithm (GA)

Start GA *Genetic*

Initialize number of generation, G *Algorithm*

Generation of initial population, *pop* *(GA)*

For generation, $i = 1{:}G$

Calculate fitness value of particle weight, $p(z_i|x_i)$ for each chromosome, ch_i

Parent selection using roulette select method:

Start

I. Calculate sum of group fitness value, $F = \sum_{i=1}^{chp} ch_i$ where chp is chromosome population

II. Calculate choice probability; P_i: $P_i = \frac{ch_i}{F}$ for each chromosome, ch_i

III. Calculate cumulative probability; Q_i: $Q_i = \sum_{i=1}^{chp} P_i$ of each chromosome, ch_i

IV. **For** number of replication individual

- Generate random number, r between 0 and 1
- Compare Q_i with r in turn and select first individual Q_i as replication object if those Q_i is more than r
- Repeat both above steps until replication individual number required be satisfied

End

End

Crossover:

Start

I. Random select two parents particles $\left\{x_i^{(a)}, x_i^{(b)}\right\}$ from collection and generate two new particles based on following definition:

$$\begin{cases} x_i^{(an)} = \xi x_i^{(a)} + (1-\xi) x_i^{(b)} \\ x_i^{(bn)} = \xi x_i^{(b)} + (1-\xi) x_i^{(a)} \end{cases}$$

where $\xi = [0,1]$ is random number between 0 and 1 of uniform distributed.

II. Replaced parent particles with new particles using Metropolis algorithm:

- Accept new particle, $x_i^{(an)}$ if $p\left(z_i \middle| x_i^{(an)}\right) > max\left\{p\left(z_i \middle| x_i^{(a)}\right), p\left(z_i \middle| x_i^{(b)}\right)\right\}$
- Else, accept with probability of $p\left(z_i \middle| x_i^{(an)}\right) / max\left\{p\left(z_i \middle| x_i^{(a)}\right), p\left(z_i \middle| x_i^{(b)}\right)\right\}$
- Use same process for new particle, $x_i^{(bn)}$

End

Mutation:

Start

I. Random select one parents particles $\left\{x_i^{(c)}\right\}$ from collection and generate new particles based on following definition:

$x_i^{(cn)} = x_i^{(c)} + \gamma$

where γ is random number of Gaussian distributed with covariance value of σ^2 and mean value of 0.

Fig. 2. GA-PSO-FastSLAM approach

II. Accept new particle, $x_i^{(cn)}$ if $p\left(z_i \middle| x_i^{(cn)}\right) > p\left(z_i \middle| x_i^{(c)}\right)$

III. Else, accept with probability of $p\left(z_i \middle| x_i^{(cn)}\right) / p\left(z_i \middle| x_i^{(c)}\right)$

End

End

End GA

---

5) Draw new unweight particle set using importance resampling *Generic FastSLAM*

---

6) Position estimation optimization using particle swarm optimization (PSO)

Start PSO

Compute mean of particle position — *Particle Swarm Optimization (PSO)*

Initialize iteration Number, M

Initialize particle size, N

For each particle, i = 1:N

Initialize particle initial position, $[x_i,y_i]$

Initialize particle initial velocity, V

End

For iteration number, m = 1:M

For each particle, i = 1:N

Compute fitness value, *fv* (Particle weight value)

End

Assign the *gbest* (Mean of particle position) as global attractor based on following definition:

$gbest_{m+1}$ = mean($[x_i,y_i]_m$) || $[x_i,y_i]_{m+1}$ - $[x_i,y_i]_m$ ||

For each particle, i = 1:N

- Assign *pbest* of the N-th of particle as personal best based on following definition:
  $pbest_{m+1}$ = || $[x_i,y_i]_{m+1}$ - $[x_i,y_i]_m$ ||
- Calculate particle N-th velocity, V based on following definition:
  $V_i(m+1) = hV_i(m)+qr(pbest_i(m)-[x_i,y_i](m))+qr(gbest(m)-[x_i,y_i](m))$
  where h is inertia factor, q is acceleration factor and r is uniform distribution random numbers on interval (0,1).
- Update particle N-th position, $[x_i,y_i]$ based on following definition:
  $[x_i,y_i]_{m+1} = [x_i,y_i](m) + V_i(m+1)$

End

End

End PSO

End Algorithm

Fig. 2. (continued)

Based on Fig. 2, the proposed approach started with generic FastSLAM algorithm steps (Step 1 to 3) where it starts with drawing new robot pose sample for each particle, update landmark EKFs of observed features for each particle and followed by calculating important weight for each particle.

In Step 4, GA process takes places where number of generation is initialized and initial population is generated. Next, GA operator takes places where for each generation number, four sub processes were performed. First sub process is calculating fitness value of particle weight for each chromosome. Second sub process is selecting parent using roulette select method to produce a number of individuals. Third sub process is generating two new particles based on parent particles by crossover operator. Then, these new generated particles replaced the current selected parent particles. Last

sub process is generating new particles based in parent particles using mutation operator. These new particle then replaced the current selected parent particle. These four sub processes are repeated for a specified number of generations and until the GA process is terminated. Once GA process is completed, it is followed by Generic FastSLAM algorithm process to draw a new unweight particle set using importance resampling (Step 5).

After that, PSO process (Step 6) is performed where in the first step, five parameter that are mean of particle position, PSO iteration number and three initial value of particle size, position and velocity were initialized. Next step is performing two sub processes by PSO operator for each iteration. The first sub process is calculating fitness value for each particle and selecting particle that has highest particle weight as global attractor. Second sub process is selecting particle personal best and calculating its velocity. These sub processes are repeated for a specified number of iterations until the PSO process is terminated. Step 1 to 6 was repeated until robot finished its exploration in the assigned environment map.

4 Experiment

4.1 Experiment Setup

Software used to conduct the experiment is Matlab and tool function used is a SLAM toolbox developed by [12]. In order to validate the proposed approach performance, a simulated environment map used by [5] is adopted and reconstructed in this work. Based on the reconstructed environment map, the proposed approach is evaluated and compared with other approaches in [1, 4, 5]. The simulated environment is two-dimensional map that contains landmarks (obstacles) and robot waypoint (robot path). Figure 3 shows the simulation environment map.

To ensure that the conducted experiment is performed properly, an appropriate parameter setting is needed. This is to avoid any unexpected error occurs during

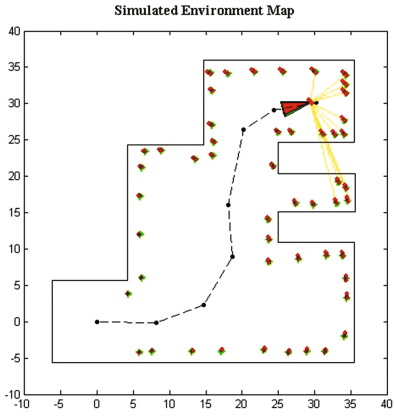


Fig. 3. Reconstructed simulated environment map based on [5]

Table 1. Experiment parameter setup

Parameter	Value(s)	Unit(s)
Number of simulation	50	Simulation
Robot velocity	1	m/s (Meter per Second)
Robot wheelbase size	4:2	Length:Width in m (Meter)
Time interval	0.05	s (Second)
Number of particle	100	Particle

simulation experiment. The parameter setup is applied to all approaches to ensure that the comparison made is fair. Table 1 shows the experiment parameter setup and its corresponding value.

4.2 Experiment Result

The performance evaluation of the GA-PSO-FastSLAM approach was evaluated using root mean square error (RMSE) analysis. The analysis is comparing the RMSE result of this work (GA-PSO-FastSLAM) with three approaches namely FastSLAM [1], GA-FastSLAM [4] and PSO-FastSLAM [5]. The RMSE analysis is conducted based on two performance indicators that are robot and landmark set position. The evaluation is only focusing on capabilities of the approaches in estimating the positions but not on the execution time (computational complexity). The RMSE is computed at every time interval and it is collected during the simulation. Figures 4 and 5 show the result analysis while Table 2 summarizes the result.

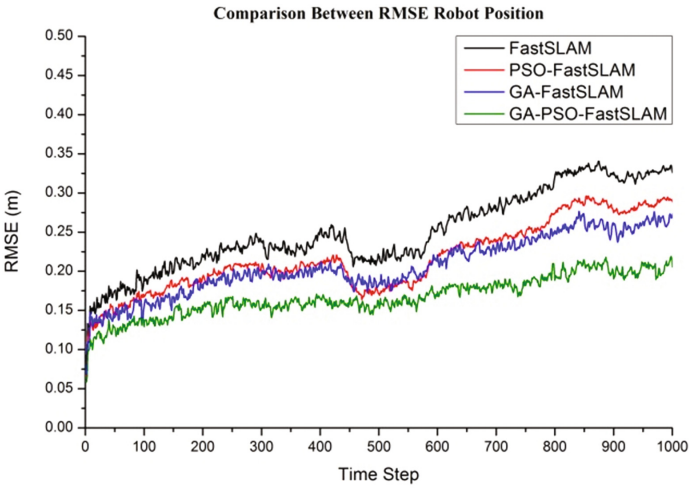


Fig. 4. Robot position RMSE analysis

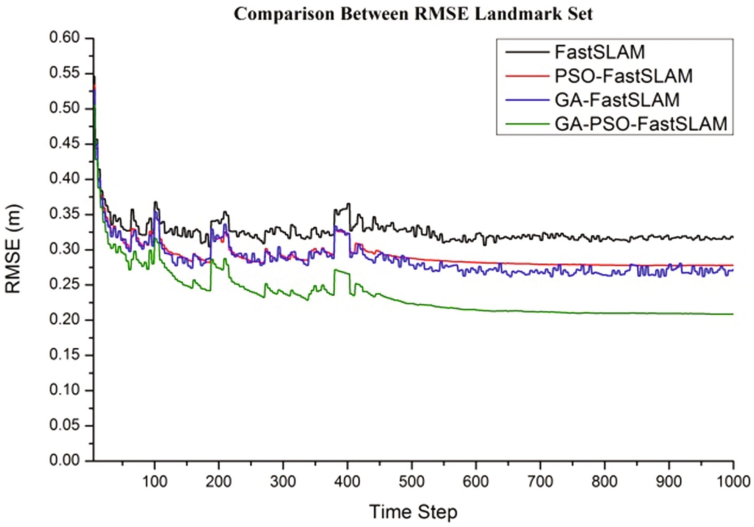


Fig. 5. Landmark set position RMSE analysis

Table 2. Average value of RMSE performance

Approach	RMSE robot position	RMSE landmark set position
FastSLAM	0.2511	0.3253
PSO-FastSLAM	0.2172	0.2919
GA-FastSLAM	0.2069	0.2859
GA-PSO-FastSLAM	0.1673	0.2359

Based on the result and analysis, it shows that the proposed GA-PSO-FastSLAM approach produce a lower RMSE values of robot and landmark set position compared to other approaches. This proves that the proposed approach able to reduce the error occurs in FastSLAM and its associated algorithms for robot and landmark set estimation. Hence, the proposed approach able to reduce the particle depletion problem and maintain the particle diversity.

5 Conclusion

In this work, GA-PSO-FastSLAM is presented that hybridize two soft computing techniques which are genetic algorithm and particle swarm optimization to optimize the FastSLAM performance. The aim is to minimize the particle depletion problem occurs in FastSLAM. The experiment is conducted in simulated environment map with several configuration and setup. The performance of GA-PSO-FastSLAM was evaluated using RMSE analysis and it is compared with four approaches [1, 4, 5]. The result discussed shows an improvement in FastSLAM performance and it is able to minimized the

particle depletion problem. For future work, the improvement of the proposed approach in terms of reducing computational complexity can be done by introducing a new promising solution.

Acknowledgment. This work was supported by Fundamental Research Grant Scheme (FRGS) under grant number R.J130000.7828.4F860 funded by Ministry of Education (MOHE) under the Malaysian government for Faculty of Computing, Universiti Teknologi Malaysia (UTM).

References

1. Montemerlo, M., Thrun, S., Koller, D., Wegbreit, B.: FastSLAM: a factored solution to the simultaneous localization and mapping problem. In: 18th National Conference on Artificial Intelligence, Menlo Park, CA, USA, pp. 593–598 (2002)
2. Montemerlo, M., Thrun, S.: FastSLAM, vol. 27, 1st edn. Springer, Heidelberg (2007)
3. Bailey, T., Nieto, J., Nebot, E.: Consistency of the FastSLAM algorithm. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 424–429 (2006)
4. Xia, Y.-M., Yang, Y.-M.: An improved FastSLAM algorithm based on genetic algorithms. In: Qi, L. (ed.) ISIA 2010. CCIS, vol. 86, pp. 296–302. Springer, Heidelberg (2011). doi:10.1007/978-3-642-19853-3_43
5. Lee, H.-C., Park, S.-K., Choi, J.-S., Lee, B.-H.: PSO-FastSLAM: an improved FastSLAM framework using particle swarm optimization. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 2763–2768 (2009)
6. Gongyuan, Z., Yongmei, C., Feng, Y., Quan, P.: Particle filter based on PSO. In: International Conference on Intelligent Computation Technology and Automation, pp. 121–124 (2008)
7. Zou, Z., Cai, Z., Chen, B.: An improved FastSLAM method based on niche technique and particle swarm optimization. In: Control and Decision Conference, China, pp. 2414–2418 (2011)
8. Choi, M., Sakthivel, R., Chung, W.K.: Neural network-aided extended Kalman filter for SLAM problem. In: Conference on Robotics and Automation, Italy, pp. 1686–1690 (2007)
9. Dai, X.F., Hao, B., Shao, L.: Self-organizing neural networks for simultaneous localization and mapping of indoor mobile robots. In: Intelligent Networks and Intelligent Systems, China, pp. 115–118 (2008)
10. Mahrani, M.: Hybrid genetic algorithm and particle swarm optimization model for simultaneous localization and mapping problem. Ph.D. thesis, Universiti Teknologi Malaysia (2016)
11. Zhang, H., Dai, X.: Soft computing technique for simultaneous localization and mapping of mobile robots. In: International Conference on E-Product, E-Service and E-Entertainment (ICEEE), Henan, China, pp. 1–4 (2010)
12. Bailey, T.: SLAM Simulation Toolbox, 3 April 2015. http://wwwpersonal.acfr.usyd.edu.au/tbailey/software/slam_simulations.htm

Three Case Studies Using Agglomerative Clustering

Rodrigo C. Camargos[1(✉)] and Maria do Carmo Nicoletti[1,2]

[1] Faculdade Campo Limpo Paulista (FACCAMP), Campo Limpo Paulista, SP, Brazil
{rodrigocamargos,carmo}@cc.faccamp.br
[2] Universidade Federal de São Carlos (UFSCar), São Carlos, SP, Brazil

Abstract. Finding a data clustering in a data set is a challenging task since algorithms usually depend on the adopted inter-cluster distance as well as the employed definition of cluster diameter. The work described in this paper approaches a well-known agglomerative clustering algorithm named AGNES (*Agglomerative Nesting*), in regards to its performance on three case studies namely, datasets formed by clusters of different sizes, uneven inter-cluster distances and diameters. Clustering results are evaluated using three well-known indexes, Dunn, Davies-Bouldin and Rand. Results obtained with K-means were used for comparison purposes. The experiments were conducted divided into three case studies. Their results suggest that AGNES and K-means have similar performance as far as identifying clusters with different sizes and inter-cluster distances, however, AGNES obtained the best results when dealing with clusters having both, different sizes and diameters.

Keywords: Unsupervised machine learning · Agglomerative clustering · Case studies in clustering

1 Introduction

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) focused mainly on investigations and proposals of new formalisms and computational algorithms, aimed at proving theoretical support as well as implementing automatic learning by computers. Over the past few decades, many ideas on how to enable automatic learning have been presented, discussed and implemented. Among the many ways ML can be implemented, the so called clustering algorithms are a particular group of (unsupervised) algorithms which aim at organizing sets of data points into groups, having data exploratory processes in sight. In the literature one can find numerous works proposing new clustering algorithms as well as different clustering taxonomies. Although taxonomies aim at organizing such algorithms into categories, due to the fact that usually they adopt different criteria for grouping them, they are not necessarily, compatible to each other (see, for instance, those suggested in [1–4]).

As pointed out in [5], efficient clustering techniques are considered a challenge, mainly due to the fact that there is no external supervision, which implies knowing nothing about the internal structure of point sets (such as spatial distribution, volume, density, geometric shapes of clusters, etc.). In such a scenery automatic learning becomes an exploratory task, aiming at identifying which are the groups of data points that are statistically separable (or not), which are the most obvious clusters and how they relate to what is aimed at to

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_7

discriminate, in an attempt to expose the underlying structure of the data, based only on their descriptions, generally given as a vector of attribute values.

The main focus of this work is an empirical research and evaluation of the AGNES (*AGglomerative NESting*) algorithm [3], taking into account data point sets with different characteristics, cluster sizes and inter-cluster distances. For the experiments a collection of data points was artificially created. Seven data point sets were used for evaluating the performance of AGNES, having the K-means algorithm [6] as baseline. The experiments were organized into three case studies, depending on the characteristics of the data point sets.

The remainder of this paper has four more sections. Section 2 comments on the main characteristics of the agglomerative approach and gives a high-level pseudocode of the AGNES algorithm, which has been implemented and used in the experiments. Section 3 describes the data used in the experiments, organized as three case studies. Section 4 briefly introduces the validation indices Dunn, Davies-Bouldin and Rand, used for evaluating the experiment results, followed by the set of experiments related to each case study, discussing their results and presenting some comparative analysis. Section 5 resumes the work done and highlights some conclusions.

2 AGNES (*AGglomerative NESting*) Algorithm

For a given data set containing N data points to be clustered, agglomerative hierarchical clustering algorithms usually start with N clusters (each single data point is a cluster of its own); the algorithm goes on by merging two individual clusters into a larger cluster, until a single cluster, containing all the N data points, is obtained. Obviously, the algorithm can have another stopping criteria, such as that of ending when a clustering containing a user-defined number of clusters (k) is obtained.

Figure 1 presents a high level pseudocode of the AGNES algorithm which, at each iteration, chooses two clusters to be merged, based on the shortest Euclidean distance between the clusters formed so far. The many clustering agglomerative algorithms found in the literature can be organized taking into account the way the inter-cluster distance is defined i.e., what definition is used to compute the shortest distance between all pairs of clusters. Among the various ways, three are particularly popular and are defined next. Given two clusters X and Y, let d(x,y) denotes the distance between two data points.

(1) *Single Linkage*, the distance between two clusters X and Y is the shortest distance between two data points, $x \in X$ and $y \in Y$, formally represented by Eq. (1).

$$d_{SL}(X, Y) = \min_{x \in X,\, y \in Y} d(x, y) \tag{1}$$

(2) *Complete Linkage*, the distance between two clusters X and Y is the farthest distance between two data points, $x \in X$ and $y \in Y$, formally represented by Eq. (2).

$$d_{CL}(X, Y) = \max_{x \in X,\, y \in Y} d(x, y) \tag{2}$$

(3) *Average Linkage* or *UPGMA*, where the distance between two clusters is the mean distance between data points of each cluster in one cluster to every point in the other cluster, formally represented by Eq. (3).

$$d_{AL}(X, Y) = 1/(|X| \times |Y|) \times (\sum_{x \in X} \sum_{y \in Y} d(x, y)) \quad (3)$$

```
procedure AGNES (X,K,AC_t)
Input: X = {P_1, P_2,...,P_N} % dataset with N patterns
Output: AC_t
begin
t ← 0
AC_t ← {{P1}, {P2},...,{PN}} % initial clustering
Nro_C ← N                    %initial number of clusters
CI ← N+1        % cluster index for new created clusters
repeat
  t ← t + 1
  among all possible pairs of clusters {C_r,C_s} in AC_t-1,
  find one {C_i,C_j} such that g(C_i,C_j) = min_r,s g(C_r,C_s),
  where g is a dissimilarity function (distance)
  New_C ← C_i ∪ C_j % create a new cluster
  Nro_C  ← Nro_C − 1
  C_CI  ← New_G
  CI ← CI +1
  AC_t  ← (AC_t-1 − {C_i,C_j}) ∪ {C_CI}
until Nro_C ≤ K
end
```

Fig. 1. A customized version of AGNES pseudocode, based on [1, 3].

For the experiments described in Sect. 4, AGNES was implemented using the UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*), for inter-cluster distance, where all pair-wise distances contribute equally.

3 Data Description

The experiments described in Sect. 4 had three different focuses of empirical investigations: (1) sizes of clusters; (2) inter-cluster distances and (3) diameter of clusters. For addressing each focus, a corresponding case study was conducted, having two-dimensional point sets prepared according to the particular focus intended. For all the experiments a total of 7 synthetic sets of data points were created.

Case Study I (CS-I:Squares) uses three point sets created having their clusters at different distances between themselves, (a) Square1, (b) Square3 and (c) Square5. The only difference between the three point sets Square1, Square3 and Square5 is the degree of overlap between the four clusters that define each point set. In Square1, the clusters touch each other but hardly overlap, whereas in Square5 the overlap is such that there is little density difference when moving from one cluster to the next, as can be seen in

Fig. 2. Point sets Square1, Square3 and Square5 can be described as square arrangements of four clusters of equal size and spread, each cluster being a Gaussian distribution around a central point; they differ from each other only in relation to the inter-cluster distances. The number of clusters, the size of the clusters and the average and standard deviation vectors of each cluster were previously defined for both point sets: Square (CS-I) and Sizes (CS-II). Square and Sizes were created, based on the ones used in [7], having in mind to investigate the sensitivity of an agglomerative clustering algorithm to the inter-cluster distance as well as the increase of the overlapping between clusters.

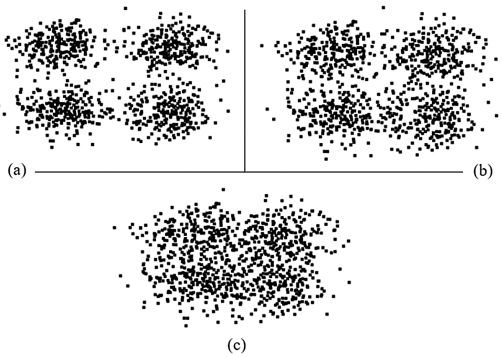


Fig. 2. Case *Study I* - Point sets formed by clusters with different inter-cluster distances (a) Square1, (b) Square3 and (c) Square5.

Case Study II (CS-II: Sizes) also uses three point sets, each created having four clusters. The differences among the three point sets rely on the sizes (number of data points) of their corresponding clusters, (a) Sizes1, (b) Sizes3 and (c) Sizes5, as shown in Fig. 3. The three point sets have been created based on the Square1 (Fig. 2(a)), where there were changes in the relative cluster sizes such that the ratio of the smaller to its immediate larger cluster was 2, 6 and 10, respectively. It is important to notice, though, that the spread of the clusters has been kept constant. Sizes has been created for investigating the sensibility of the algorithm to clusters of different sizes.

Case Study III (CS-III:Aggregation). The seventh point set, named *Aggregation*, (shown in Fig. 4) consists of 7 spherical-shaped clusters, (labeled 1, 3, 4, 5, 6 and 7 in Fig. 4) and 1 non-spherical cluster (labeled 2 in Fig. 4). The point set has been created based on the one used in [8] for experiments related to the clustering aggregation problem. Such set of points has characteristics that are known to create difficulties for many agglomerative algorithms, such as narrow "bridges" between clusters, uneven-sized clusters, clusters with different diameters, etc. On the one hand, agglomerative clustering algorithms based on the nearest neighbor, such as *single linkage*, tend to join groups that touch each other, such as the clusters 3 and 6 as well as 4 and 7 in Fig. 4. On the other hand, agglomerative clustering algorithms based on furthest neighbor, such as *complete linkage*, tend to break large clusters (such as cluster 4 in Fig. 4), based on

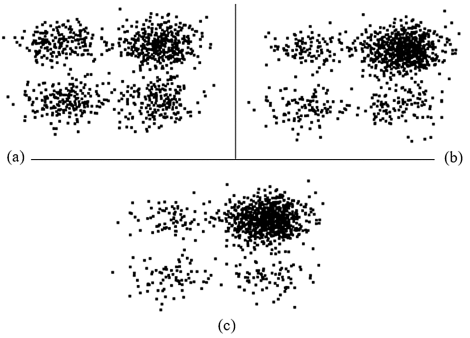


Fig. 3. Case *Study II* - Point sets used in the experiments, formed by clusters of different sizes (number of points). (a) Sizes1, (b) Sizes3 and (c) Sizes5.

the diameters of the small ones, such as clusters 1, 5 and 7. Table 1 presents a summary of the point sets involved in the experiments, describing their basic characteristics.

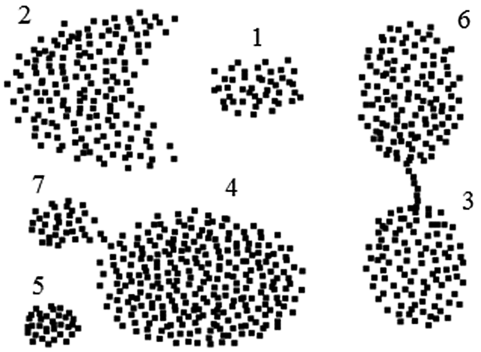


Fig. 4. Case *Study III* - Point set (*Aggregation*) formed by clusters with different diameters. Each cluster has been labeled for reference.

Table 1. Summary of the seven synthetic point sets. #NP: no. of points (size), #NC: no. of clusters. (*)No. of points taking into account the label ordering in the *Aggregation* data.

Point sets	#NP	#NC	Sizes of groups
Square1	1000	4	250-250-250-250
Square3	1000	4	250-250-250-250
Square5	1000	4	250-250-250-250
Sizes1	1000	4	400-200-200-200
Sizes3	1000	4	667-111-111-111
Sizes5	1000	4	769-77-77-77
Aggregation	788	7	45-170-102-273-34-130-34(*)

4 Validation Indices, Experiments, Results and Analysis

This section presents the empirical results obtained by AGNES algorithm [3], considering the three case studies described in the previous section, involving seven data point sets. For the experiments described in this section, AGNES and K-means [6] were implemented in C# and run under a Microsoft Windows platform. The K-Means algorithm was used for comparative purposes only. The results from running both algorithms using the same input data are presented in the following three tables, where the best results are bold faced. To make a fair comparison the number of clusters (K) supplied by the user, to both algorithms, was the same. It is important to mention that all instances in the seven point sets were previously labeled (i.e., has a class attribute associated) in order to allow for conducting external validation. The inherent separation easily perceived between the clusters in the point sets was employed for class assignment.

The results obtained were evaluated using two internal validation indexes: the Dunn's index (D) [9] and the Davies-Bouldin index (DB) [10]. To quantify the number of data points incorrectly assigned (using the original classes previously assigned, and part of the description of each data point in all the seven point sets), the Rand index (R) [11] external validation index was also used.

Consider the following notation. S - point set clustered into clustering C_1, C_2,…, C_{NC}; |S| - number of data points in S; C_i - ith cluster of the clustering; n_i - number of data points in C_i; E_x, E_y - two data points, and let the distance between the two data points, E_x, E_y, be represented as dist(E_x, E_y). The Dunn's index (D) is defined by Eq. (4).

$$\begin{gathered} D = \min_i \left\{ \min_j (A/B) \right\}, \text{where} \\ A = \min_{Ex \in Ci, Ex \in Cj} \operatorname{dist}\left(E_x, E_y\right) \text{ and } B = \max_k \{ \max_{Ex,Ey \in Ck} \operatorname{dist}\left(E_x, E_y\right) \} \end{gathered} \tag{4}$$

For defining the Davies-Bouldin index (DB) as in [10], first consider s_i be a measure of dispersion of a cluster C_i (*i.e.*, a measure of its spread around its mean vector) and let $d(C_i,C_j) = d_{ij}$ be the dissimilarity between two clusters, using an appropriate dissimilarity measure (*e.g.*, distance). Consider the similarity index R_{ij}, between C_i and C_j be given by: $R_{ij} = (s_i + s_j)/d_{ij}$, provided that d_{ij} is symmetric. Let R_i be defined as $R_i = \max_{j=1,..,NC, j \neq i} R_{i,j}$, i = 1,…, NC. Then, the Davies-Bouldin index DB is defined by Eq. (5).

$$DB = 1/NC \times \sum_{i=1,\ldots NC} R_i \tag{5}$$

In data clustering the value of the Rand index [11] can be approached as a measure of the similarity between two data clusterings. For the experiments presented next, one of the clusterings will be the one induced by the clustering algorithm and the other, the one provided externally, by previously assigning a class to each data point, in each of the seven point sets considered. So, in a general setup, given the notation previously introduced and considering that one of the clustering of the point set S is given as $O = \{O_1, O_2, \ldots, O_{NO}\}$ and the other as $C = \{C_1, C_2, \ldots, C_{NC}\}$, the Rand index is composed by the following values:

(i) a: number of pairs of points in S that are in the same set in O and in the same set in C;
(ii) b: number of pairs of points in S that are in different sets in O and in different sets in C;
(iii) c: number of pairs of points in S that are in the same set in O and in different sets in C and
(iv) d: number of pairs of points in S that are in different sets in O and in the same set in C. So the Rand index (R) is given by Eq. (6).

$$R = (a + b)/(a + b + c + d) \tag{6}$$

Intuitively (a + b) can be thought of as the number of agreements between O and C and (c + d) as the number of disagreements between O and C.

In the following tables, (+) and (−) report the best and the worst result of K-Means, respectively. As far as the point sets of *CS-I:Squares* are concerned, the experiment results shown in Table 2 indicate that both algorithms, AGNES and K-Means had similar performance. It is important to mention, however, that AGNES performed slightly better than K-means when taking into account the K-Means worst-case results. Nevertheless, considering the CS-I results, AGNES and K-Means algorithms share similar performances when considering clusters with uneven-sizes and different inter-cluster distances.

Table 2. Clustering results for case study *CS-I:Squares* – Point set in Fig. 2.

Clustering algorithm	Point set	D	DB	R
AGNES	Square1	**0.06**	**0.42**	**0.974**
AGNES	Square3	0.03	0.55	0.934
AGNES	Square5	0.02	0.79	0.789
K-means (+)	Square1	**0.05**	**0.40**	**0.980**
K-means (−)	Square1	0.05	0.42	0.980
K-means (+)	Square3	0.01	0.52	0.950
K-means (−)	Square3	0.01	0.52	0.947
K-means (+)	Square5	0.01	0.61	0.887
K-means (−)	Square5	0.01	0.58	0.880

The numbers obtained from experiments in *CS-II:Sizes* shown in Table 3 suggest that, as the size (i.e., number of points) of the four clusters change, AGNES and the K-Means performances also change.

The three validation indexes used to evaluate the quality of clusterings induced by AGNES suggest different outcomes. The values of the D index implies that AGNES achieved its best performance on the Sizes5 data point set; however, the DB and the R indexes suggest that the best results were obtained when running AGNES on Sizes1. Taking into account only clustering results from AGNES, its best performance was achieved having Size1 as input. Comparing AGNES and K-means results obtained in CS-II:Sizes, it is obvious that K-Means had the best performance as far as the values of indexes DB and R are concerned.

Table 3. Clustering results for case study *CS-II:Sizes* – Point set in Fig. 3.

Clustering algorithm	Point set	D	DB	R
AGNES	Sizes1	0.04	**0.43**	**0.974**
AGNES	Sizes3	0.03	0.51	0.936
AGNES	Sizes5	**0.05**	0.44	0.963
K-means (+)	Sizes1	0.03	**0.40**	**0.982**
K-means (−)	Sizes1	0.03	0.47	0.972
K-means (+)	Sizes3	0.03	0.43	0.972
K-means (−)	Sizes3	0.03	0.47	0.972
K-means (+)	Sizes5	0.02	0.46	0.966
K-means (−)	Sizes5	0.00	0.69	0.499

Case Study III (*CS-III:Aggregation*) focuses on clustering experiments having, as input, the *Aggregation* point set, shown in Fig. 4. The values of the three validation indexes applied to the clustering results obtained from several experiments using *Aggregation* as input are presented in Table 4.

The experiments were conducted following a different methodology than the one used in the previous two case studies. It was decided to initially consider *Aggregation* as having only clusters 1 and 6 and then, gradually grow the point set by adding to it some of its clusters, until reaching the 7 total clusters it effectively has. By doing so it was expected that the experiment would allow to investigate what point set configuration would have more impact on AGNES performance.

The *Aggregation* point set is initially formed by clusters 1 and 6; both clusters are different with regard to their sizes, diameters and shapes, but are well separated and their inter-cluster distance promotes a good performance of both, AGNES and K-means. The values of the Rand index in Table 4 suggest that both algorithms correctly identified clusters 1 and 6. Next, the clustering results shown in Table 4 are from running AGNES and K-means on the previous *Aggregation* point set, added with cluster 3. Such addition did not decrease AGNES performance, in spite of the "bridge" between clusters 3 and 6. The K-Means, however, failed to identify some points from cluster 3. In sequence, the *Aggregation* was modified again, by replacing cluster 3 and 6 with cluster 2. In spite of cluster 1 and 2 having different shapes, which eventually could create difficulties for some clustering algorithms, their differences in shape were not substantial enough and AGNES identify both correctly. The K-Means, however, could not fully identify points from cluster 2 due to their distance from the cluster's centroid.

The next modification of *Aggregation*, at this point having cluster 1 and 2, was done by adding to it the cluster 6. AGNES was able to correctly detect the three clusters (1, 2 and 6) while K-Means failed again for identifying points in cluster 2. *Aggregation* suffered another addition, this time of cluster 3. With this configuration both algorithms, AGNES and K-Means, did not identify correctly the four clusters; yet, AGNES was the one that came closer. Next, by adding cluster 4 to the previous *Aggregation*, the K-Means still had trouble to fully detect all clusters, while AGNES recovered its optimal performance, since with the addition of cluster 4, the average distance between the five clusters changed. Moreover, it can be easily noticed that, by adding cluster 5 to *Aggregation*,

Table 4. Clustering results for case study *CS-III:Aggregation* – Point set in Fig. 4.

Clustering algorithm	Point set	D	DB	R
AGNES	Clusters 1 and 6	0.35	0.48	1
K-means (+)	Clusters 1 and 6	0.35	0.48	1
K-means (−)	Clusters 1 and 6	0.35	0.48	1
AGNES	Clusters 1, 3 and 6	0.04	0.33	**1**
K-means (+)	Clusters 1, 3 and 6	0.04	**0.32**	0.998
K-means (−)	Clusters 1, 3 and 6	0.04	**0.32**	0.998
AGNES	Clusters 1 and 2	**0.04**	0.33	**1**
K-means (+)	Clusters 1 and 2	**0.04**	**0.32**	0.998
K-means (−)	Clusters 1 and 2	0.08	0.54	0.972
AGNES	Clusters 1, 2 and 6	**0.33**	**0.33**	**1**
K-means (+)	Clusters 1, 2 and 6	0.08	0.34	0.993
K-means (−)	Clusters 1, 2 and 6	0.08	0.35	0.989
AGNES	Clusters 1, 2, 3 and 6	0.04	0.41	**0.993**
K-means (+)	Clusters 1, 2, 3 and 6	0.04	**0.36**	0.989
K-means (−)	Clusters 1, 2, 3 and 6	0.04	**0.36**	0.989
AGNES	Clusters 1, 2, 3, 4 and 6	**0.04**	0.39	**1**
K-means (+)	Clusters 1, 2, 3, 4 and 6	0.03	**0.38**	0.997
K-means (−)	Clusters 1, 2, 3, 4 and 6	0.02	0.51	0.900
AGNES	Clusters 1, 2, 3, 4, 5 and 6	**0.04**	**0.34**	**0.998**
K-means (+)	Clusters 1, 2, 3, 4, 5 and 6	0.03	0.41	0.928
K-means (−)	Clusters 1, 2, 3, 4, 5 and 6	0.03	0.49	0.889
AGNES	All 7 clusters	**0.04**	**0.33**	**0.998**
K-means (+)	All 7 clusters	0.03	0.44	0.927
K-means (−)	All 7 clusters	0.03	0.50	0.919

the average cluster distances computed by AGNES was negatively affected. In spite of that, AGNES still had a much better performance than K-means. Finally, *Aggregation* was restored to its original seven clusters, as in Fig. 4. Notice that the addition of cluster 7 did not decrease AGNES performance and the clustering it induced was very close to the optimal result, as confirmed by its Rand index of 0.998.

5 Conclusions

This paper addresses the use of the agglomerative hierarchical clustering algorithm AGNES, in unsupervised tasks involving 7 point sets, grouped into three case studies: the *CS-I:Squares*, which focuses on inter-cluster distance, the *CS-II:Sizes*, with focus on clusters of different sizes and the *CS-III:Aggregation*, involving mainly different shapes and diameters. As far as both case studies, *CS-I:Squares* and *CS-II:Sizes*, are concerned, AGNES and K-Means results were similarly evaluated; however, in the *CS-I:Squares* AGNES had a slightly better performance than K-means when taking into account K-Means worst-case results. In the third case study, the *CS-III:Aggregation*, in

most experiments, the best results were obtained with AGNES. In spite of AGNES being more strongly affected by the inter-cluster distance than any of the other chosen characteristics, such as size, diameter or shapes (except for elongated clusters where *single linkage* based algorithms usually have better performance), the algorithm was still very robust considering a combination of all the chosen characteristics.

Acknowledgments. The authors would like to thank CAPES, CNPq and FACCAMP.

References

1. Theodorides, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Elsevier, Amsterdam (2009)
2. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. **31**(3), 264–323 (1999)
3. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, Hoboken (1990)
4. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann-Elsevier, Amsterdan (2012)
5. Aggarwal, C.C., Reddy, C.K. (eds.): Data Clustering: Algorithms and Applications. Chapman and Hall/CRC, Boca Raton (2013)
6. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, vol. 1, pp. 281–297 (1967)
7. Handl, J., Knowles, J.: Evolutionary multiobjective clustering. In: Yao, X., et al. (eds.) PPSN 2004. LNCS, vol. 3242, pp. 1081–1091. Springer, Heidelberg (2004). doi: 10.1007/978-3-540-30217-9_109
8. Gionis, A., Manilla, H., Tsaparas, P.: Clustering aggregation. ACM Trans. Knowl. Disc. Data **1**, 30 p. (2007). Article 4
9. Dunn, J.: Well separated clusters and optimal fuzzy partitions. J. Cybern. **4**, 95–104 (1974)
10. Davies, D.L., Boudin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-1**, 224–227 (1979)
11. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**, 846–850 (1971)

Historic Document Image De-noising Using Principal Component Analysis (PCA) and Local Pixel Grouping (LPG)

Han-Yang Tang, Azah Kamilah Muda(✉), Yun-Huoy Choo, Noor Azilah Muda, and Mohd Sanusi Azmi

Computational Intelligence and Technologies Lab (C.I.T Lab), Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia
hanyang14@gmail.com, {azah,huoy,azilah,sanusi}@utem.edu.my

Abstract. In this paper, an approach of principal component analysis (PCA) with local pixel grouping (LPG) is used to de-noising the noisy historical document image. This technique ensures the preservation of historic document image local structure. This is due to block matching based LPG which carries out classification to allow only the sample blocks with similar contents used in the calculation for PCA transform estimation. Such an LPG procedure ensures that the image local features can be well preserved after the noise removing process in the PCA domain. The LPG-PCA de-noising procedure will repeat one more times with adaptively adjusted noise level to further improve the performance of de-noising the historic document image. The experiment results show that LPG-PCA model has good results in de-noising historical document image.

Keywords: Historical documents · Local pixel grouping (LPG) · Principal component analysis (PCA)

1 Introduction

Historic document is very important to preserve the originality of an information. It needs to be transformed into computer readable format. However, a lot of the historic document is blurry during the process of scanning, transmission or conversion the document into digital form, the images may be contaminated with defects. According to definition suggested by Baird [1]: "By degradation (or defects), we mean every sort of less than ideal properties of real document images". Just to name a few, marginal noise which appears in a dark region around the document image, salt and pepper noise that appear in document image caused by dirt on the document, background noise happens when the document images have degradations like show through effects, interfering strokes, background spots and uneven contrast. Therefore, an appropriate efficient de-noising method is needed to remove the noise from historical document images and improve their quality.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_8

In this paper we concentrate on de-noising historical document image with background defect which generally called show through or bleed through or interference strokes effect. This degradation is due to ink's seeping through the paper of historic documents after stored for a long time. The result is that content from the back side appear as noise on the front side which will drastically reduce the readability of the documents if the interference acts in a significant way.

A proposed solution to this problem is a two stage Principal Component Analysis (PCA) based de-noising method with local pixel grouping (LPG). This method has two stages. The first stage will produce an initial estimation of the image by removing most of the noise of the document images while the second stage will further improve the output from the first stage. These two stages have almost same procedure except the parameter of the noise level. The main idea of this method is the LPG process during the second stage will improved accuracy due to the significant reduce of noise in the first stage. The result is a final de-noising result that is visually much better than the original document image.

The rest of paper is organized as follows. Section 2 presents related works on denoising historic document. Section 3 presents the LPG-PCA denoising in detail while Sect. 4 outlines the experimentation process and performance measurement. Section 4.3 depicts the result while Sect. 5 draws the conclusion.

2 Related Work

Many methods have proposed by other researchers on handling background degradation in document image that can separated into five major groups:

1. Thresholding based method
2. Histogram based method
3. Morphology based method
4. Fuzzy Logic based method
5. Genetic Algorithm based method

Literature shows that thresholding techniques have a very high preference from researchers in handling the degraded historical document image due to its simplicity and high performance. Thresholding method can divided into two main categories, which are Global thresholding method and local thresholding method. Global thresholding technique is found by Otsu in 1979 [2]. However, if the document illumination or contrast is not evenly throughout the document, global thresholding will not provide good performance and an adaptive thresholding technique need to be employed. Mitra [3] proposes a new optimum thresholding technique which performs better than the well known Otsu's thresholding technique. He shows the importance of pre-processing the document image before it is subjected to thresholding.

Histogram based method is a popular algorithm for contrast enhancement on historical document image due to its effectiveness and simplicity [4]. Histogram equalization can be categorized into 2 methods: global and local histogram equalization. Global histogram equalization is simple and fast but its contrast enhancement power is relatively

low. While local histogram equalization can enhance overall contrast more effective, but required a very high computation complexity due to its fully overlapped sub-blocks. Partially Overlapped Sub-Block Histogram Equalization (POSHE) was proposed by Kim et al. [5]. In this method, the image is divided into blocks then in each block histogram equalization is done. This method achieves better performance in contrast enhancement because of it use local feature extraction.

Morphological based method use mathematical morphology to enhance uneven backgrounds. The operator is a great tool to analyze and process shapes with structural features like area, borders. This method search for nose patterns which appear as a shadow in background with defined structuring elements. Then, thickening algorithm will thicken the objects in a noisy binary image. Followed by pruning algorithm properly clean up the thickened to obtain image without spurs. Nomura et al. [6] has proposed the Shadow Location and Lightening method. This method use thickening to highlight features that cause the shadow in images, then use pruning to remove the shadow.

Next is Fuzzy Logic based method which enhancing image quality by using fuzzy logic operators which is based on mapping grey levels of image to fuzzy space, and defining an appropriate membership function requires prior knowledge and experience. Image enhancement with fuzzy operators employs weighting features proportional to some image features, like average intensity to increase contrast. The aim is to generate an image of higher contrast than the original image. Tizhoosh [7] proposed a fuzzy approach to image enhancement by using a contrast intensification operator which can improve the contrast in the image. After that, Parvathi et al. [8] proposed another method to improve the Tizhoosh's algorithm. An improved algorithm is proposed where it is used an intensification operator on the first and second types of Intuitionist Fuzzy Sets (IFSs).

Genetic Algorithm based method is one of the optimization or searching algorithm which categorized under Evolutionary Algorithm. Background can has complex and variety in degradation. It is difficult to separate characters from the background (binarization). In order to improve the binarization, a well-known filter such as Fourier transform, Gabor filter and wavelet transform can be used. However, it is difficult for a single filtering technique to deal with a variety of degradation. Kohmura [9] has proposed a Genetic Algorithm approach to construct an optimal sequence of image in order to extract characters from different source of color images. Deborah et al. [10] has proposed Genetic Algorithm to estimate the degradation function of an image. The degradation model has a degradation function that together with an additive noise to operate on an input image in order to produce a degraded image. However, the optimization procedure using Genetic Algorithm is rather slow because every fitness evaluation requires the comparison of two images.

3 LPG-PCA De-noising Algorithm

3.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) can identify patterns in the data where similarities and differences are highlighted [11]. By transforming the original dataset into PCA

domain and preserving only the several most significant principal components, the noise and unimportant information can be removed. PCA commonly used as a dimensionality reduction technique which highlights the dominant parts while diminishing the noise in pattern recognition. In [11], a PCA based method was proposed for image de-noising by using a moving window to calculate the local statistics, from which the local PCA transformation matrix was estimated. However, this method applies PCA directly to the noisy image without data selection which caused many noise remained in the de-noised outputs.

PCA will express data in more compact and meaningful form and efficiently choose the most relevant information into lower dimensional space and the less important information will be represented by a larger number of basis vectors. According to this, PCA can be viewed as represent the relevant data as linear. Let $X = [x_1, x_2, x_{3\ldots\ldots} x_m]^T$, an m-component vector variable. The sample matrix of X shown as follows:

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \ldots & x_1^n \\ x_2^1 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \vdots & \vdots \\ x_m^1 & x_m^2 & \ldots & x_m^n \end{bmatrix} \tag{1}$$

Where x_i^j, where $j = 1, 2 \ldots . n$, represent the discrete sample of variable x_i, where $i = 1, 2 \ldots . m$.

The mean of a row is calculated as follows:

$$\mu_i = \frac{1}{n} \sum_{j-1}^{n} x_i(j) \tag{2}$$

This can compute the average value of pixel. The centralized matrix of x_i can be written as $\bar{x}_i = x_i - \mu_i = [\bar{x}_i^1 \bar{x}_i^2 \cdots \bar{x}_i^n]$ where $\bar{x}_i^j = x_i^j - \mu_i$. Finally, we calculate the covariance matrix of centralized data set by using the formula

$$C_x = \frac{1}{n} \overline{xx}^T \tag{3}$$

The goal of PCA is to generate an orthonormal transformation matrix P to de-correlate the centralized matrix. For example, $\bar{Y} = P\bar{X}$ to make the covariance matrix $\bar{Y}$ will become diagonal. Since the covariance matrix, C_x is symmetrical, it written as $C_x = \beta_{\bar{Y}} \wedge_{\bar{Y}} \beta_{\bar{Y}}^T$. The diagonal matrix $\wedge$ contains the eigenvectors of C with decreasing order. The matrix β represents the orthonormal matrix such that $P = \beta^T$. Therefore, the diagonal matrix shown as follows:

$$\wedge = (\frac{1}{n}) \overline{YY}^T \tag{4}$$

Therefore, the signal and noise in the document image can better differentiate in the PCA domain (Fig. 1).

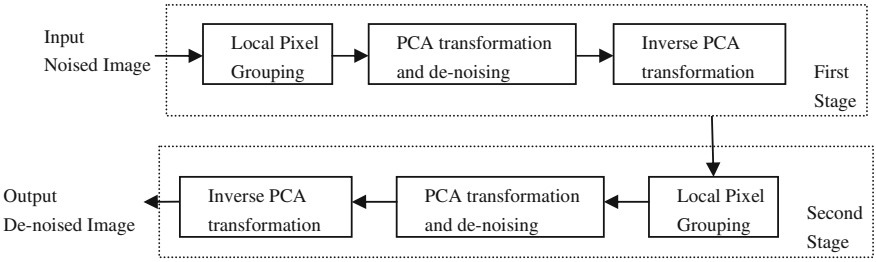


Fig. 1. The two stage LPG-PCA de-noising scheme

3.2 Local Pixel Grouping (LPG)

The intensity value of a pixel can be partially correlating by a smaller number of pixels surrounding it. Hence, the dependency of pixel is creates more locally than globally. An individual independent grouping of pixels provides better noise removal among the pixels. Hence, by looking for a potential source group, we will be able to find out which of the group individual pixels belong to the original image.

Due to this, local pixel grouping will select $K \times K$ blocks as a kernel from the given $L \times L$ training block, as shown in Fig. 2 reprinted from [17] authored by C.P Kumar at year 2004. For an underlying pixel to be de-noised, we set a $K \times K$ window centered on it and denote by $x = [x_1 \cdots x_m]^T$ and $m = K^2$, the vector containing all the components within the window. The noisy vector of the document image has shown as $x_v = \mathrm{x} + v$, where v is the additive noise adding to the original image and show as $v = [v_1 \cdots v_m]^T$.

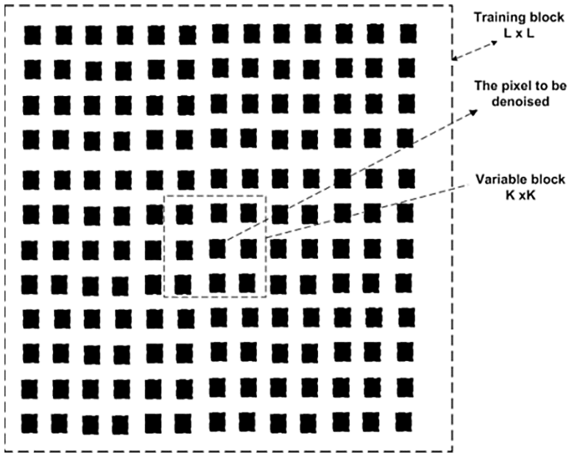


Fig. 2. Local pixel grouping [17]

The total number of training sample is $(L - K + 1)^2$. However, there can be very different blocks of given central $K \times K$ blocks in the $L \times L$ training window. If we take all the $K \times K$ samples as the training samples, it will lead to inaccurate estimation of the

PCA transformation matrix and will result a worst output image that have more noise. Therefore, in order to select and group a training sample that similar to the central $K \times K$ block, block matching method is being employ which are simplest yet very efficient method compared to the other classification method like correlation based matching, K-mean clustering and etc. [12].

3.3 LPG-PCA Based De-noising First Stage

In this stage, we will first obtain the centralized matrix of X_k which shown as $\bar{x}_k = x_k - \mu_k$ and $\bar{x}_v = \bar{x} + v$. With the linear model $\bar{x}_v = \bar{x} + v$, we can calculate the covariance matrix $C_{\bar{x}_v} = \frac{1}{n}\bar{x}_v\bar{x}_v^T = \frac{1}{n}\left(\overline{xx}^T + \bar{x}v^T + v\bar{x}^T + vv^T\right)$. In this case, $\bar{x}$ and v are uncorrelated, and thus $C_{\bar{x}_v} = \frac{1}{n}(\overline{xx}^T + vv^T) = C_{\bar{x}} + C_v$ where $C_{\bar{x}} = (\frac{1}{n})\overline{xx}^T$ and $C_v = \left(\frac{1}{n}\right)vv^T$. Next we can decompose $C_{\bar{x}} = \beta_{\bar{x}} \wedge_{\bar{x}} \beta_{\bar{x}}^T$ where $C_{\bar{x}}$ is the m x m orthonormal eigenvector matrix is and $\wedge_{\bar{x}}$ is the diagonal eigenvalue matrix. Since $C_{\bar{x}}$ is an orthonormal matrix, we can write C_v as $C_v = \left(\sigma^2 I\right)\beta_{\bar{x}}\beta_{\bar{x}}^T = \beta_{\bar{x}}C_v\beta_{\bar{x}}^T$. Thus, we will have $C_{\bar{X}_v} = \beta_{\bar{x}}\left(\wedge_{\bar{x}} + \sigma^2 I)\beta_{\bar{x}}^T\right) = \beta_{\bar{x}} \wedge_{\bar{x}} \beta_{\bar{x}}^T$.

Since the eigenvector of $C_{\bar{x}}$ and $C_{\bar{X}_v}$ are same, we can directly compute on $\beta_{\bar{x}}$ by decomposing $C_{\bar{X}_v}$ instead of $C_{\bar{x}}$. Then the orthonormal PCA transformation matrix for $C_{\bar{x}}$ is set as $P_{\bar{x}} = \beta_{\bar{x}}^T$. Then we applying $P_{\bar{x}}$ to dataset $\bar{x}_v$, we will have $\bar{Y}_u = P_{\bar{x}}\bar{x}_u = P_{\bar{x}}\bar{x} + P_{\bar{x}}v = \bar{Y} + v_Y$. Where $\bar{Y} = P_{\bar{x}}\bar{x}$ is the de-correlated dataset for $\bar{x}$ and $v_Y = P_{\bar{x}}v$ is the transformed noise dataset for v. Since $\bar{Y}$ and v_Y noise are uncorrelated, we can easily derive that the covariance matrix of $\bar{Y}_u$ to be $C_{\bar{y}_u} = \frac{1}{n}\bar{Y}_u\bar{Y}_u^T = C_{\bar{Y}} + C_{u_Y}$. Hence, in the PCA transformed domain, it will preserving only the several most significant principal components, the noise and unimportant information can be removed.

3.4 LPG-PCA Based De-noising Refinement in Second Stage

In this stage, we will obtain a de-noised image in which most of the noise is removed using the proposed LPG-PCA de-noising procedure in first stage. However we can still see some noise remaining in the de-noising output. There are mainly two reasons which weaken the de-noising performance. First is because the strong noise in the original dataset will leads to LPG errors which will results in estimation bias of the covariance matrix. Second is because of estimation bias of PCA transformation matrix which also caused by strong noise in original dataset. Therefore, it is necessary to further process the de-noising output from the first stage for a better noise reduction output. Since the noise has been much removed in the first stage of LPG-PCA de-noising, the problem we stated before can be solved. The estimation of covariance matrix and LPG accuracy process will improve in the following LPG-PCA de-noising procedure of the second round.

4 Experimentation

4.1 Datasets

In order to evaluate the proposed approach, several samples of degraded image documents have been taken as datasets from digital library of Bibliotheque Nationale de France which is the national library of France. These document images were written in English and mostly have some background degradation. The most common degradation of the image document is shown through or bleeds through noise.

4.2 Assessment Metrics

The quality of an image can be examined objectively as well as subjectively. For subjective evaluation, the image has to be observed by a human expert [13]. However the human visual system cannot do pixel by pixel evaluation of a given image, this causing the exact quality of image is difficult to determine. Generally, the objective quality evaluation methods are faster and more cost effective than the subjective methods. The evaluation method we used for objective evaluation is Peak signal to noise ratio (PSNR) and Structural Similarity index (SSIM).

Peak Signal noise ratio is non-distance pixel based evaluation measurement which is given in decibel value (dB). PSNR can measure the intensity between two images by measuring the ratio of the peak signal to noise and the difference between two images. This ratio is often used as a quality measurement between the original and enhanced image. The higher the PSNR value, the better the quality of the enhanced image.

Structural Similarity index proposed in [14] is one of the most commonly use method to measure the image visual quality assessment. SSIM can reflect the structure similarity between target image and reference image and measuring the similarity between these two images.

Other than PSNR and SSIM, OCR is also included in the experiment. The OCR software used for the comparison was FREE OCR 5.41 which is a Windows OCR program including a Windows compiled Tesseract free OCR engine. The accuracy metric used to rank the performance of proposed method is based on OCR Accuracy [15]. It should be noted that the OCR accuracy is equal to 100 if perfect match is achieved between the texts produced by the OCR and the ground truth text. The OCR accuracy is as follows:

$$OCR_Accuracy = 100\,(\frac{Correct\ characters}{Ground\ truth\ characters}) \tag{5}$$

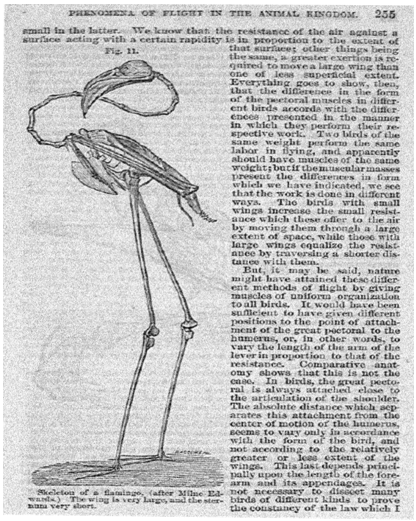
Where Correct characters = Ground Truth characters – (Insertion + Substitutions + Deletions).

4.3 Experimental Result

In this section, we will first verify the noise removal result of the proposed LPG-PCA method. Table 1 lists the PSNR and SSIM measures of the noisy image, first stage and second stage de-noising output. We can see that most of the noise is removed after the first stage because the PSNR measures are much improved. However, for some image the second stage is not improved much at PSNR measures, the SSIM measures, which can better reflect the image visual quality, have much improved after the second stage de-noising refinement (Fig. 3).

Table 1. PSNR (dB) and SSIM results of noisy image and the de-noised images of proposed first stage and second stage LPG-PCA method

	Noisy image PSNR and SSIM result	First stage PSNR and SSIM result	Second stage PSNR and SSIM result
Image1	22.1102 (0.6513)	26.7199(0.8436)	26.9315(0.8775)
Image2	22.1102 (0.4726)	29.3435 (0.8200)	29.6428 (0.8633)
Image3	22.1102 (0.4615)	29.2487 (0.8115)	29.5502 (0.8542)
Image4	22.1102 (0.6186)	27.3067 (0.8388)	27.3461 (0.8574)
Image5	22.1102 (0.7086)	26.5001 (0.8711)	26.5301 (0.8807)
Image6	22.1102 (0.6803)	26.7094 (0.8694)	26.7725 (0.8844)
Image7	22.1102 (0.5221)	28.8438 (0.8429)	29.1598 (0.8827)
Image8	22.1102 (0.4305)	29.8014 (0.8107)	30.1882 (0.8634)
Image9	22.1102 (0.5009)	29.0114 (0.8478)	29.3662 (0.8971)
Image10	22.1102 (0.5712)	28.0957 (0.8469)	28.3061 (0.8712)


PHENOMENA OF FLIGHT IN THE ANIMAL KINGDOM. 255

small in the latter. We know that the resistance of the air against a surface acting with a certain rapidity is in proportion to the extent of that surface; other things being the same, a greater exertion is required to move a large wing than one of less superficial extent. Everything goes to show, then, that the difference in the form of the pectoral muscles in different birds accords with the differences presented in the manner in which they perform their respective work. Two birds of the same weight perform the same labor in flying, and apparently should have muscles of the same weight; but if the muscular masses present the differences in form which we have indicated, we see that the work is done in different ways. The birds with small wings increase the small resistance which these offer to the air by moving them through a large extent of space, while those with large wings equalize the resistance by traversing a shorter distance with them.

But, it may be said, nature might have attained these different methods of flight by giving muscles of uniform organization to all birds. It would have been sufficient to have given different positions to the point of attachment of the great pectoral to the humerus, or, in other words, to vary the length of the arm of the lever in proportion to that of the resistance. Comparative anatomy shows that this is not the case. In birds, the great pectoral is always attached close to the articulation of the shoulder. The absolute distance which separates this attachment from the center of motion of the humerus, seems to vary only in accordance with the form of the bird, and not according to the relatively greater or less extent of the wings. This last depends principally upon the length of the forearm and its appendages. It is not necessary to dissect many birds of different kinds to prove the constancy of the law which I

Fig. 11.

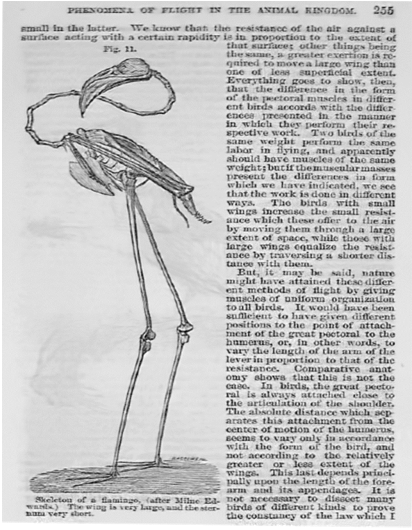
Skeleton of a flamingo, (after Milne Edwards.) The wing is very large, and the sternum very short.

Fig. 3. Image10 noisy image

We then compare the OCR accuracy. Table 2 lists the OCR accuracy of the noisy image, the first stage de-noise image and the second stage de-noise image. We can see that the result shows that there is much better OCR accuracy after the first stage compared to the noisy image. The second stage de-noise image also shows a little improved on the OCR accuracy compared to the first stage (Fig. 4).

Table 2. OCR accuracy results of noisy image and the de-noised images of proposed first stage and second stage LPG-PCA method

	Noisy image OCR accuracy result	First stage OCR accuracy result	Second stage OCR accuracy result
Image1	65.3186%	76.2881%	76.6759%
Image2	19.7107%	42.6763%	46.8354%
Image3	17.1748%	33.9161%	35.3147%
Image4	18.6539%	30.3408%	32.2925%
Image5	40.3475%	55.8531%	57.3495%
Image6	41.2592%	49.4358%	53.2349%
Image7	19.3485%	41.4583%	44.3496%
Image8	20.3742%	40.5735%	45.3862%
Image9	18.1353%	31.3486%	36.5862%
Image10	18.3429%	30.9797%	35.4326%


PHENOMENA OF FLIGHT IN THE ANIMAL KINGDOM. 255

small in the latter. We know that the resistance of the air against a surface acting with a certain rapidity is in proportion to the extent of that surface; other things being the same, a greater exertion is required to move a large wing than one of less superficial extent. Everything goes to show, then, that the difference in the form of the pectoral muscles in different birds accords with the differences presented in the manner in which they perform their respective work. Two birds of the same weight perform the same labor in flying, and apparently should have muscles of the same weight; but if the muscular masses present the differences in form which we have indicated, we see that the work is done in different ways. The birds with small wings increase the small resistance which these offer to the air by moving them through a large extent of space, while those with large wings equalize the resistance by traversing a shorter distance with them.

But, it may be said, nature might have attained these different methods of flight by giving muscles of uniform organization to all birds. It would have been sufficient to have given different positions to the point of attachment of the great pectoral to the humerus, or, in other words, to vary the length of the arm of the lever in proportion to that of the resistance. Comparative anatomy shows that this is not the case. In birds, the great pectoral is always attached close to the articulation of the shoulder. The absolute distance which separates this attachment from the center of motion of the humerus, seems to vary only in accordance with the form of the bird, and not according to the relatively greater or less extent of the wings. This last depends principally upon the length of the forearm and its appendages. It is not necessary to dissect many birds of different kinds to prove the constancy of the law which I

Fig. 11.

Skeleton of a flamingo, (after Milne Edwards.) The wing is very large, and the sternum very short.

Fig. 4. Image 10 s stage output

5 Conclusion

In this paper, we have investigated the performance of Principal Component Analysis (PCA) and Local Pixel Grouping (LPG) in handling noise de-noising on historical document image. In order to preserve the local image structure when de-noising, we need to model a pixel and its neighbors as a vector variable, then use PCA technique to convert the de-noising of the pixel into estimation of variable from its noisy observation. However the PCA transformation matrix needs adaptive train from the local window of the image which consider as noisy variable. Therefore, the block matching based LPG will classify and select sample of local window to ensure that the only similar sample block are used in the PCA transformation estimation. The LPC-PCA algorithm will repeat one more time for better an image characteristic preservation. Our experiment results have demonstrated the effectiveness of LPG-PCA algorithm on smoothing noise of historic document image.

Acknowledgments. This work was supported by PJP High Impact Research Grant (S01473-PJP/2016/FTMK/HI3) from Universiti Teknikal Malaysia Melaka (UTeM), Malaysia.

References

1. Baird, H.S.: State of the art of document image degradation modeling. In: IAPR 2000 Workshop on Document Analysis System, Brazil, December (2000)
2. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
3. Mitra, A.: Restoration of noisy document images with an efficient bi-level adaptive thresholding. Int. J. Comput. Intell. (IJCI) **2**(2), 118–123 (2005)
4. Kim, Y.-T.: Contrast enhancement using brightness preserving bi-histogram equalization. IEEE Trans. Consum. Electron. **43**(1), 1–8 (1997)
5. Kim, J.-Y., Kim, L.-S., Hwang, S.-H.: An advanced contrast enhancement using partially overlapped sub-block histogram equalization. IEEE Trans. Circ. Syst. Video Technol. **11**, 475–484 (2001)
6. Nomura, S., Yamanaka, K., Shiose, T., Kawakami, H., Katai, O.: Morphological preprocessing method to thresholding degraded word images. Pattern Recogn. Lett. **30**(8), 729–744 (2009)
7. Tizhoosh, H.R.: Fuzzy Image Processing. Springer, Berlin (1997)
8. Parvathi, R., Jayanthi, S., Palaniappan, N., Devi, S.: Intuitionistic fuzzy approach to enhance text documents. In: Proceedings 3rd IEEE International Conference on Intelligent Systems (IEEE IS 2006), pp. 733–737 (2006)
9. Kohmura, H., Wakahara, T.: Determining optimal filters for binarization of degraded characters in color using genetic algorithms. In: Proceedings of 18th International Conference on Pattern Recognition, vol. 3, pp. 661–664 (2006)
10. Deborah, H., Arymurthy, A.M.: Image enhancement and image restoration for old document image using genetic algorithm. In: Proceedings of Second International Conference on Advances in Computing, Control and Telecommunication Technologies (ACT 2010), pp. 108–112 (2010)

11. Muresan, D.D., Parks, T.W.: Adaptive principal components and image denoising. In: Proceedings of the 2003 International Conference on Image Processing, 14–17 September 2003, vol. 1, pp. I101–I104 (2003)
12. Zhang, L., Dong, W.S., Zhang, D., Shi, G.M.: Two-stage image denoising by principal component analysis with local pixel grouping. Pattern Recogn. **43**, 1531–1549 (2010)
13. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice Hall, Englewood Cliffs (1989)
14. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
15. Rice, S.V.: Measuring the accuracy of page-reading systems. Ph.D. dissertation, Department of Computer Science, University of Nevada, Las Vegas (1996)
16. Zhang, L., Dong, W., Zhang, D., Shi, G.: Two-stage image denoising by principal component analysis with local pixel grouping. Pattern Recogn. **43**(4), 1531–1549 (2010)
17. Kumar, C.P.: Study of image de-noising using PCA local pixel grouping. IJITR **2**(6), 1581–1589 (2014)

A Robust and Optimally Pruned Extreme Learning Machine

Ananda L. Freire(✉) and Ajalmar R. Rocha Neto

Department of Teleinformatics, Federal Institute of Ceará,
Fortaleza, Ceará 60050-215, Brazil
anandalf@gmail.com, ajalmar@ifce.edu.br

Abstract. In recent years, the interest in the study of outlier robustness properties in Extreme Learning Machines (ELM) has grown. Most of the published works uses a more robust estimation method than the commonly adopted ordinary least squares. Nevertheless, the ELM network offers other challenges that also influence its robustness properties, such as the number of hidden neurons and the computational stability of the hidden layer's output matrix. That being said, we propose here two networks: ROP-ELM and ROPP-ELM that address the three aforementioned problems at once, in a combination of a pruning method, a cost function based on ℓ_1-norm and the addition of a biologically plausible mechanism named Intrinsic Plasticity.

Keywords: Extreme Learning Machine · Robust regression · Pruning · Batch Intrinsic Plasticity

1 Introduction

The Extreme Learning Machine (ELM) [1] is a popular artificial neural network due to its good generalization performance and fast training period by randomly selecting the weights and biases of the hidden neurons. This feature allows the problem of estimating the output weights to be cast as a linear model, which is then analytically determined by finding a least squares solution [2]. Nevertheless, it also offer some challenges.

As any single layer feedforward network (SLFN), the number of hidden neurons must be determined. A small network may not provide a good performance due to its limited information processing which leads to a too simple model, while a large one may lead to overfitting and poor generalization. The most common strategy is to resort to trial-and-error or exhaustive search strategies [3].

Another issue is related to the random nature of the input-to-hidden-layer weights (hidden weights, for short). The random choice of weights values may generate an ill-conditioned matrix of hidden neurons' activations **H**. In this case, the resulting output weight vector becomes numerically unstable due to the high

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_9

norm of its weights and, hence, make the whole network very sensitive to data perturbation (e.g. outliers) [4].

The third issue concerns the ELM's outlier robustness property, which also suffers influence of the aforementioned problems. Regression and classification problems are often contaminated by noise which affects the modelling accuracy as well as the estimated parameters [5].

Against this background, in the past few years, the interest in robust ELM networks has grown. The implementation of robustness feature in ELM networks have been usually accomplished by choosing an estimation method that is less sensitive to outliers than the ordinary least squares estimation. Some works, such as [6–9], adopt M-estimator algorithms. The [10] uses a rank-based Wilcoxon approach to solve the resulting linear system and [11,12] adopt a ℓ_1-norm loss function to enhance robustness.

Considering that ℓ_1-norm is more robust to outliers, produces sparser models than ℓ_2-norm [11] and that the study of robust ELM's architecture design is still in its infancy, our first proposition intends to cover those challenges. Named ROP-ELM, it adopts the ℓ_1-norm loss function and associates it with a known pruning method, named OP-ELM [13], in order to tackle the problem of the design of the hidden layer.

To also improve the numerical stability of the solutions, our second contribution, called ROPP-ELM, uses a method which adapts the activation function's parameters (slope and bias) to force the output of the hidden neurons to follow an exponential distribution. Named Batch Intrinsic Plasticity (BIP), it maximizes the networks information transmission, caused by the high entropy of the distribution and also leads to good generalization properties [14]. Our proposals have been tested on real-world regression problems with promising results.

This paper is organized as follows: Sect. 2 presents a brief overview of the used networks. In Sect. 3, we describe the methodology adopted for training and testing. Section 4 presents the results and Sect. 5 provides our conclusions.

2 Algorithms

2.1 Extreme Learning Machine

ELM is a two layer network with random and fixed weights matrix on the hidden layer, $\boldsymbol{W}_{hdd} \in \mathbb{R}^{p \times q}$, where p is the number of input attributes and q the number of hidden neurons [2]. The output of the hidden layer is given by $\mathbf{h}_n = \phi\left(\boldsymbol{W}_{hdd}^T \mathbf{u}_n + \mathbf{b}_n\right)$, where $\mathbf{h}_n \in \mathbb{R}^q$, $\mathbf{u}_n \in \mathbb{R}^p$ is the current input vector, $\mathbf{b}_n \in \mathbb{R}^q$ receives the biases and ϕ is a logistic activation function.

During the first step of training phase, all inputs from the training sequence $((\mathbf{u}_n, \mathbf{d}_n), n = 1...N)$ are presented to the network and the corresponding network states $(\mathbf{h}_n, \mathbf{d}_n)$ are harvested in $\mathbf{H} \in \mathbb{R}^{N \times q}$ and $\mathbf{D} \in \mathbb{R}^{N \times m}$ matrices respectively, where $\mathbf{d}_n$ is the desired output. For the second and last part of training, we compute the output weights $\boldsymbol{W} \in \mathbb{R}^{q \times m}$, which connects the hidden layer to the output neurons, as a linear regression problem, given by $\mathbf{Y} = \mathbf{H}\,\boldsymbol{W}$.

The ordinary least square (OLS) solution of the linear system created is given by the Moore-Penrose generalized inverse: $\boldsymbol{W} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{D}$.

In several real-world problems the matrix $\mathbf{H}^T\mathbf{H}$ can be singular. In fact, a near singular $\mathbf{H}^T\mathbf{H}$ (yet invertible) matrix is also a problem because it can lead to numerically unstable results. Huang *et al.* [1] has solved this issue using Singular Value Decomposition (SVD) approach to compute the Moore-Penrose pseudo-inverse instead of $\left((\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\right)$ [7]. That method can support either full or not full column rank matrices [7], which gives to ELM some computational robustness property. Unfortunately, it is also computationally expensive when dealing with large datasets and may produce an unreliable solution when the training data is corrupted by outliers.

2.2 Optimally Pruned Extreme Learning Machine

Optimally Pruned Extreme Learning Machine (OP-ELM) comprises three main steps: SFLN construction, ranking the neurons using Multiresponse Sparse Regression (MRSR) and selection of the number of neurons using a Leave-One-Out (LOO) validation.

In the first step, a SLFN is constructed with a large number of hidden layer neurons. The OP-ELM methodology allows the use and combination of three different sorts of kernels instead of only sigmoid kernels. The possible types are linear, sigmoid and Gaussian.

The second step consists in building a MRSR that maps the outputs of the hidden layer to the targets. MRSR can be seen as an extension of the Least Angle Regression (LARS) algorithm and hence is actually a variable ranking technique, rather than a selection one [13]. More details about the MRSR can be found in [15].

Once the neurons are ranked, the decision for the best number of neurons in the model is taken by a LOO validation method. However, the LOO method can be very time consuming, if the data set is large. To overcome this drawback, the authors use the PREdiction Sum of Squares (PRESS) statistics, which can give a direct and exact formula for the calculation of the LOO [13].

2.3 Outlier-Robust Extreme Learning Machine

Outlier-robust extreme learning machine (ORELM) [12] is an ELM which has adopted an ℓ_1-norm loss function for problems contaminated by outliers. In order to solve the resulting optimization problem, the authors substituted the OLS algorithm by the augmented Lagrange multipliers (ALM) method, with its function given by:

$$L_\mu(\boldsymbol{E}, \boldsymbol{W}, \boldsymbol{\lambda}) = \|\boldsymbol{E}\|_1 + \frac{1}{C}\|\boldsymbol{W}\|_2^2 + \boldsymbol{\lambda}^T(\mathbf{D} - \mathbf{Y} - \boldsymbol{E}) + \frac{\mu}{2}\|\mathbf{D} - \mathbf{Y} - \boldsymbol{E}\|_2^2, \quad (1)$$

where $\boldsymbol{E} = (\mathbf{D} - \mathbf{Y})$, $\boldsymbol{\lambda} \in \mathbb{R}^N$ is a vector of the Lagrange multiplier and μ is a penalty parameter, determined as $\mu = 2N/\|\mathbf{D}\|_1$. The ALM algorithm estimates

the optimal solution ($\boldsymbol{E}$, $\boldsymbol{W}$) and the Lagrange multiplier $\boldsymbol{\lambda}$ by iteratively minimizing the augmented Lagrangian function and admits the following explicit solution:

$$\boldsymbol{W}_{k+1} = \left(\mathbf{H}^T\mathbf{H} + \frac{2}{C\mu}\mathbf{I}\right)^{-1} \mathbf{H}^T \left(\mathbf{D} - \boldsymbol{E}_k + \frac{\boldsymbol{\lambda}_k}{\mu}\right), \tag{2}$$

where $I \in \mathbb{R}^{q\times q}$ is an identity matrix and $\boldsymbol{E}_{k+1}$ is given by

$$\boldsymbol{E}_{k+1} = \max\{|\mathbf{D} - \mathbf{H}\,\boldsymbol{W}_{k+1} + \boldsymbol{\lambda}_k/\mu| - 1/\mu, 0\} \circ \text{sign}\{\mathbf{D} - \mathbf{H}\,\boldsymbol{W}_{k+1} + \boldsymbol{\lambda}_k/\mu\}, \tag{3}$$

where $\circ$ represents the element-wise multiplication.

The most computational expensive part of this method is mainly spent on solving the inverse ($\mathbf{H}^T\mathbf{H} + \frac{2}{C\mu}\mathbf{I}$), nevertheless, it is pre-calculated before the beginning of the iterations. This contributes to the ALM algorithm to become a fast and robust estimation method. For implementation details, see [12,16].

2.4 ROP-ELM

Given the growing interest in robust ELM networks and the challenge of network architecture design inherent to any SLFN, our first contribution, named *Robust and Optimally Pruned ELM* (ROP-ELM), combines a known pruning algorithm, OP-ELM, with a fast and robust estimation method. Figure 1 shows its main steps: the SLFN is first built using the ELM approach (random initialization of hidden weights and biases); then a MRSR algorithm is used to rank the neurons of the hidden layer; finally the selection of the optimal number of neurons is performed using a Leave-One-Out criterion. The used least squares solution for the output weights, which is sensible to outliers in OP-ELM algorithm, is then substituted by the ALM algorithm described in [12].

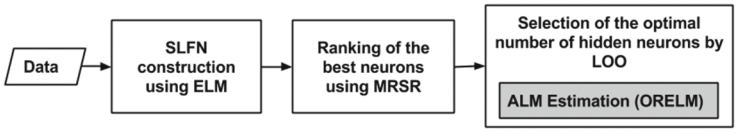


Fig. 1. ROP-ELM main steps.

However, as OP-ELM does not take into account the numerical stability of output weights before pruning, it may produce some solutions with high values for output weights. In order to improve this feature, our next proposition introduces the use of Batch Intrinsic Plasticity (BIP), which is described below.

2.5 Batch Intrinsic Plasticity

BIP [14] is an unsupervised learning rule performed only once as a pre-training. It is based on a biologically plausible mechanism, that adapts bias (b_j) and slope

(a_j) of the hidden neurons activation function, tuning them into more suitable regimes, maximizing information transmission and acting as a feature regularizer [14]. This is accomplished by forcing the j-th hidden neuron activation with a logistic activation function (Eq. 4) into a desired exponential distribution f_{des}.

$$h_{jn} = \frac{(1 - \exp(-a_j x_{jn} - b_j))}{(1 + \exp(-a_j x_{jn} - b_j))} \tag{4}$$

For each hidden neuron, all the incoming synaptic sum $\mathbf{x}_j = \boldsymbol{W}_{hdd_j}^T \mathbf{U}$ is collected, where $\mathbf{U} = (\mathbf{u}(1), ..., \mathbf{u}(N))$. Then random targets $\mathbf{T}_{f_{des}} = (t_1, ..., t_N)^T$, from the desired exponential output distribution, and the collected stimuli $\mathbf{x}_j$ are drawn in ascending order.

The model $\Phi(\mathbf{x}_j) = (\mathbf{x}_j^T, (1, ..., 1)^T)$ is built so we can calculate: $(a_j, b_j)^T = (\Phi(\mathbf{x}_j)^T \Phi(\mathbf{x}_j) + \gamma I)^{-1} \Phi(\mathbf{x}_j)^T f^{-1}(\mathbf{t}_{f_{des}})$, where f^{-1} is the inverse of the activation function. Thus, $\gamma > 0$ is the regularization parameter and $I \in \mathbb{R}^{2\times 2}$ is an identity matrix [14].

2.6 ROPP-ELM

As aforementioned, our second contribution adds to the ROP-ELM algorithm the batch intrinsic plasticity procedure. It improves computational issues resulting from a ill-posed $\mathbf{H}$ by avoiding saturated, almost linear or constant hidden neurons. Besides, since every hidden layer neuron is updated separately, pruning procedures can also be used without recomputation of the already obtained biases and slopes. That being said, the resulting algorithm is named *Robust and Optimally Pruned with Intrinsic Plasticity ELM* (ROPP-ELM) and its procedure is described in Fig. 2.

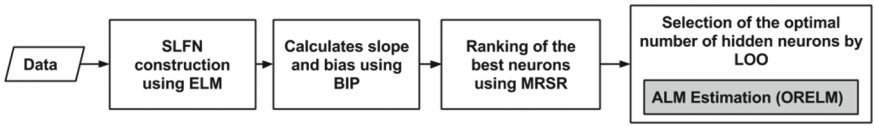


Fig. 2. ROPP-ELM main steps.

3 Methodology

All experiments involved basically two steps, as illustrated in Fig. 3. First, the data is randomized and then separated in 10 almost even sets of samples. We will use the idea of cross validation here to guarantee that all samples were used to test the algorithms at least one time, which results into ten independent repetitions.

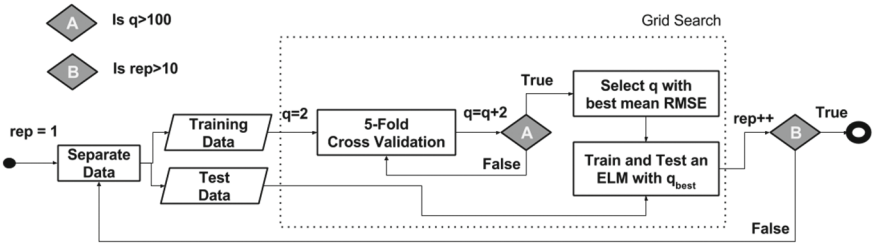


Fig. 3. A 10-fold and 5-fold cross validation combined with grid search scheme.

At each repetition, the second step depends on the method to be assessed. For ELM and ORELM, we applied a grid search, where the number of hidden neurons was increased from 2 to 100 neurons, in steps of 2. For each number of hidden neurons, it was performed a 5-fold cross validation. After all grid was covered, the number of neurons that returned the smallest mean validation error is chosen. Later, the network using the chosen number of hidden neurons is trained with training plus validation datasets, and then tested with unseen data. For the OP-based algorithms, the grid search is unnecessary. Therefore, we applied only the external 10-fold cross validation.

In order to evaluate the robustness of each algorithm, specific levels of noise were added to the target outputs of training sets. These levels allow us to control and observe the behaviour of the chosen algorithms with increasing distortion and if their performances will degrade or show robustness to it. However, it is important to highlight that the testing set has no added noise but the one embedded in the dataset itself. By doing that, we can evaluate if the algorithm has learned the model or it has been distorted by the outliers.

Inspired by [7], the contamination rates are 0%, 10%, 20% and 30% of training data. Accordingly with the percentage of contamination, $K \in 1, .., N$ training output samples are randomly selected. To each one of them, is added an error value drawn from an uniform distribution $\Delta_k \sim U[-1, 1]$.

We highlight that all networks received the same sets of data with the same contaminated samples.

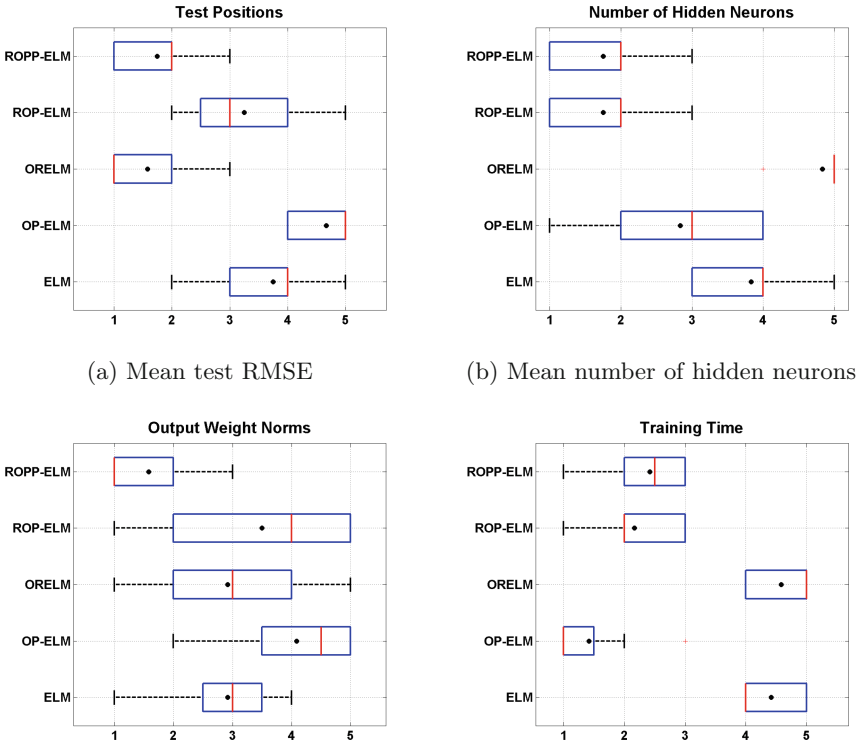
4 Results

The tests were performed with 3 different datasets regarding regression problems: Breast Cancer[1], Boston Housing and Body fat[2]. Their attributes were scaled to $[0, 1]$ and their target values to $[-1, 1]$. All the experiments were carried out in the Matlab 8.3 (2014a) environment running in Intel(R) Core(TM) i3 4130 processor with the speed of 3.40 GHz.

[1] UCI machine learning repository (2013), http://archive.ics.uci.edu/ml.
[2] Statlib: Data, software and news from the statistics community (1989), http://lib.stat.cmu.edu/index.php.

All networks use hyperbolic tangent as activation function. For the BIP procedure, the estimation of the activation function's parameters, as shown in Eq. 2.14, needs a regularization parameter $\gamma = 10^{-2}$ and the desired exponential distribution mean $\mu_{exp} = 0.2$. For ORELM the maximum number of iterations is 20 and the regularization parameter $C = 2^{-40}$. OPELM starts with 100 hidden neurons and, since the data has been already scaled, we do not set the internal normalization. We adopted the OP-ELM Toolbox v1.1 [17] and ORELM Toolbox v2.0 [16] implementations.

To evaluate the network's performance, we compare the tests RMSE, the resulting number of hidden neurons, the output weights norms and their training time. We highlight here that the training time includes the grid search period for ELM and ORELM, since it is an essential step to find a suitable network.



(a) Mean test RMSE

(b) Mean number of hidden neurons

(c) Mean output weight norms

(d) Training time

Fig. 4. Summary of ranking evaluation for all testing scenarios.

In order to facilitate an overview of all results, we ranked the networks by sorting their results in ascending order considering all test scenarios. This means that the smaller these values are, the better is the network performance.

Figure 4 shows the ranking distribution for each network, where the black dots represent the mean and the red lines are the medians. In terms of test RMSE, it is clear that ELM and OP-ELM, which are not robust to outliers, have the worst performance. As expected, the robust networks show better results. The ORELM network provides, most of the time, the smallest RMSE, followed by ROPP-ELM and ROP-ELM.

In contrast, ORELM has the lowest ranking on the other evaluation measures, indicating that it needs the largest amount of hidden neurons, needs a larger training time (due to grid search), and has a larger output weights norm than ELM and ROPP-ELM.

OP-ELM has the smallest training time and also provides the third smallest number of hidden neurons. However, its resulting network has the largest output weight norms and also did not exceed, as aforementioned, in RMSE performance.

Even though our contributions, ROP-ELM and ROPP-ELM, have the second and third best RMSE, they provide the smallest number of hidden neurons. ROPP-ELM, in particular, stands out by also producing the smallest output weights norms with a shorter training time than ORELM.

Tables 1, 2 and 3 present in details the mean and standard deviation of the 10-fold cross validation results of those evaluation measures, in all different contamination scenarios. The best results are highlighted.

Table 1. Body Fat: results for each contamination scenario.

		ELM	**OP-ELM**	**ORELM**	**ROP-ELM**	**ROPP-ELM**
0%	q	72.0 ± 16.92	38.0 ± 14.76	70.6 ± 10.83	42.5 ± 16.54	**28.0 ± 7.15**
	RMSE	0.039 ± 0.035	0.045 ± 0.030	0.036 ± 0.039	0.065 ± 0.072	**0.035 ± 0.019**
	$\|\|\mathbf{w}\|\|$	0.94 ± 0.19	1.66 ± 1.78	0.75 ± 0.09	2.75 ± 6.65	**0.46 ± 0.04**
	Time	0.946 ± 0.210	**0.058 ± 0.007**	0.763 ± 0.191	0.067 ± 0.014	0.064 ± 0.008
10%	q	18.4 ± 4.60	20.5 ± 7.62	53.8 ± 8.40	**14.0 ± 5.68**	14.5 ± 5.99
	RMSE	0.078 ± 0.042	0.067 ± 0.017	**0.044 ± 0.042**	0.047 ± 0.010	0.052 ± 0.016
	$\|\|\mathbf{w}\|\|$	0.88 ± 0.20	2.68 ± 2.92	0.79 ± 0.15	1.24 ± 2.28	**0.40 ± 0.07**
	Time	0.733 ± 0.106	**0.076 ± 0.007**	0.746 ± 0.060	0.105 ± 0.032	0.080 ± 0.023
20%	q	14.0 ± 3.13	10.5 ± 6.43	40.8 ± 6.27	**7.0 ± 3.50**	7.0 ± 3.50
	RMSE	0.079 ± 0.027	0.101 ± 0.071	**0.033 ± 0.028**	0.071 ± 0.017	0.070 ± 0.018
	$\|\|\mathbf{w}\|\|$	0.88 ± 0.15	0.60 ± 0.41	0.71 ± 0.27	2.12 ± 5.19	**0.43 ± 0.10**
	Time	0.765 ± 0.157	0.091 ± 0.033	0.646 ± 0.020	**0.086 ± 0.017**	0.087 ± 0.010
30%	q	9.6 ± 2.46	9.0 ± 5.16	33.0 ± 6.55	9.0 ± 6.15	**5.5 ± 1.58**
	RMSE	0.112 ± 0.044	0.127 ± 0.102	**0.041 ± 0.030**	0.088 ± 0.044	0.090 ± 0.036
	$\|\|\mathbf{w}\|\|$	0.94 ± 0.53	6.82 ± 18.82	0.71 ± 0.16	0.78 ± 0.27	**0.41 ± 0.04**
	Time	0.766 ± 0.091	**0.060 ± 0.007**	0.820 ± 0.041	0.062 ± 0.005	0.070 ± 0.013

Table 2. Breast Cancer: results for each contamination scenario.

		ELM	**OP-ELM**	**ORELM**	**ROP-ELM**	**ROPP-ELM**
0%	q	76.2 ± 10.13	**43.5 ± 13.34**	70.4 ± 13.29	45.0 ± 8.16	56.0 ± 8.76
	RMSE	0.072 ± 0.028	0.095 ± 0.037	0.081 ± 0.029	0.085 ± 0.015	**0.043 ± 0.009**
	$\|\|\mathbf{w}\|\|$	**0.51 ± 0.06**	0.54 ± 0.13	0.56 ± 0.11	0.52 ± 0.06	0.54 ± 0.06
	Time	0.591 ± 0.080	**0.059 ± 0.014**	0.571 ± 0.015	0.109 ± 0.110	0.074 ± 0.008
10%	q	29.2 ± 8.18	28.0 ± 11.35	50.6 ± 9.89	23.0 ± 9.78	**22.5 ± 4.86**
	RMSE	0.124 ± 0.053	0.126 ± 0.041	0.080 ± 0.026	0.111 ± 0.020	**0.067 ± 0.013**
	$\|\|\mathbf{w}\|\|$	0.61 ± 0.07	0.61 ± 0.15	0.63 ± 0.07	9.36 ± 27.88	**0.54 ± 0.09**
	Time	0.497 ± 0.030	**0.071 ± 0.009**	0.555 ± 0.021	0.079 ± 0.007	0.085 ± 0.014
20%	q	21.0 ± 4.55	24.5 ± 10.39	38.6 ± 8.17	**18.0 ± 7.89**	20.5 ± 5.99
	RMSE	0.153 ± 0.059	0.162 ± 0.058	0.115 ± 0.026	0.115 ± 0.018	**0.080 ± 0.015**
	$\|\|\mathbf{w}\|\|$	0.64 ± 0.13	60.16 ± 188.71	0.63 ± 0.12	0.61 ± 0.18	**0.56 ± 0.07**
	Time	0.545 ± 0.053	**0.054 ± 0.006**	0.648 ± 0.048	0.066 ± 0.006	0.062 ± 0.008
30%	q	19.2 ± 4.92	23.5 ± 13.34	31.6 ± 10.70	**12.0 ± 2.58**	13.5 ± 5.30
	RMSE	0.171 ± 0.037	11.568 ± 35.981	**0.111 ± 0.034**	0.186 ± 0.142	0.117 ± 0.041
	$\|\|\mathbf{w}\|\|$	0.72 ± 0.14	$3\times10^4\pm11\times10^4$	**0.57 ± 0.07**	3.52 ± 9.10	0.61 ± 0.12
	Time	0.516 ± 0.022	0.067 ± 0.025	0.586 ± 0.031	**0.066 ± 0.007**	0.067 ± 0.006

Table 3. Boston Housing: results for each contamination scenario.

		ELM	**OP-ELM**	**ORELM**	**ROP-ELM**	**ROPP-ELM**
0%	q	91.4 ± 4.43	**52.0 ± 12.29**	92.8 ± 5.90	53.0 ± 12.74	70.0 ± 11.55
	RMSE	0.059 ± 0.018	0.065 ± 0.017	0.056 ± 0.012	0.067 ± 0.011	**0.046 ± 0.006**
	$\|\|\mathbf{w}\|\|$	0.80 ± 0.13	2.46 ± 6.07	**0.77 ± 0.11**	29.31 ± 90.59	1.32 ± 0.47
	Time	0.745 ± 0.031	**0.151 ± 0.023**	1.005 ± 0.129	0.164 ± 0.028	0.193 ± 0.039
10%	q	36.2 ± 7.15	32.0 ± 13.37	81.2 ± 13.57	**18.5 ± 5.80**	20.0 ± 8.50
	RMSE	0.094 ± 0.011	0.088 ± 0.012	**0.062 ± 0.010**	0.075 ± 0.009	0.067 ± 0.013
	$\|\|\mathbf{w}\|\|$	0.71 ± 0.12	99.46 ± 234.26	0.82 ± 0.09	**0.51 ± 0.08**	0.64 ± 0.18
	Time	1.028 ± 0.154	**0.123 ± 0.030**	0.971 ± 0.147	0.154 ± 0.020	0.168 ± 0.019
20%	q	29.4 ± 5.08	30.5 ± 18.17	71.6 ± 17.02	15.0 ± 7.45	**14.5 ± 6.85**
	RMSE	0.104 ± 0.020	0.429 ± 0.946	**0.074 ± 0.015**	0.089 ± 0.019	0.085 ± 0.023
	$\|\|\mathbf{w}\|\|$	0.74 ± 0.11	599.04 ± 1×10^4	0.82 ± 0.18	**0.64 ± 0.52**	0.68 ± 0.19
	Time	1.141 ± 0.234	**0.149 ± 0.013**	1.116 ± 0.228	0.151 ± 0.025	0.175 ± 0.022
30%	q	27.6 ± 5.87	21.5 ± 8.18	58.0 ± 11.74	15.5 ± 14.23	**13.5 ± 3.37**
	RMSE	0.102 ± 0.018	0.113 ± 0.025	**0.072 ± 0.016**	0.094 ± 0.010	0.081 ± 0.016
	$\|\|\mathbf{w}\|\|$	0.69 ± 0.10	0.64 ± 0.28	0.73 ± 0.10	362.84 ± 1×10^4	**0.58 ± 0.15**
	Time	1.053 ± 0.286	0.185 ± 0.067	1.088 ± 0.192	0.177 ± 0.034	**0.177 ± 0.021**

5 Conclusion

In this paper, we combine a robust ELM, named ORELM with a pruning method in order to achieve robust solutions with smaller and stabler networks. The two resulting contributions are ROP-ELM and ROPP-ELM, that differs from each other by the use of a batch intrinsic plasticity method (BIP).

The results illustrate how the ELM network and OP-ELM, which do not provide any kind of robust estimation method, has their performance worsened as the training data is increasingly contaminated by outliers.

As Tables 1, 2 and 3 illustrate, our contributions offer solutions from 20% to 85% less hidden neurons than ORELM, while maintaining smaller output weights norm and less training time. Although the test RMSE is slightly larger than the ORELM, the results are still robust to outliers and, with a smaller set of hidden neurons and output norms that guaranties a stable response of the network, the ROPP-ELM proposition offers solutions especially suitable for environments with low computational resources.

As for future work, we will explore different robust estimations methods and other pruning algorithms, as well as evaluate their performances when dealing with outliers of different distributions.

Acknowledgments. The authors would like to thank the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) for the financial support.

References

1. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing **70**(1–3), 489–501 (2006)
2. Huang, G.B., Wang, D., Lan, Y.: Extreme learning machines: a survey. Int. J. Mach. Learn. Cybern. **2**(2), 107–122 (2011)
3. Yang, Y., Wang, Y., Yuan, X.: Bidirectional extreme learning machine for regression problem and its learning effectiveness. IEEE Trans. Neural Netw. Learn. Syst. **23**(9), 1498–1505 (2012)
4. Zhao, G., Shen, Z., Man, Z.: Robust input weight selection for well-conditioned extreme learning machine. Int. J. Inf. Technol. **17**(1), 1–13 (2011)
5. Beliakov, G., Kelarev, A., Yearwood, J.: Robust artificial neural networks and outlier detection. Technical report. CoRR (2011)
6. Huynh, H.T., Won, Y., Kim, J.J.: An improvement of extreme learning machine for compact single-hidden-layer feedforward neural networks. Int. J. Neural Syst. **18**(5), 433–441 (2008)
7. Horata, P., Chiewchanwattana, S., Sunat, K.: Robust extreme learning machine. Neurocomputing **102**, 31–34 (2013). (ELM 2011)
8. Freire, A., Barreto, G.: A robust and regularized extreme learning machine. In: ENIAC 2014, São Carlos (Brazil), October 2014
9. Barreto, G.A., Barros, A.L.B.: A robust extreme learning machine for pattern classification with outliers. Neurocomputing **176**, 3–13 (2016)
10. Xie, X.L., Bian, G.B., Hou, Z.G., Feng, Z.Q., Hao, J.L.: Preliminary study on wilcoxon-norm-based robust extreme learning machine. Neurocomputing **198**, 20–26 (2016)
11. Balasundaram, S., Gupta, D., Kapil, : 1-norm extreme learning machine for regression and multiclass classification using newton method. Neurocomputing **128**, 4–14 (2014)
12. Zhang, K., Luo, M.: Outlier-robust extreme learning machine for regression problems. Neurocomputing **151**, 1519–1527 (2015). Part 3

13. Miche, Y., Sorjamaa, A., Bas, P., Simula, O., Jutten, C., Lendasse, A.: Op-elm: optimally pruned extreme learning machine. IEEE Trans. Neural Networks **21**(1), 158–162 (2010)
14. Neumann, K., Steil, J.: Optimizing extreme learning machines via ridge regression and batch intrinsic plasticity. Neurocomputing **102**, 23–30 (2013). (ELM 2011)
15. Similä, T., Tikka, J.: Multiresponse sparse regression with application to multidimensional scaling. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3697, pp. 97–102. Springer, Heidelberg (2005). doi:10.1007/11550907_16
16. Zhang, K., Luo, M.: Orelm matlab toolbox. https://www.mathworks.com/matlabcentral/fileexchange/50877-outlier-robustextreme-learning-machine-for-regression-problems
17. Lendasse, A., Sorjamaa, A., Miche, Y.: Op-elm matlab toolbox. http://www.cis.hut.fi/projects/tsp/index.php?page=research&subpage=downloads

Investigating the Effect of Combining Text Clustering with Classification on Improving Spam Email Detection

Doaa Hassan(✉)

Department of Computers and Systems, National Telecommunication Institute, 5 Mahmoud El Miligy Street, 6th district-Nasr City, Cairo, Egypt
doaa@nti.sci.eg

Abstract. Nowadays emails have been an easy and fast tool of communication among people. As a result, filtering unsolicited/spam emails has become a very important challenge to achieve. Recently there has been some research work in text mining that combines text clustering with classification to improve the classification performance. In this paper, we investigate the effect of combining text clustering using K-means algorithm with various supervised classification mechanisms on improving the performance of classification of emails into spam or non-spam. The conjunction of clustering and classification mechanisms is carried out by adding extra features from the clustering step to the feature space used for classification. Our results show that combining K-means clustering with supervised classification by this methodology does not always improve the classification performance. Moreover, for the cases that the classifiers performance is improved by clustering, we found that the performance of classifiers in terms of accuracy is slightly increased with a very small amount that does not meet the increase in the time taken for building a learning model that combines both mechanisms. The result of our experiment has been shown using the Enron-Spam datasets.

Keywords: Text clustering · Classification · Spam email detection

1 Introduction

In the recent years, spam emails have been a huge problem that spreads widely on the Internet. They are represented as unsolicited (junk) messages that are sent to a large number of users. As a result, the identification of spam email from the non spam one has been a great challenge to achieve, once the message has been received. Otherwise, a large number of spam emails might be received and result in a full email inbox that causes a blocking of the user's email account.

The text clustering and classification have been used in data mining research field as great methodologies to identify spam emails [3,6]. The former does grouping of spam emails in one cluster with a spam label, while grouping the non-spam

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_10

ones into another cluster with a non spam label. Thus each new received email is specified as spam or non-spam by measuring the similarity between its vector and the centroid vector of each cluster. Finally the label of the most similar cluster is assigned to the new email. As for text classification, the basic idea is to analyze a large corpus of emails and extract some syntactic emails that are used to train a classifier to learn a detection model that distinguish the spam emails from the ham ones.

Recently, there has been some research work in text mining that proposed combining text clustering with classification to improve the classification performance [4]. However, such an approach has not been applied to spam email detection in an explicit form that illustrates if the cluster label is used as a class label for a classification or as an ordinary feature in the feature space used by the classifier to identify spam emails. To this end, in this paper, we investigate the effect of combining text clustering using k-means algorithm [20] with various supervised classification mechanisms on improving the performance of spam email detection. The conjunction of clustering and classification is carried out by adding the clustering label as an attribute to the feature vector used by the classifier to learn the detection model that is able to identify spam emails. Our results show that combining k-means clustering with supervised classification using this methodology does not always improve the classification performance. Among all classifiers, only the conjunction of logistic regression with K-means clustering has been found to outperform slightly using logistic regression on its own. However, we found that such an improvement does not meet the increase in the time required to build the learning model that combines both the clustering and classification algorithms, particularly if the conjunction is applied to large datasets.

The structure of this paper is organized as follows: in Sect. 2, we provide an overview of text mining and discuss the related work. In Sect. 3, we present our approach for applying a combination of text clustering and classification on Enron datasets [9]. In Sect. 4, we present the experimental settings as well as the performance evaluation results. Finally, in Sect. 5 we provide conclusions with some directions for future work.

2 Background

2.1 Overview of Text Mining

The basic idea of text mining is to break the document into words and then treat each word as an attribute in the feature vector of the machine learning model [1]. The value of each attribute is determined using term weighting; a parameter that refers in the simplest case to the number of occurrence of a specific word **W** in the content of the document d. A common term weighting scheme is the $TF * IDF$ weighting scheme [12] which is defined by:

$$TF * IDF(t,d) = TF(t,d) * IDF(t)$$

where $TF(t,d)$ refers to the number of times that term t (i.e., word) occurs in document d, and $IDF(t)$ is the inverse document frequency that is defined by:

$$IDF(t) = log_2 \frac{N}{DF(t)}$$

where N is the number of documents in the dataset and $DF(t)$ refers to the number of documents containing the term t at least once.

2.2 Related Work

I. Alsmadi and I. Alhami [8] developed algorithms for performing clustering and classification for large text collection of a large set of personal emails. They combined K-means clustering with SVM classification to improve classification performance for the purpose of subject and folder and classifications. Their approach classifies each email into three categories/clusters that are produced using K-means clustering: personal, professional and other based on content of email, its header and its sender. However this approach was not applied to improve the classification for the purpose of spam detection explicitly.

M. Basavaraju and R. Prabhakar [7] presented an approach that converts emails document into a vector form then applying clustering to two different training sets, one belonging to spam and the other one belonging to non-spam, where the cluster representatives belong to two different classes spam or non spam. Finally they applied classification using Nearest neighbor and K-nearest neighbor classifiers to assigns each unlabeled document in the test set to its nearest labeled neighbor cluster. However, this approach uses the cluster label as class label for classification. This is different from our approach which uses cluster label as an attribute in the feature space used for the classification of email document into spam or non spam.

N. O. F. Elssied et al. [10] proposed a mechanism for e-mail spam detection based on hybrid of SVM and K-means clustering algorithms that involves one more input parameter in the feature vector of SVM classifier: the number of clusters. However this mechanism is not a text mining one and it was tested on the spambase datatset [11] that has a predefined number of features.

Our approach follows the one proposed by A. Kyriakopoulou and T. Kalamboukis [5]. It combines text clustering with classification to improve the performance of classifying users'posts in social bookmarking systems into spam or non-spam. Thought their approach is very closed to us as they augment the dataset with meta-features originated from the clustering step, it was not applied to detect spam emails. Also it combines clustering with SVM classification mechanism only and did not highlight combining clustering with other classification mechanisms. Moreover, it does not take into account the increase in the time required to build the classification model in case of combining clustering with classification.

3 Main Methodology

The first step in our approach is to parse all email documents in the email dataset that consists of spam and non spam emails and calculate the frequency of words needed for building the word frequency table. First, we have selected words of frequency greater than or equal 1 in the whole email dataset in one trial as this frequency is set in the default settings of Weka StringToWordVector filter. Next we change the default settings of word frequency by selecting words of frequencies above 50 or 100 in other two trials. Using each one of both frequencies allows to filter/capture the words that are highly repeated in the email document. In each trial, we have used the $TF * IDF$ weighting scheme introduced in the previous section as a measure to word frequency. Also, we applied stoplist removal mechanism [19] on the generated word frequency table to exclude some of the generic terms/words (i.e., stoplist words) in all email documents such as: the, is, you, I, an, this, and, for, to, etc. Those terms are excluded as they may not be relevant features in the classification process. Also we apply stemming removal mechanism [15] to reduce words into their root forms by removing prefix, suffixes and infixes. The word frequency table represents the email matrix where rows represent the emails documents samples and columns represent the $TF * IDF$ weight of each word (i.e., feature vectors) in addition to a column that represent the class label of each training sample. This matrix can be an input to either the clustering or classification algorithms.

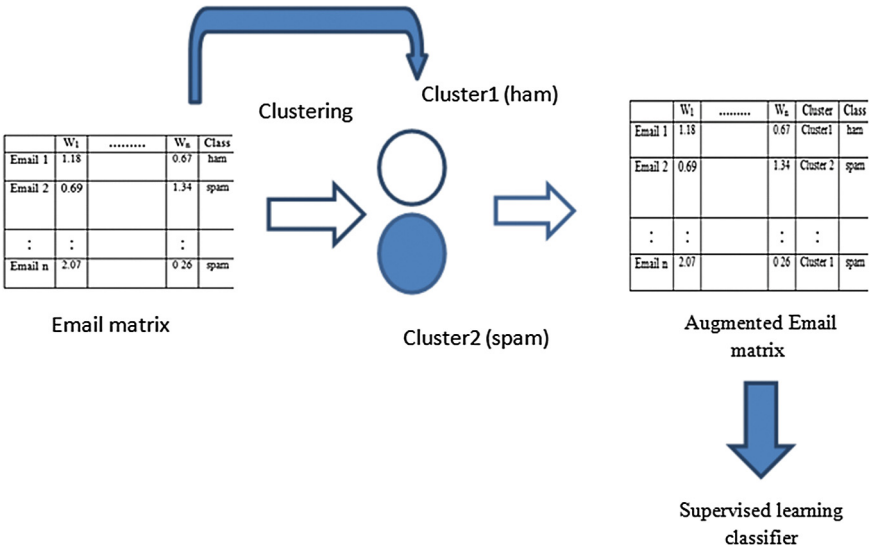


Fig. 1. The conjunction of clustering and classification for the purpose of spam email detection.

Next, for each trial, we have build a learning model that combines K-means clustering algorithm and various classifications algorithms including those that are widely used for text classification such as Naive Bayes, support vector machine (SVM) and logistic regression (LR) in addition to decision tree (DT) [14]. Figure 1 shows the conjunction of clustering and classification in one trial for the purpose of email spam detection. The following steps summarize the procedure of combining clustering with classification for identifying spam emails in each trial:

- Apply the K-means clustering algorithm to the generated email matrix as described above to divide the email dataset based on the values of the words frequencies for each sample into two clusters: non-spam and spam.
- Augment the original email matrix with an extra feature, called cluster that has been originated from the clustering step. This attribute is different from the class label of the training data and it has two possible values: cluster1 to indicate that the email document belongs to non-spam emails. Cluster2 that indicates the opposite. Table 1 shows an example of the augmented email matrix.
- Train and test the classifier with the expanded email matrix generated in the previous step in 10-fold cross validation, where 90% of dataset is randomly selected for training and the remaining 10% is kept for testing in 10 repeated iterations.

Table 1. An example of an email matrix augmented with the cluster feature.

	W_1	.....	W_n	Cluster	Class
$Email_1$	1.18		0.67	$Cluster_1$	Ham
$Email_2$	0.69		1.34	$Cluster_2$	Spam
:	:		:	:	
$Email_n$	2.07		0.26	$Cluster_1$	Spam

4 Basic Experiment

4.1 Experimental Settings

We have used Weka [13], a free data mining software for email text preprocessing and performing email text mining with clustering and classification.

The first step in emails text preprocessing is to convert the dataset of emails documents samples into Attribute-Relation File Format (ARFF) using Weka TextDirectorytoArff converter [16]. Second, the Weka StringToWordVector filter [17] is used to convert string attribute into a set of attributes representing word occurrence/frequency (i.e., weights of each word in the email matrix introduced in the previous section). This filter has various options that enable us to do the following:

– use Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme to determine the value of each attribute.
– apply stemming mechanism for reducing words into their root form by removing their prefixes, suffixes and infixes.
– determine the minimum frequency of words.
– remove stopword words (e.g., the generic redundant terms/words in all email documents such as: you, the, is, etc.).
– get the number of output words produced by breaking the email document into words rather than indicating their presence or absence by boolean 0 or 1.

For performing email text mining using a conjunction of classification and clustering, we have used Weka FilteredClassifier [17] that enables us to run any arbitrary supervised classifier on data that has been passed through two filters [18]: StringToWordVector and AddCluster. The later is a filter that allows us to apply a specific clustering algorithm and add a new nominal attribute to the feature space. This attribute represents the cluster assigned to each instance in the email matrix by the clustering algorithm.

We have experimented using the freely available Enron-Spam collection. This collection consists of six collections: Enron1, Enron2, Enron3, Enron4, Enron5, and Enron6. Each collection has a different number of spam and non spam emails. Enron1 has 1500 spam emails versus 3672 non-spam ones. Enron2 has 1496 spam emails versus 4361 non-spam ones. Enron3 has 1500 spam emails versus 4012 non-spam ones. Enron4 has 4500 spam emails versus 1500 non-spam ones. Enron5 has 3675 spam emails versus 1500 non-spam ones. Finally, Enron6 has 4500 spam emails versus 1500 non-spam ones. We have used 10-fold cross validation to divide each collection randomly in 10 sets, where 9 sets is used for training and the remaining set is used for testing in 10 iterations process. We have conducted three experiments on each dataset of the collection: one with setting minTermFreq option of StringToWordVector filter to be equal to 1 (the default setting), another one with making it equal to 50 and the third one with setting equal to 100. The three experiments have been conducted on a Windows 7 laptop machine with 2.6 GHZ processor Intel core (TM)i5 and 4 G Memory Rams.

4.2 Performance Evaluation and Results

The overall accuracy has been used as a measure for evaluating the classification performance either when it is combined with clustering or when considering it only for predicting email documents as non-spam or spam.

Table 2 shows the performance of experimental classifiers in three trials of different minimum word frequency, either when performing classification alone or in conjunction with clustering according to our proposed methodology introduced in Sect. 3. By investigating the average of results over the three trials as shown in the last row of Table 2, we can notice the following:

– In contrast to the case with other classifiers, the conjunction of logistic regression (LR) with clustering (i.e., C-LR) outperforms using LR alone with an average overall accuracy over the three trials of 93.99% for the former (as highlighted in bold font) and 93.79% for the later.

Table 2. The performance of experimental classifiers with and without clustering.

Dataset	Classification accuracy									
	MinTermFreq = 1									
	NB	C-NB	LR	C-LR	SVM	C-SVM	KNN	C-KNN	DT	C-DT
Enron1	93.15%	93.15%	91.80%	92.17%	96.80%	96.73%	90.55%	90.60%	94.43%	94.43%
Enron2	97.22%	97.22%	92.79%	93.24%	98.52%	98.58%	96.00%	95.96%	95.90%	95.90%
Enron3	93.68%	93.68%	99.43%	99.60%	99.73%	99.73%	99.75%	99.75%	99.75%	99.75%
Enron4	89.48%	89.48%	96.74%	97.08%	97.74%	97.74%	93.69%	93.72%	95.59%	95.56%
Enron5	97.03%	97.03%	96.32%	95.99%	97.96%	97.99%	92.16%	92.12%	94.61%	94.61%
Enron6	91.06%	91.06%	87.53%	87.98%	96.91%	96.60%	94.52%	94.52%	94.71%	94.71%
Average	93.60%	93.6%	94.1%	94.34%	97.94%	97.9%	94.45%	94.45%	95.83%	95.83%
MinTermFreq = 50										
Enron1	94.19%	94.19%	89.62%	89.92%	97.35%	97.38%	89.71%	89.71%	94.19%	94.17%
Enron2	97.22%	97.22%	91.99%	91.76%	98.70%	98.70%	96.21%	96.23%	96.09%	96.07%
Enron3	93.68%	93.68%	99.43%	99.60%	99.73%	99.73%	99.75%	99.75%	99.75%	99.75%
Enron4	91.41%	91.41%	95.85%	95.56%	97.72%	97.72%	92.89%	92.90%	95.46%	95.38%
Enron5	96.88%	96.88%	95.17%	95.32%	97.96%	97.92%	90.04%	90.00%	94.57%	94.57%
Enron6	91.38%	91.38%	86.91%	87.60%	96.66%	96.79%	95.28%	95.28%	94.02%	94.21%
Average	94.13%	94.13%	93.16%	93.29%	98.02%	98.04%	93.98%	93.98%	95.68%	95.69%
MinTermFreq = 100										
Enron1	93.15%	93.15%	91.80%	92.17%	96.80%	96.73%	90.55%	90.60%	94.43%	94.43%
Enron2	97.22%	97.22%	92.79%	93.24%	98.52%	98.58%	96.00%	95.96%	95.90%	95.90%
Enron3	93.68%	93.68%	99.43%	99.60%	99.73%	99.73%	99.75%	99.75%	99.75%	99.75%
Enron4	89.48%	89.48%	96.74%	97.08%	97.74%	97.74%	93.69%	93.72%	95.59%	95.56%
Enron5	97.03%	97.03%	96.32%	95.99%	97.96%	97.99%	92.16%	92.12%	94.61%	94.61%
Enron6	91.06%	91.06%	87.53%	87.98%	96.91%	96.60%	94.52%	94.52%	94.71%	94.71%
Average	93.60%	93.6%	94.1%	94.34%	97.94%	97.9%	94.45%	9.45%	95.83%	95.83%
Total average	93.78%	93.78%	93.79%	**93.99%**	**97.97%**	**97.95%**	94.29%	94.10%	95.78%	95.78%

- It is notable that neither LR on its own nor in conjunction with clustering (C-LR) achieves the best average overall accuracies among all other classifiers. SVM and (C-SVM) have been found to outperform other classifiers and achieve the best accuracies of 97.97% and 97.92% respectively as highlighted in bold font.
- Though the accuracy of (C-LR) outperforms that of LR, the average time required to build the learning model over the three trials for C-LR classifier is the highest among all classifiers as shown and highlighted in bold font in Table 3 that represents the time required for building the learning model. For sake of clarity, only the average time for building the learning model in the three trials at three as well as the total average over the three trials are only shown in the table.
- The improvement in the performance in case of (C-LR) does not seem to be proportional to the significant increase in the time required to build the learning model that combines the clustering and classification algorithms as shown in Table 3. As that table shows, the average time required to build the learning model over the three trials in case of LR classifier on its own is 25.51 s, while for C-LR is 243.39 s. This difference in time should be taken into account more if the conjunction of clustering and classification is applied to large datasets.

Table 3. The average time to build the classification model with and without clustering for each trial as well as the total average of time over the three trial.

Time to build learning model										
MinTermFreq = 1										
	NB	C-NB	LR	C-LR	SVM	C-SVM	KNN	C-KNN	DT	C-DT
Average	6.96	61.848	41.22	613.01	4.96	63.59	4.23	55.33	41.55	106.66
MinTermFreq = 50										
Average	7.07	51.61	20.11	72.52	5.91	45.57	4.75	46.05	39.36	78.81
MinTermFreq = 100										
Average	5.85	38.59	15.2	44.64	6.52	32.24	4.37	35.93	21.19	36.86
Total average	6.63	50.68	**25.51**	**243.39**	5.80	47.13	4.45	45.77	34.03	74.11

5 Conclusions and Future Work

We have investigated the effect of combining K-means clustering with various classification algorithms from different categories on improving the classification performance for the purpose of spam emails detection. The conjunction of clustering and classification extends the feature vectors used by the classifier with an extra feature originated from the clustering step that determines the cluster to which each email document sample belongs (i.e., either spam or non spam). We have presented our experimental results on the Enron datasets; an industry-standard collection of email data. The results of the experiments conducted show that the conjunction of K-means clustering approach with LR classifier achieves a slight improvement among other classifiers on the average over the use of a standard LR classifier on its own. However such an improvement does not seem to be proportional to the significant increase in the time required to build the learning model that combines both clustering and classification algorithms, particularly when applying this conjunction to large datasets.

As a future work, we are looking forward to investigating the effect of combining other alternatives to the k-means algorithm (such as those presented in [2]) with various classification algorithms on improving the classification performance. Also, we would like to find a methodology for decreasing the time required for building the learning model that combines both clustering and classification.

References

1. Weiss, S.M., et al.: Overview of text mining. In: Weiss, S.M., Indurkhya, N., Zhang, T. (eds.) Fundamentals of Predictive Text Mining, Chap. 1, pp. 1–12. Springer, London (2010)
2. Hamerly, G., Elkan, C.: Alternatives to the k-means algorithm that find better clusterings. In Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM) (2002)
3. Sasaki, M., Shinnou, H.: Spam detection using text clustering. In: Proceedings of the 2005 International Conference on Cyberworlds (CW 2005), pp. 316–319 (2005)

4. Kyriakopoulou, A., Kalamboukis, T.: Using clustering to enhance text classification. In: Proceedings of SIGIR 2007, Amsterdam, The Netherlands (2007)
5. Kyriakopoulou, A., Kalamboukis, T.: Combining clustering with classification for spam detection in social bookmarking systems. In: ECML/PKDD Discovery Challenge (2008)
6. Tretyakov, K.: Machine learning techniques in spam filtering. In: Data Mining Problem-oriented Seminar, Institute of Computer Science, University of Tartu, pp. 60–79 (2004)
7. Basavaraju, M., Prabhakar, R.: A novel method of spam mail detection using text based clustering approach. Int. J. Comput. Appl. **5**(4), 15–25 (2010)
8. Alsmadi, I., Alhami, I.: Clustering and classification of email contents. J. King Saud Univ. Comput. Inf. Sci. **27**(1), 46–57 (2015)
9. Klimt, B., Yang, Y.: The enron corpus: a new dataset for email classification research. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 217–226. Springer, Heidelberg (2004). doi:10.1007/978-3-540-30115-8_22
10. Elssied, N.O.F., Ibrahim, O., Abu-Ulbeh, W.: An improved of spam e-mail classification mechanism using k-means clustering. J. Theor. Appl. Inf. Technol. **60**(3), 568–580 (2014)
11. Set, Spam Base Data. https://archive.ics.uci.edu/ml/datasets/Spambase
12. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. J. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
14. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining, 1st edn. Addison-Wesley, Boston (2005)
15. Frakes, W.B., Baeza-Yates, R.: Information Retrieval: Data Structures and Algorithms. Prentice-Hall Inc., Upper Saddle River (1992)
16. Text categorization with WEKA. https://weka.wikispaces.com/Text+categorization+with+WEKA
17. Bouckaert, R.R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., Scuse, D.: WEKA Manual for Version 3-6-8. University of Waikato, Hamilton, New Zealand (2012)
18. Hidalgo, J.M.G.: Text mining in WEKA: chaining filters and classifiers, January 2013
19. Sarukkai, R.R.: Foundations of Web Technology. The Springer International Series in Engineering and Computer Science, vol. 698. Springer, Heidelberg (2002)
20. Teknomo, K.: K-means clustering tutorial. http://people.revoledu.com/kardi/tutorial/kMean/index.html. Accessed July 2007

Radial Basis Function Neural Networks for Datasets with Missing Values

Diego P. Paiva Mesquita and João Paulo P. Gomes(✉)

Department of Computer Science, Federal University of Ceará, Fortaleza, Ceará, Brazil
{diegoparente,jpaulo}@lia.ufc.br

Abstract. Radial Basis Function Neural Networks (RBFNN) are among the most popular supervised learning methods and showed significant results in various applications. Despite is applicability, RBFNNs basic formulation can not handle datasets with missing attributes. Aiming to overcome this problem, in this work, the RBFNN is modified to deal with missing data. For that, the expected squared distance approach is used to compute the RBF Kernel. The proposed approach showed promising results when compared to standard missing data strategies.

Keywords: Neural networks · Missing data · RBF neural networks

1 Introduction

The absence of some attributes in feature vectors is a common problem in many real world applications [1]. Aiming at managing this problem, several options are available, ranging from discarding samples to complex uncertainty management techniques [2]. By far, discarding samples is the most common strategy [3]. In the so-called listwise deletion, instances that have missing attributes are eliminated and the analysis is performed using the rest of the examples. Although commonly used, listwise deletion can lead to performance decrease when the number of instances with missing attributes is increased [4].

Another usual approach consists on replacing the missing values with plausible ones. These broad set of techniques are called imputation methods. A naive method is obtained when replacing each missing attribute by its average value estimated from the complete components of the dataset. Another method of imputation is obtained by searching for the nearest neighbors among only the fully known patterns [5]. This strategy may be ineffective when a majority of the data samples have missing components [6]. In [7], the authors propose a maximum likelihood estimation based approach using the Expectation-Maximization algorithm (EM). In the EM algorithm for datasets with missing values a multivariate Gaussian distribution is fit to the data and the missing values are imputed with its expected value given the observed attributes. An estimate of the variance is also possible by calculating the expected variance given the observed attributes.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_11

Apart from discarding examples and imputing missing values, recent works proposed variants of well known machine learning methods that can handle datasets with missing attributes in its formulation. In [8], the authors propose a modified risk function that incorporates a probability distribution for missing data in the Support Vector Machine formulation. Eirola *et al.* [6], proposed a method for pairwise distance estimation of vectors with missing values with application on K-NN classifiers. A similar approach was used on the Minimal Learning Machine (MLM, [9]) in [10] since MLM uses a distance matrix built from the training data.

In this paper, we propose a variant of the Radial Basis Function Neural Network (RBFNN) for datasets with missing values. The proposed method is based on the estimation of the squared distances used on the RBF Kernel. Several experiments were conducted with simulated and real world datasets and the performance of the proposed RBFNN was compared to other commonly used approaches for missing data.

The remainder of this paper is organized as follows. Section 2 presents a brief introduction to the formulation of the RBFNN. Section 3 presents the proposed variant of the RBFNN. Description of the experiments is presented in Sect. 4 while the results are shown in Sect. 5. Finally, conclusions and future works are presented in Sect. 6.

2 RBF Neural Networks

For given D-dimensional point X in the input space, the output of a RBF network with L centroids with locations in $\{C_1, \cdots, C_L\}$ and kernel width γ is given by

$$\hat{Y} = \sum_{l=1}^{L} w_l K(X, C_l) = \sum_{l=1}^{L} w_m exp\left\{ -\frac{\gamma}{2} \|X - C_l\|_2^2 \right\} \tag{1}$$

Consider a training set $\mathcal{D} = \{\mathcal{X} \in \mathbb{R}^{N \times D}, \mathcal{Y} \in \mathbb{R}^{N \times P}\}$ such that the n-th line of $\mathcal{X}$ is a point X_n in the input space and the n-th line of $\mathcal{Y}$ is its respective output Y_n. Once the centroid locations and the kernel width, the network's hyper-parameters, have been set, the weights $w = [w_1, \cdots, w_L]^T$ are learnt in order to minimize Eq. (2):

$$\sum_{n=1}^{N} \left(Y_n - \sum_{l=1}^{L} w_l K(X_n, C_l) \right)^2 \tag{2}$$

this can be done using the *Ordinary Least-Squares* (OLS) estimate:

$$w = (\mathcal{H}^T \mathcal{H})^{-1} \mathcal{H}^T \mathcal{Y} \tag{3}$$

in which $\mathcal{H} \in \mathbb{R}^{N \times L}$ such that $\mathcal{H}_{n,l} = K(X_n, C_l)$.

A detailed description of the RBFNN can be found in [11].

3 Computing $\mathcal{H}$ in the Presence of Missing Data

In the presence of missing data, to perform the computations required in the training step of a RBF Neural Network as described in Sect. 2, we first need to estimate the matrix $\mathcal{H}$. That could be achieved by imputing the missing entries, and thus performing the calculations presented in Eq. 3.

It is worth noting that this procedure does not consider the inherent uncertainty in the process of estimating the missing values. An alternative solution could be to estimate of the squared distances between the input vector and each centroid and use such estimate to calculate $\mathcal{H}$. Although this approach still does not directly estimate the expected value of $\mathcal{H}$, we believe that estimating the squared distances may lead to more precise estimates of $\mathcal{H}$ when compared to estimating the missing values directly. This can be explained by the work of Eirola *et al.* [6] that showed that the standard imputation procedure lead to underestimated squared distances.

In our proposal, we estimate the expected value of $\|X_n - C_l\|_2^2$ with the Expected Squared Distance (ESD) method introduced by Eirola *et al.* [6]. Assuming that data is Missing-at-Random (MAR) [1] and that the samples originate from some probability distribution, statistical techniques were applied to find an expression for the expectation of the ESD between samples. In the subsequent subsection, we describe the ESD method and how to proceed with its computation.

3.1 Expected Squared Distance (ESD) Calculation

Consider $\alpha, \beta \in \mathbb{R}^D$ drawn from a same multivariate probability distribution, but possibly with deleted entries. It is reasonable to estimate the squared euclidean norm between α and β as

$$\mathrm{E}[\|\alpha - \beta\|_2^2] = \sum_{i=1}^{D} \mathrm{E}[(\alpha_i - \beta_i)^2] \tag{4}$$

Let M_α and M_β be sets containing the index of the missing entries of α and β, respectively. We can further expand (4) into

$$\begin{aligned} \mathrm{E}[\|\alpha - \beta\|_2^2] = &\sum_{i \notin M_\alpha \cup M_\beta} (\alpha_i - \beta_i)^2 + \sum_{i \in M_\alpha \setminus M_\beta} \mathrm{E}[(\alpha_i - \beta_i)^2] \\ &+ \sum_{i \in M_\beta \setminus M_\alpha} \mathrm{E}[(\alpha_i - \beta_i)^2] + \sum_{i \in M_\alpha \cap M_\beta} \mathrm{E}[(\alpha_i - \beta_i)^2] \end{aligned} \tag{5}$$

By performing simple expected value calculations, we can develop the terms comprising the entries $i \in M_\alpha \setminus M_\beta$ into.

$$\begin{aligned} \mathrm{E}[(\alpha_i - \beta_i)^2] &= \mathrm{E}[\alpha_i^2 + \beta_i^2 - 2\alpha_i\beta_i] = E[\alpha_i^2] + \beta_i^2 - 2\mathrm{E}[\alpha_i]\beta_i \\ &= \mathrm{E}[\alpha_i^2] - \mathrm{E}[\alpha_i]^2 + \mathrm{E}[\alpha_i]^2 + \beta_i^2 - 2\mathrm{E}[\alpha_i]\beta_i \\ &= (\mathrm{E}[\alpha_i] - \beta_i)^2 + \mathrm{Var}[\alpha_i] \end{aligned} \tag{6}$$

By symmetry the terms involving entries $i \in M_\beta \backslash M_\alpha$ resume to

$$\mathrm{E}[(\alpha_i - \beta_i)^2] = (\alpha_i - \mathrm{E}[\beta_i])^2 + \mathrm{Var}[\beta_i] \tag{7}$$

In case both α_i and β_i are missing, assuming these are uncorrelated ($\mathrm{E}[\alpha_i\beta_i] = \mathrm{E}[\alpha_i]\mathrm{E}[\beta_i]$) we have

$$\begin{aligned} \mathrm{E}[(\alpha_i - \beta_i)^2] &= \mathrm{E}[\alpha_i^2] + \mathrm{E}[\beta_i]^2 - 2\mathrm{E}[\alpha_i]\mathrm{E}[\beta_i] \\ &= \mathrm{E}[\alpha_i^2] - \mathrm{E}[\alpha_i]^2 + \mathrm{E}[\beta_i^2] - \mathrm{E}[\beta_i]^2 + \mathrm{E}[\alpha_i]^2 + \mathrm{E}[\beta_i]^2 - 2\mathrm{E}[\alpha_i]\mathrm{E}[\beta_i] \\ &= (\mathrm{E}[\alpha_i] - \mathrm{E}[\beta_i])^2 + \mathrm{Var}[\alpha_i] + \mathrm{Var}[\beta_i] \end{aligned} \tag{8}$$

The results can be summarized leading to the following formulation.

$$\mathrm{E}[\|\alpha - \beta\|_2^2] = \|\mathrm{E}[\alpha] - \mathrm{E}[\beta]\|_2^2 + \sum_{i=1}^{D} \mathrm{Var}[\alpha_i] + \sum_{i=1}^{D} \mathrm{Var}[\beta_i] \tag{9}$$

If we assume that the distribution from which both α and β were drawn is a multivariate normal distribution, we can use the Expectation Maximization (EM) algorithm to estimate a mean vector μ and a covariance matrix Σ based on training data. We can then use μ and Σ to substitute each expected value by its conditional mean and substitute the variance of each entry by its respective entry in the conditional covariance matrix.

Specifically for the RBF network, we are interested in estimating the squared distance between a feature vector X_n and a centroid location C_l. Thus, the more general identity in Eq. (9) resumes to:

$$\mathrm{E}[\|X_n - C_l\|_2^2] = \|\mathrm{E}[X_n] - C_l\|_2^2 + \sum_{i=1}^{D} \mathrm{Var}[X_{n,i}] \tag{10}$$

which when used in the RBF kernel formula in order to approximate $K(X_n, C_l)$ yields

$$\begin{aligned} K(X_n, C_l) &\approx exp\Big\{ -\frac{\gamma}{2}\mathrm{E}[\|X_n - C_l\|_2^2] \Big\} \\ &= exp\Big\{ -\frac{\gamma}{2}\Big(\|\mathrm{E}[X_n] - C_l\|_2^2 + \sum_{i=1}^{D} \mathrm{Var}[X_{n,i}] \Big) \Big\} \\ &= exp\Big\{ -\frac{\gamma}{2}\|\mathrm{E}[X_n] - C_l\|_2^2 \Big\} \cdot exp\Big\{ -\frac{\gamma}{2} \sum_{i=1}^{D} \mathrm{Var}[X_{n,i}] \Big\} \end{aligned} \tag{11}$$

It is interesting to notice that the only real difference between the using the ESD to estimate $K(X_n, C_l)$ and the result obtained by the direct substitution of the entries in X_n by its EM estimates is the presence of a multiplicative term involving the variances in Eq. (11), which propagates some uncertainty on the operation being performed using the imputed values.

4 Performance Evaluation

Standard techniques to deal with missing data include imputing the missing entries with the sample mean or EM estimates. A more naive approach is to simply drop the examples that contain missing data. To evaluate the performance of the proposed algorithm, we compare it with RBFs trained on datasets treated with these techniques.

To carry the experiments, 4 different regression datasets from the UCI machine learning repository [12] were chosen. In addition to that, a dataset was generated using the DC motor simulated model available at Matlab's Simscape toolbox. The regression task is a DC motor parameter estimation problem. The input vector comprise a set of time delayed measurements of current, voltage and output torque. The output is the armature resistance value. The problem of armature resistance estimation may be used as part of a health monitoring procedure for DC motors. The generated problem is a nonlinear regression task with highly correlated attributes. Details about each data set are presented in Table 1.

Table 1. Datasets characteristics

Dataset	Attributes	Instances
Concrete compression	8	1030
Parkinson disease	21	5875
Breast cancer	32	194
Forest-fire	4	517
DC motor	30	1250

To simulate the impact of the number of missing values for each method, we gradually increase the number of examples with missing attributes in each data set. These attributes are deleted in a manner to guarantee that data are Missing at Random [1]. The experiments were repeated 10 times. At each trail, the data was split between training and test sets (70% and 30%, respectively).

For all experiments, the RBFs consisted of $L = 50$ centroids whose locations where chosen using k-means. Following a procedure described in [11], γ was set to $\sqrt{2L}/\mathcal{K}$, where $\mathcal{K}$ is the distance between the furthest centroid locations. The adopted performance measure was the Root Mean Square Error (RMSE).

5 Results

Figure 1 presents the RMSE for all datasets as the number of examples with missing data increase. Each entry in a plot is the mean RMSE achieved in 10 repetitions.

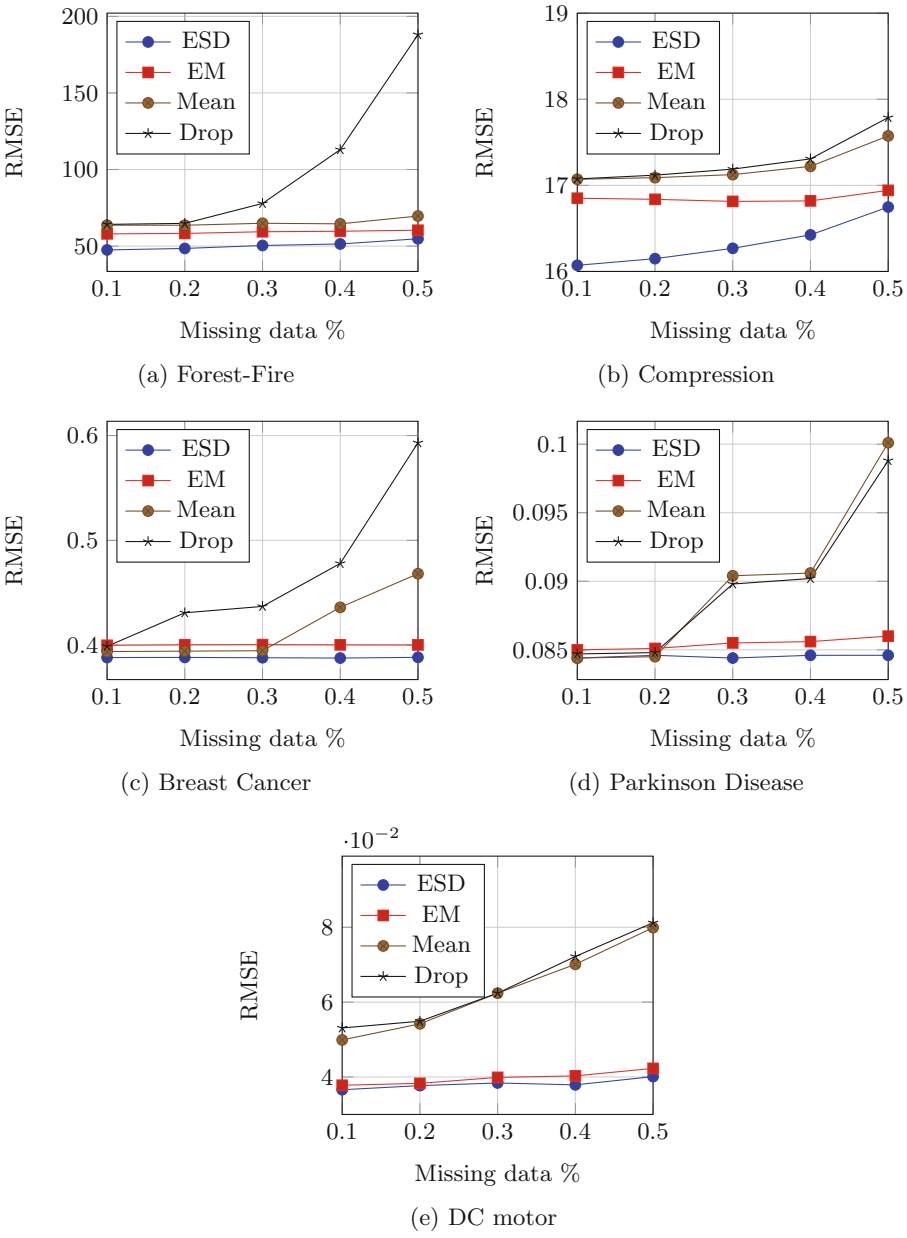


Fig. 1. RMSE for all datasets

As expected, more naive strategies lead worst results. Using the mean value for each missing feature (Mean) and discarding examples (Drop) achieved the worst results for all data sets, while using the ESD to approximate the RBF kernel (ESD) presented the best performance, followed by using the EM estimate (EM) to directly impute values.

It is interesting to remember the proposed approach relies on a model of the data, that was assumed to be multivariate normal. Therefore, there is reason to believe our method might profit from the use of a Gaussian Mixture Model as it usually presents better fit to data than a single Gaussian.

6 Conclusion

This work presented an extension of RBFNN for datsets with missing values. The proposed approach uses the Expected Squared Distance (ESD) as an intermediate step to estimate the value of RBF kernel between a centroid location and an input point, potentially with missing attributes. It is worth noting that the proposed algorithm can handle not only missing data on the training set but also capable of making decision under new data with missing attributes.

Results showed that RBF networks using ESD achieved the best overall results when compared to different standard strategies to deal with missing data such as sample mean imputation and EM imputation. Future works include the use of a Gaussian Mixture as a model of the training data as reported in [13] and the development of more precise procedures for RBF Kernel estimation.

Acknowledgments. The authors acknowledge the support of CNPq (Grant 402000/2013-7).

References

1. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 2nd edn. Wiley Interscience, New York (2002)
2. Oliveira, P.G., Coelho, A.L.V.: Genetic versus nearest-neighbor imputation of missing attribute values for RBF networks. In: Köppen, M., Kasabov, N., Coghill, G. (eds.) ICONIP 2008. LNCS, vol. 5507, pp. 276–283. Springer, Heidelberg (2009). doi:10.1007/978-3-642-03040-6_34
3. Qi, Y., Miche, Y., Eirola, E., van Heeswijk, M., Séverin, E., Lendasse, A.: Regularized extreme learning machine for regression with missing data. Neurocomputing **102**, 45–51 (2013)
4. Garcia-Laencina, P.J., Sancho-Gomez, J.-L., Figueiras-Vidal, A.R.: Pattern classification with missing data: a review. Neural Comput. Appl. **19**, 253–282 (2010)
5. Hruschka, E.R., Hruschka, E.R., Ebecken, N.F.F.: Evaluating a nearest-neighbor method to substitute continuous missing values. In: Gedeon, T.T.D., Fung, L.C.C. (eds.) AI 2003. LNCS (LNAI), vol. 2903, pp. 723–734. Springer, Heidelberg (2003). doi:10.1007/978-3-540-24581-0_62
6. Eirola, E., Doquire, G., Verleysen, M., Lendasse, A.: Distance estimation in numerical data sets with missing values. Inf. Sci. **240**, 115–128 (2013)
7. Meng, X.L., Rubin, D.B.: Maximum likelihood estimation via the ECM algorithm. Biometrika **80**, 267–278 (1993)
8. Pelckmans, K., De Brabanter, J., Suykens, J.A.K., De Moor, B.: Handling missing values in support vector machine classifiers. Neural Netw. **18**, 684–692 (2005)
9. de Souza Junior, A.H., Corona, F., Miche, Y., Barreto, G.A., Lendasse, A.: Minimal learning machine: a novel supervised distance based approach for regression and classification. Neurocomputing **164**, 34–44 (2015)

10. Mesquita, D.P.P., Gomes, J.P.P., Souza Jr., A.H.: A minimal learning machine for datasets with missing values. In: Arik, S., Huang, T., Lai, W.K., Liu, Q. (eds.) ICONIP 2015. LNCS, vol. 9489, pp. 565–572. Springer, Heidelberg (2015). doi:10.1007/978-3-319-26532-2_62
11. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. Pearson Education Press, Singapore (2001)
12. Frank, A., Asuncion, A.: UCI Machine Learning Repository University of California. School of Information and Computer Sciences, Irvine (2010)
13. Eirola, E., Lendasse, A., Vandewalle, V., Biernacki, C.: Mixture of gaussians for distance estimation with missing data. Neurocomputing **131**, 32–42 (2015)

Evaluation Method for an Adaptive Web Interface: GOMS Model

Rebai Rim(✉), Maalej Mohamed Amin, Mahfoudhi Adel, and Abid Mohamed

ENIS, Embedded Computer System Laboratory, University of Sfax, Sfax, Tunisia
rebai_rima@yahoo.fr, m.maalej@gmail.com, adel.mahfoudhi@fss.rnu.tn, mohamed.abid@enis.rnu.tn

Abstract. Adaptive web interfaces help users to perform tasks in an application and also construct a user's preferences model to best serve them in the future. This paper presents usability evaluation of adaptive web interface which focuses on how users can learn to achieve their goals. First, we present our adaptive Web interface using a Bayesian networks approach. Then, a formal GOMS model approach was applied to the evaluation of our user interface for a specialized web application. The evaluation shows that the adaptive user interface was more comfortable than the fixed user interface.

Keywords: User interface · Bayesian network · Evaluation · GOMS model

1 Introduction

Adapting a web interface to different types of users is important to improve the usability of applications. Two techniques used for adaptation are adaptive navigation and adaptive presentation. Adaptive navigation is to customize the ways in which users perform their tasks in the application. Adaptive presentation is to customize the content presented to the user. The user's preferences and the user's objectives are at the intersection of issues of adaptive presentation and adaptive interface. Thus, users have knowledge, objectives, preferences, abilities and learning style. The information about the user is called a user model [9,12,15].

We focus on the use of Bayesian models to make inferences about the preferences of users. We found that Bayesian networks can be used to represent uncertainty in user modeling [8,18] and can be effective in diagnosing a user's preferences. Thus, Bayesian user models provide an infrastructure for building new services in web applications. So, when users navigate through the Web interface, they don't encounter problems in finding the desired information. To evaluate the web application, many evaluation methods have been employed [5,12,20,24]. In this paper, we used GOMS model to evaluate our adaptive web interface. This adaptive user interface was constructed using a Bayesian networks approach. The results proved the effectiveness of Bayesian approach for user modeling.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_12

This paper was organized as follows: in Sect. 2, some existing work were introduced. The Bayesian user model was presented in Sect. 3. A discussion of the evaluation of our user interface using GOMS model is given in Sect. 4 and we conclude in Sect. 5.

2 Related Work

User models is used to adapt user interfaces, to present user's needs and user's preferences. They can be used for increasing the usability of services and for personalization purposes. User models has been the subject of many studies. The research field of user models began to grow in the 80's with the work in the Intelligent Tutoring Systems field [3,21].

Significant web applications have been taken in personalized services in books [17], online research papers [16] and e-commerce [26].

Bayesian network is as a effective approach for using uncertainty in context awareness [7]. Korpipaa et al. [13] have represented the user using the nave Bayes classifier. Hong et al. [10] have proposed an approach for a context sensitive messenger that deducts the objectives of the user. In the work of Horvitz [11], they have proposed a system that predicts user attention.

In a recent work [12], In-Jee Song et al. have constructed a model for the ubiquitous family environment to implement an adaptive user interface using Bayesian networks. In [14], using Bayesian approach, different systems have been proposed to predict the next teaching activity based on the level of knowledge of students.

The evaluation of interactive and adaptive interfaces taking into account the ergonomics of use become an important research field. Evaluation is to check that users can use the service and that they like it and can achieve their objectives. If the interactive interface meets the criteria (performance, usability, ...), it is accepted; otherwise, it must be used for a new evaluation.

The evaluation, in this view, is considered as a specific characteristic that influences the quality of a Web application. In recent years, many studies have interested in evaluations in the field of web applications [1,2,4].

In this paper, as result of evaluation with GOMS model, our Bayesian model offers a better correspondence between user's preferences and presented information. We were able to build a specific models in order to propose the web pages having the most adapted form and content for each user.

3 Bayesian Approach for User Modeling

The structure of the Bayesian network was constructed after a sufficient user data was obtained. Because the number of available users is small, we used the method of cross-validation. We defined the structure of Bayesian model using the simulated annealing algorithm [23]. Figure 1 shows the structure of our model.

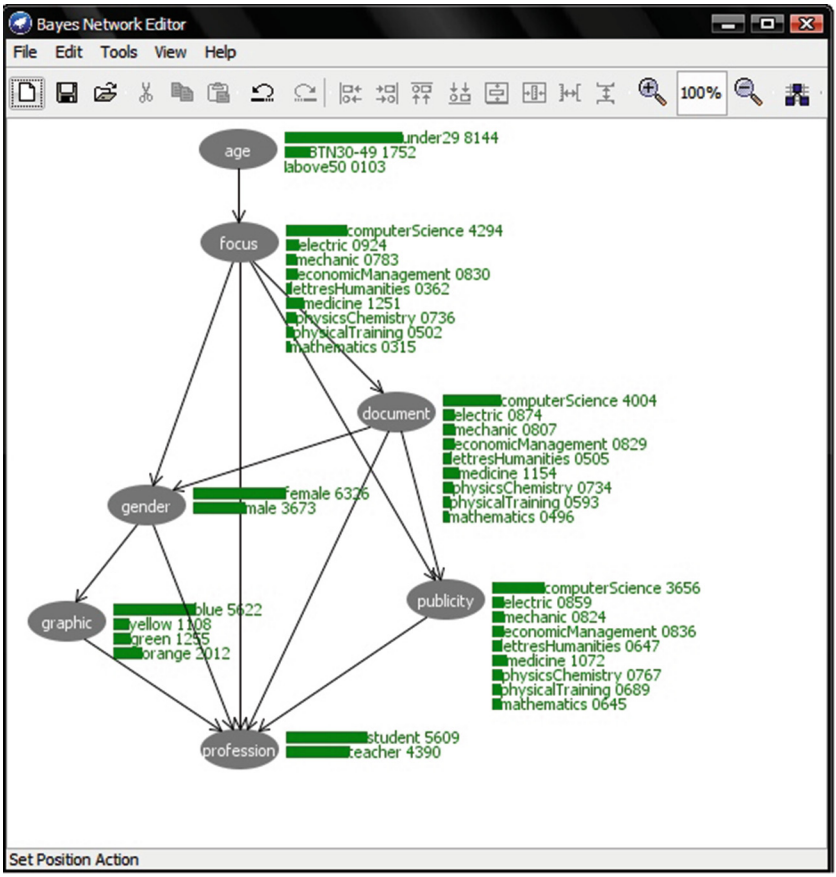


Fig. 1. The structure of Bayesian model.

The Bayesian user model can predict user's goals and user's preferences and calculate the adaptations to be made [19]. The collected data and the structure of the Bayesian model are used for the inference algorithm. The inference algorithm consists of how to choose the predicted preference from a probability distribution.

In our Bayesian user model, the variables profession, age, focus and gender are presented as evidence to infer user's preferences [22]. Thus, we construct this probability using the Junction Tree algorithm. An attribute with a higher probability have more functions to generate an adaptive user interface than an attribute with a lower probability. This adaptive user interface allows finding thesis topics, conferences, publishing courses and exchanging documents between teachers and students (Fig. 2).

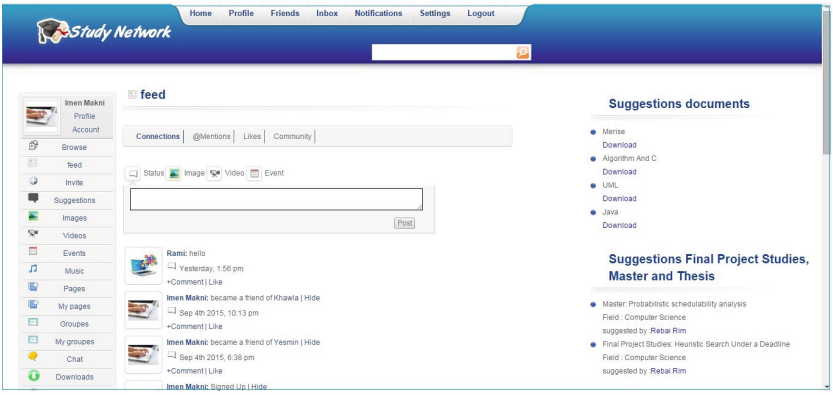


Fig. 2. The adaptive user interface

4 Evaluation Method: GOMS Model

In this section, we evaluate the proposed approach with GOMS model and we present the evaluation results. A GOMS model approach, proposed by Card, Moran, and Newell [6], is a description of the user's knowledge in order that a user perform tasks on a system. This approach has identified significant usability problems integrated into the procedures by which users interact with the interface.

To evaluate the usability of our adaptive interface, we assumed 5 situations that happen in the adaptive and fixed interface. 126 user's participated in this evaluation. Thus, using GOMS model, we evaluated the adaptive interface based on the required time to perform the predefined tasks in these situations. The situations and the associated tasks are presented in Table 1.

Then, we compared the required time for users to complete the tasks in these situations provided in Table 1 in the adaptive and fixed interface. The results of comparison are presented in Fig. 3. Table 2 presents the reduction rate of the time required when using the proposed approach in the adaptive user interface. Then, in the adaptive interface, we obtain 44.91% reduction rate in the time required compared to the fixed user interface.

The inference, in the Bayesian model, was conducted to infer the user's objectives (document, publicity, graphic, ...). An attribute with a higher probability is supposed to have more functions to generate an adaptive user interface than an attribute with a lower probability.

Then, we make a survey for each situation and we compared the result in the survey with the result of each situation completed in the proposed interface. The same 126 users was participated in the survey.

We calculated the percentage of correct attributes with high probability for each situation. We show in Fig. 4 that the results were good where the high probability of attributes were correctly presented with about 74%.

Table 1. The situations and detailed tasks for usability test.

No	Situation	Detailed task for situation
1	Find a conference	Click the icon suggestion → Search among the conferences proposed what is appropriate
2	Find a proposal for laboratory	Click the icon suggestion → Search among the laboratories proposed what is appropriate
3	Download a course	Click the icon document → Search among the courses offered what is appropriate → Download the course
4	Find a proposal for thesis	Click the icon suggestion → Search among the thesis topics proposed what is appropriate
5	Find pages or groups	Click the icon pages or groups → Give the specialty of study → Search among the pages or the groups offered what is appropriate

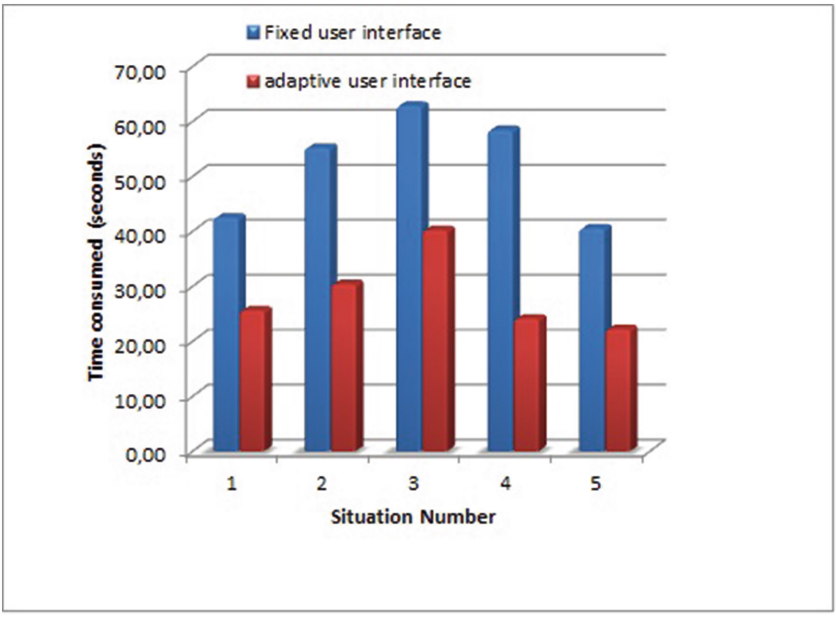


Fig. 3. Time to complete the tasks in the situations.

Table 2. Reduction rate of the time intended to perform the tasks.

No. situation	1	2	3	4	5	Avg
Reduction rate of time required	39.76%	44.92%	35.98%	58.73%	45.18%	44.91%

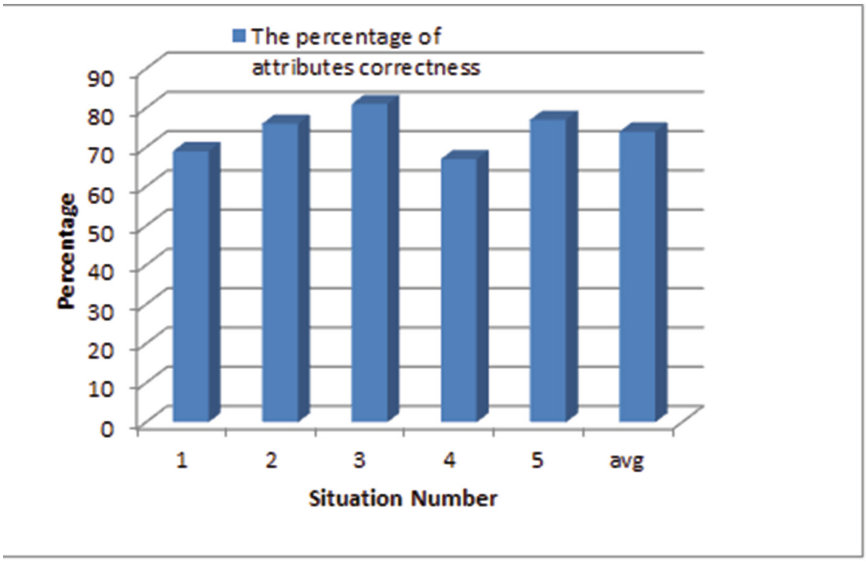


Fig. 4. The percentage of attributes correctness.

To test the effectiveness of our approach, we use also measures that are applied to data sets: precision, recall and F1 Metrics [25]. The definition of these measures is:

$$Recall = \frac{\text{True positive}}{\text{True positive} + \text{false negative}} \tag{1}$$

$$Precision = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \tag{2}$$

$$F1 = \frac{2 * \text{Recall} * \text{Presicion}}{\text{Recall} + \text{Presicion}} \tag{3}$$

In our work, the true positive is the number of users who have correctly inferred preferences. Figure 5 shows that the best results were obtained with the adaptive user interface. Using the Bayesian approach, the F1 measure is 90%. Thus, we can conclude that the Bayesian approach provides a maximum F1 and a significant result in the quality of the web interface.

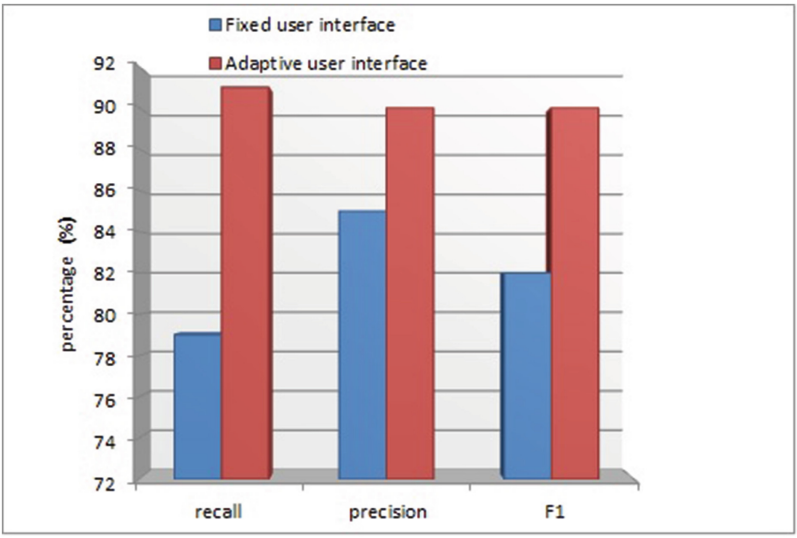


Fig. 5. Measures of both the adaptive and fixed user interface.

5 Conclusion

In this paper, we used Bayesian networks to compute the probability distribution over the expected preferences. Then, we evaluated the adaptive user interface with the GOMS model. Comparing the fixed user interface to the adaptive user interface, we found that the proposed interface was more comfortable for use than the fixed user interface and we achieved a 45% reduction rate in the time required to complete the tasks when using the adaptive user interface.

Then, we needed to improve the model with a better training. A log of the users' activities was used to specify the Bayesian network. In order to improve the system evaluation we will include some qualitative measures on the user experience and satisfaction from a verbal user feedback.

References

1. Alva, M.E.O., Martínez, P.A.B., Cueva, L.J.M., Sagástegui, C.T.H., López, P.B.: Comparison of methods and existing tools for the measurement of usability in the web. In: Lovelle, J.M.C., Rodríguez, B.M.G., Gayo, J.E.L., Puerto Paule Ruiz, M., Aguilar, L.J. (eds.) ICWE 2003. LNCS, vol. 2722, pp. 386–389. Springer, Heidelberg (2003). doi:10.1007/3-540-45068-8_70
2. Batra, S., Bishu, R.R.: Web usability and evaluation: issues and concerns. In: Aykin, N. (ed.) UI-HCII 2007. LNCS, vol. 4559, pp. 243–249. Springer, Heidelberg (2007). doi:10.1007/978-3-540-73287-7_30
3. Brown, A.L.: Learning, remembering and understanding. In: Flavell, J.H., Markham, E.M. (eds.) Carmichael's Manual of Child Psychology. Wiley, New York (1982)

4. Cunliffe, D.: Developing usable web sites - a review and model. Internet Res. Electron. Networking Appl. Policy **10**, 295–307 (2000)
5. Card, S.K., Pirolli, P., Van Der Wege, M.M., Morrison, J.B., Reeder, R.W., Schraedley, P.K., Boshart, J.: Information scent as a driver of Web behavior graphs: results of a protocol analysis method for Web usability. In: Proceedings of ACM CHI 2001, pp. 498–504 (2001)
6. Card, S., Moran, T.P., Newell, A.: The Psychology of Human Computer Interaction. Lawrence Erlbaum Associates, Hillsdale (1983)
7. Dey, A.K.: Understanding and using context. Pers. Ubiquit. Comput. **5**, 2024 (2001)
8. Fink, J., Kobsa, A.: User modeling for personalized city tours. Artif. Intell. Rev. **18**, 33–74 (2002)
9. Fischer, G.: User modeling in human-computer interaction. J. User Model. User-Adap. Inter. **11**, 65–86 (2001)
10. Hong, J.H., Yang, S.I., Cho, S.B.: ConaMSN: a context aware messenger using dynamic Bayesian networks with wearable sensors. Expert Syst. Appl. **37**, 4680–4686 (2010)
11. Horvitz, E., Kadie, C.M., Paek, T., Hovel, D.: Models of attention in computing and communications: from principles to applications. Commun. ACM **46**(3), 52–59 (2003)
12. In-Jee, S., Sung-Bae, C.: Bayesian and behavior networks for context-adaptive user interface in a ubiquitous home environment. Int. J. Expert Syst. Appl. **40**, 1827–1838 (2013)
13. Korpipaa, P., Koskinen, M., Peltola, J., Makela, S.M., Seppanen, T.: Bayesian approach to sensor-based context awareness. Pers. Ubiquit. Comput. **7**(2), 113–124 (2003)
14. Mayo, M., Mitrovic, A.: Optimising ITS behaviour with Bayesian networks and decision theory. Int. J. Artif. Intell. Educ. **12**, 124–153 (2001)
15. McTear, M.F.: User modelling for adaptive computer systems: a survey of recent developments. Artif. Intell. Rev. **7**, 157–184 (1993)
16. Middleton, S., Shadbolt, N., De Roure, D.: Ontological user profiling in recommender systems. ACM Trans. Inf. Syst. **22**, 54–88 (2004)
17. Mooney, R.J., Roy, L.: Content-based book recommending using learning for text categorization. In: Proceedings of the Fifth ACM Conference on Digital Libraries, San Antonio, TX, pp. 195–204. ACM (2000)
18. Nguyen, L., Do, P.: Combination of Bayesian network and overlay model in user modeling. In: Allen, G., Nabrzyski, J., Seidel, E., Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2009. LNCS, vol. 5545, pp. 5–14. Springer, Heidelberg (2009). doi:10.1007/978-3-642-01973-9_2
19. Pearl, J.: Probabilistic Reasoning in Intelligent Systems – Networks of Plausible Inference. Morgan Kaufman, San Mate (1988)
20. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, Chapel Hill, NC, pp. 175–186 (1994)
21. Rich, E.: Stereotypes and user modeling. In: Kobsa, A., Wahlster, W. (eds.) User Models in Dialog Systems, Chap. 2. Symbolic Computation (Artificial Intelligence), pp. 38–51. Springer, Heidelberg (1989)
22. Rim, R., Mohamed Amin, M., Adel, M.: Bayesian networks for user modeling: predicting the user's preferences. In: International Conference on Hybrid Intelligent Systems, HIS 2013, pp. 144–148 (2013)

23. Rim, R., Mohamed Amin, M., Adel, M.: Building and evaluating an adaptive user interface using a Bayesian network approach. Int. J. Comput. Sci. Inf. Secur. **7**, 548–565 (2016)
24. Scholtz, J., Laskowski, S., Downey, L.: Developing usability tools and techniques for designing and testing Web sites. In: Proceedings of HFWeb, June 1998
25. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley, Boston (2005). Commerce
26. Wei, K., Huang, J., Fu, S.: A survey of e-commerce recommender systems. In: Proceedings of the International Conference on Service Systems and Service Management, Chengdu, China, pp. 1–5 (2007)

Diversification Strategies in Differential Evolution Algorithm to Solve the Protein Structure Prediction Problem

Pedro Henrique Narloch and Rafael Stubs Parpinelli(✉)

Graduate Program in Applied Computing, Departament of Computer Science, State University of Santa Catarina, Joinville, Brazil
pedro.narloch@gmail.com, rafael.parpinelli@udesc.br

Abstract. The protein structure prediction is considered as one of the most important open problems in biology and bioinformatics due the huge amount of plausible shapes that a protein can assume. The objective of this paper is to apply the Differential Evolution (DE) algorithm employing two simple diversification strategies known as generation gap and Gaussian perturbation to solve the protein structure prediction problem in the backbone and side-chain model. To test our approaches the 1PLW, 1ZDD and 1CRN proteins were used and the standard DE algorithm was compared with DE using the diversification approaches and with some state-of-art algorithms. Also, the genotypic diversity was analyzed during the algorithm run, showing the impacts generated by the diversification mechanisms. Despite its simplicity, the proposed approaches achieved competitive results.

Keywords: Protein structure prediction · Bioinformatics · Differential evolution

1 Introduction

Proteins are macromolecules which have important biological functions in every living organism when the tree-dimensional conformation is reached. These macromolecules are composed by an unique amino acid sequence, known as the primary structure, which influences the protein to fold into a three-dimensional shape [2]. Nowadays, the methods to determine an existing protein's tertiary structure are the nuclear magnetic resonance and the crystallography X-ray [9]. Although these methods can determine the native conformation of a determined protein, they are too expensive [1].

Different representations were created to solve the protein structure prediction (PSP) problem. The prediction only by the amino acid sequence is called *Ab Initio* prediction and it's one of the most challenging problems in bioinformatics because of its complexity even for small proteins [10]. The high complexity associated with this problem is due the huge amount of plausible shapes that a

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_13

protein can assume. Hence, the protein structure prediction (PSP) is labeled as a NP-complete problem [11].

Due limitations of exact algorithms to solve this class of problems, metaheuristics became a viable way to explore the search space and find possible conformations in a plausible time. In recent literature, different approaches of Evolutionary Computing (EC) algorithms have been used to solve the PSP problem in atomic representations as in [2,6,9].

The standard Differential Evolution (DE) algorithm is a population-based EC algorithm that has been chosen to solve the PSP problem in the present work. The DE algorithm is considered a good algorithm to solve problems from continuous optimization [14]. However, it is known that it loses its diversity very quickly, increasing the chance of getting stuck in a local optimum when employed in a high multimodal problem like the PSP. Hence, with diversity control strategies it is possible to slow down the convergence, aiming to escape from local optima [4]. In this work, two simple diversification strategies are used with the DE algorithm: the Generation Gap and the Gaussian Perturbation.

Four different approaches were used to solve the PSP problem. One of them is the standard Best/1/Bin DE algorithm, while others are called as Generation Gap (GG), Gaussian Perturbation (GP) and the combination of them (GG-GP). To verify the efficiency of the proposed approaches, the genotypic diversity is analysed. With this analysis it is possible to verify the behaviour of algorithms and the impact of such strategies in the results obtained. Furthermore, results are compared with recent literature which used the same representation model and the same energy function.

The next sections are organized as follows. In Sect. 2 the PSP problem is described and related works are discussed. Section 3 explains the DE approaches to solve the PSP problem. Section 4 exhibit the results obtained in our test cases. Finally, Sect. 5 contains the conclusion of this work and some future directions.

2 Protein Structure Prediction

Proteins are made from amino acids chains where each amino acid is composed by an amino group (H_3N^+), a carboxyl group (COO^-) and a hydrogen atom attached to a central carbon (C_α) [2]. Each amino acid has a side chain attached to the C_α, distinguishing each one of the 20 different amino acids known in nature.

A protein can be depicted into four different well defined structures. The primary structure is formed by a linear sequence of amino acids, the secondary structure represents the local structure found in the backbone conformation, the tertiary structure considers the protein's final conformation (including the amino acids side chain) and determine its biological function. The quaternary structure represents interactions among proteins to accomplish specific functions.

To evaluate if a protein is near its native state, the Anfinsen's thermodynamic hypothesis declares that a native three-dimensional protein shape has the lowest

free energy. In the present work, the energy is obtained by the CHARMM force field [3] which is one of the most popular energy functions [10] shown in Eq. 1.

$$E_{total} = \sum_{bonds} K_b(b - b_0)^2 + \sum_{UB} K_{UB}(S - S_0)^2 + \sum_{angle} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\chi(1 + \cos(\eta - \delta)) + \sum_{impropers} K_{imp}(\varphi - \varphi_0)^2 + \sum_{nonbond} \epsilon[(\frac{R_{minij}}{R_{ij}})^{12} - (\frac{R_{minij}}{R_{ij}})^6] + \frac{q_i q_j}{\epsilon_1 r_{ij}} \quad (1)$$

From Eq. 1, $\mathbf{E_{total}}$: is the total energy value; **bonds** measures the energy according to the bond stretching between two atoms; Urey-Bradley (**UB**) represents the interactions between pairs of atoms; **angle** is the sum among all angles in the structure; **dihedrals** is the energy associated with the torsion angles; **impropers** values are associated to deformations of improper torsion angles; **nonbond** values are related to Van der Waals and Charge-Charge energy. Van der Waals is the energy from interactions between nonbonded angles from attraction and repulsion. Charge-Charge varies according to the distance among atoms.

Different types of protein's atomic representations emerged with different levels of abstraction. Some commonly used are: **(a)** all-atom three-dimensional coordinates; **(b)** all-heavy-atom coordinates; **(c)** backbone atom coordinates + side-chain centroids; **(d)** C_α coordinates; **(e)** backbone and side-chain torsion angles.

As this work employs the backbone and side-chain torsion angles model, it is known that each residue has a defined number of torsion angles that is needed to be optimized. Each amino acid has three backbone angles (ϕ, ψ, and ω) and a particular number of side chain angles (χ_i) as shown in Table 1. A backbone classification was employed to identify the secondary structure and

Table 1. χ angles for each amino acid

Aminoacid	χ angles
GLY, ALA, PRO	Backbone
SER, CYS, THR, VAL	χ_1
ILE, LEU, ASP, ASN, PHE, TYR, TRP	χ_1, χ_2
MET, GLU, GLN	χ_1, χ_2, χ_3
LYS, ARG	χ_1, χ_2, χ_3, χ_4

Table 2. DSSP 8-class classification.

Secondary structure	ϕ bounds	ψ bounds
H (α-helix)	$[-67°, -47°]$	$[-57°, -37°]$
B (β-bridge)	$[-130°, -110°]$	$[110°, 130°]$
E (β-strand)	$[-130°, -110°]$	$[110°, 130°]$
G (3-10-helix)	$[-59°, -39°]$	$[-36°, 16°]$
I (pi-helix)	$[-67°, -47°]$	$[-80°, -60°]$
T (turn)	$[-180°, 180°]$	$[-180°, 180°]$
S (bend)	$[-180°, 180°]$	$[-180°, 180°]$
U (undefined)	$[-180°, 180°]$	$[-180°, 180°]$

the recommended bounds for each type of structure. These angles are shown in Table 2 and are based on the full DSSP 8-class classification [12].

2.1 Related Works

Some related works apply bio-inspired algorithms, *Ab Initio* prediction and CHARMM energy function. In [9] was applied a GA with two different approaches for diversity control: the random immigrants technique, which replaces a percentage of individuals from the population for new randomly generated individuals, and an extension called simplified self-organizing random immigrants with dynamic replacement rate. In [13] A bacterial foraging optimization algorithm (BFOA) was applied for the 1PLW protein using CHARMM energy function. Although most of related works tried to improve the diversity, in [17] a GA was combined with Hill Climbing in a parallel grid environment, improving the exploitation capacity.

In [15] the NSGA-II was employed to solve the PSP problem using island models. Also, a modified PAES algorithm is proposed in [7,8] with immune inspired operators (cloning and hyper mutation). Another multi-objective approach was proposed in [18] using the DE algorithm. However, the DE was modified to be adaptive varying the mutation mechanism. This technique is known as probability matching, associating a percentage at each DE version to execute the mutation process. If the mutation process is successful, than its chance to be selected again increases. This is the approach that found the lowest energy value using CHARMM for the 1PLW, 1ZDD and 1CRN proteins. The multi-objective formulation in these works is slicing the CHARMM energy function in two terms: bonded and non-bonded.

3 Methods

In this work each individual is formed by a set of angles representing amino acids. An individual is structured as a vector and its size changes according to the number of amino acids in each protein. Figure 1 illustrates the structure of an individual for the 1PLW protein which have 5 amino acids. Note that the ω angle is not in our representation because its value is set always to 180°.

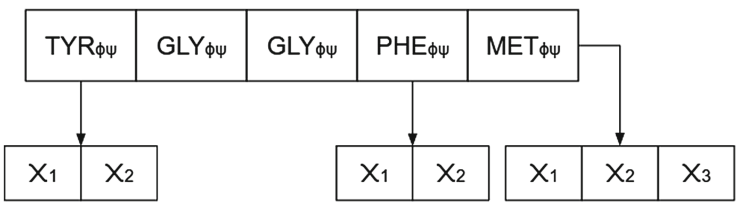


Fig. 1. Graphical representation of 1PLW individual.

The standard DE algorithm is population-based and at each new iteration an offspring is generated by a mutation operator to replace the current population if it achieves a better solution. In order to modify this routine and improve the diversity in the population, we have used the generation gap mechanism [16]. This mechanism is commonly used in EC and only a fraction of the population is replaced by the offspring according to a parameter G which varies between 0 and 1. The maintained individuals are selected at random from the current population. This model is called generation gap DE (DE_{GG}).

To create new individuals, the DE algorithm uses a mutation mechanism, combining values from different individuals. There are different approaches for mutation in DE and, in this work, the $\text{DE}_{Best/1/Bin}$ version was selected. This standard version of DE always uses the best individual in the population to combine with two other random individuals. The $\text{DE}_{Best/1/Bin}$ approach creates a new individual using $\boldsymbol{w} = \boldsymbol{x_{best}} + F \cdot (\boldsymbol{x_{rand1}} - \boldsymbol{x_{rand2}})$, where $\boldsymbol{w}$ represents the new generated individual and F is a threshold that need to be set.

Hence, another modification was done in the mutation operator of the standard $\text{DE}_{Best/1/Bin}$ algorithm. For the two randomly selected individuals, a gaussian perturbation technique is applied. With the gaussian perturbation, $\boldsymbol{x_{rand1}}$ and $\boldsymbol{x_{rand2}}$ are considered the mean and the standard deviation is defined between 0 and 1. This model is called gaussian perturbation DE (DE_{GP})

To verify how these approaches impact the diversity of solutions during the optimization process, this work uses a genotypic diversity measure for continuous domains [5]. The Eq. 2 shows how to calculate the genotypic diversity.

$$GDM = \frac{\sum_{i=1}^{N-1} ln\left(1 + \min_{j[i+1,N]} \frac{1}{D}\sqrt{\sum_{k=1}^{D}(x_{i,k} - x_{j,k})^2}\right)}{NMDF} \tag{2}$$

where D is the size of the solution vector, N is the population's size and x the individual (or solution vector). The NMDF is a normalization factor which corresponds to the maximum diversity value so far. The genotypic diversity starts with 1 which is the maximum value and when it reaches 0 it corresponds to the full convergence of the population. With this measurement it is possible to verify the diversity level during each iteration. This is an important measure to verify if the algorithm is getting trapped in a local optima and, consequently, getting a premature convergence.

Besides the function evaluation given by CHARMM and the genotypic diversity measure given by Eq. 2, there is another important metric which is considered in this work: the root mean square deviation (RMSD). The RMSD is a measure given in Å(angstrom) which compares the atomic distance between proteins and verifies if the final predicted conformation reached the native conformation. When the RMSD is near 0 means that the predicted protein is very similar to the native protein.

4 Experiments, Results and Analysis

In the current work three different proteins were used as problem instances: 1PLW, 1ZDD and 1CRN. The smallest protein used is known as *Met-Enkephalin* (1PLW) with only 5 amino acids and 22 angles to be optimized, without any well defined secondary structure. The 1ZDD is a protein which have two well defined α-helices structures and it contains 34 amino acids with 179 angles to be optimized. The biggest protein used in this work has 46 amino acids and 191 angles to be optimized, known as 1CRN. The 1CRN protein has two well defined α-helices and two β-sheets as secondary structures.

The experiments were conducted with 4 different algorithm configurations: the standard $\text{DE}_{Best/1/Bin}$, the standard DE with generation gap mechanism (DE_{GG}), the standard DE with gaussian perturbation (DE_{GP}) and DE_{GG-GP} which combines the standard DE with generation gap and gaussian perturbation. This work also compares the results obtained with another works found in the literature. All approaches use the atomic representation and the CHARMM energy function calculated with Tinker Molecular Dynamics Package.

The DE parameters used in this work are recommended by [18], with a population size of 100 individuals, the mutation factor (F) is set to 0.5, the crossover factor is 1 and the number of function evaluations is 500.000. The parameters G, which controls the generation gap was empirically set to 0.8, and the GSD which is responsible for the standard deviation was empirically set to 0.1.

For each protein and each approach 10 runs were done. Table 3 contains the results obtained for 1PLW, 1ZDD and 1CRN proteins.

The first column indicates each protein and the second column identifies each algorithm. Column 3 represents the minimum energy found in all runs and column 4 the RMSD_α from each minimum energy. Finally, column 5 represents the average minimum energy with the standard deviation. All DE approaches were developed using C++ language in an Intel core i7 with 8GB RAM.

For *Met-Enkephalin* (1PLW) all four developed DE approaches got similar results. However, the DE_{GG-GP} reached -35.82 kcal mol^{-1} with RMSD_α of 1.98Å. These values are competitive with the state-of-art ADEMO/D algorithm proposed in [18]. The lowest RMSD_α was reached by NSGA-II [15] with 1.26Å.

For 1ZDD protein the results obtained showed significant differences among all four DE approaches. It is possible to notice that the DE_{GP} and DE_{GG-GP} reached better results for minimum energy values when compared with standard $\text{DE}_{Best/1/Bin}$ and DE_{GG}. The DE_{GP} reached $-1,216.40$ kcal mol^{-1} with a RMSD of 2.36 Å, becoming competitive with the state-of-art algorithm found in literature known as ADEMO/D [18] and NSGA-II [15].

Analysing the results for 1CRN protein, the best approach was the DE_{GP} with energy value of 166.83 kcal mol^{-1}. Comparing with the four DE approaches developed in this work, the DE_{GP} was better than $\text{DE}_{Best/1/Bin}$ when considering the energy value and the standard deviation showing that the gaussian perturbation improved the results obtained. Comparing the DE_{GP} with ADEMO/D [18], our approach achieved lower energy besides the RMSD_α was bigger than the results found in the literature.

Table 3. Results obtained.

Protein	Version	Min. energy	$RMSD_\alpha$	Avg. energy
1PLW	$DE_{Best/1/Bin}$	−34.69	1.90Å	-28.58 ± 3.00
	DE_{GG}	−33.95	1.99Å	-26.35 ± 2.77
	DE_{GP}	−32.10	1.63Å	-27.96 ± 1.91
	DE_{GG-GP}	**-35.82**	1.98Å	-30.47 ± 4.44
	ADEMO/D [18]	−30.43	1.77Å	–
	BFOA [13]	−19.10	3.60Å	–
	I-PAES [7]	−20.56	2.83Å	–
	NSGA-II [15]	−22.73	**1.26Å**	–
	SSORIGA [9]	42.82	–	46.23 ± 1.64
1ZDD	$DE_{Best/1/Bin}$	−955.68	2.65Å	-508.52 ± 262.99
	DE_{GG}	−796.68	4.25Å	$95.03 \pm 1,503.92$
	DE_{GP}	$-1,216.40$	2.36Å	$-1,086.99 \pm 105.74$
	DE_{GG-GP}	$-1,156.95$	5.69Å	-983.97 ± 119.80
	ADEMO/D [18]	**-1,301.38**	**2.14Å**	–
	I-PAES [8]	$-1,052.09$	2.27Å	–
	NSGA-II [15]	$-1,218.57$	3.81Å	–
1CRN	$DE_{Best/1/Bin}$	818.04	7.89Å	$2,483.87 \pm 3,424.94$
	DE_{GG}	594.07	13.12Å	$1,608.72 \pm 1,152.38$
	DE_{GP}	**166.83**	10.71Å	288.52 ± 86.67
	DE_{GG-GP}	260.12	8.60Å	464.60 ± 133.59
	ADEMO/D [18]	253.25	6.06Å	–
	I-PAES [8]	509.09	**4.43 Å**	–
	NSGA-II [15]	262.68	7.32Å	–
	SSORIGA [9]	503.56	–	535.09 ± 20.98

Figure 2 shows both the energy and the genotypic diversity convergence over generations for all developed DE approaches. Note that all approaches converged very quickly for 1PLW protein. At generation 1,000 the energy function stabilized. However, when the genotypic diversity is plotted, there is diversity in the population after the generation 1,000 for DE_{GG-GP} while for $DE_{Best/1/Bin}$ the diversity ended earlier, given no possibility to create different individuals. Because of the small size of this protein, even using a diversity control mechanism the $DE_{Best/1/Bin}$ and DE_{GG-GP} got very similar values.

Analysing the convergence for 1ZDD protein, the energies of $DE_{Best/1/Bin}$ and DE_{GG} stabilizes around generations 2,500 and 3,000, respectively, while the energies of DE_{GP} and DE_{GG-GP} are still decreasing at generation 5,000, when all algorithms end. This behaviour is related with the diversity in the population. Note that for the approaches which have converged earlier, worst energy values were obtained and the diversity was lost prematurely. However, for DE_{GP} and

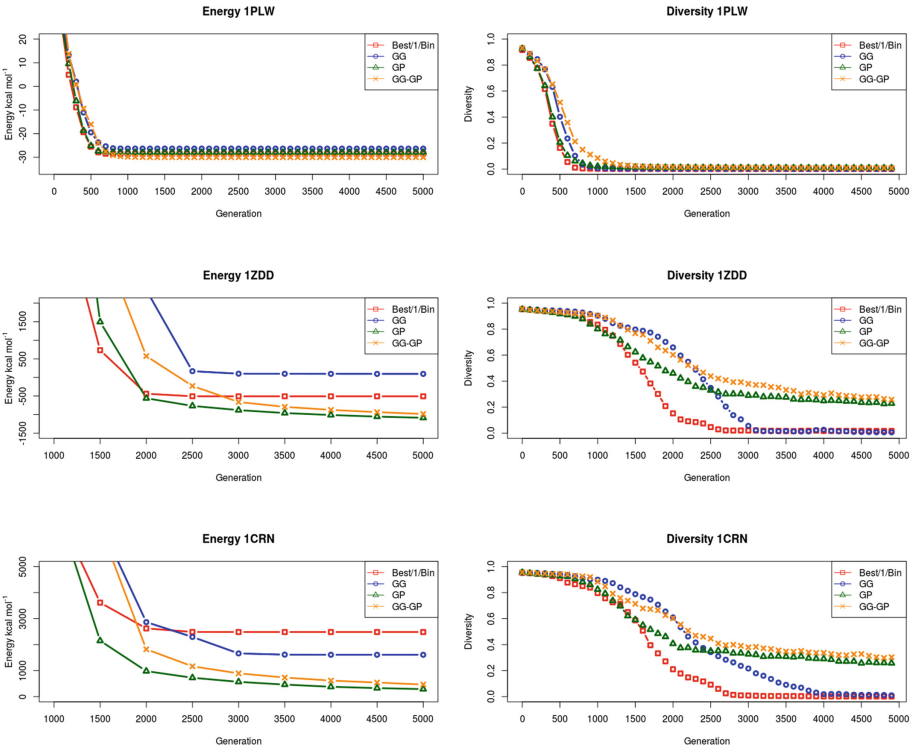


Fig. 2. Energy (left) and genotypic diversity (right) for each sequence.

DE_{GG-GP}, that have reached better energy results, the diversity is not over in the last generation.

The convergence analysis made for 1ZDD protein can also be made for 1CRN protein, where the approaches with diversity maintenance routines achieved better results avoiding premature convergence.

Overall the diversification techniques helped the standard DE to obtain lower average energy values mainly for the two biggest instances (1ZDD and 1CRN). It was verified through the GDM index that the genotypic maintenance is an important factor that need to be considered in PSP problem.

5 Conclusions and Future Research

This work applied four different DE approaches to solve some PSP problem instances using the torsion angles model and the CHARMM energy function. Two diversification strategies were used in order to avoid premature convergence: the generation gap and the gaussian perturbation.

Despite there are many works in the literature solving some PSP problem, none of them made an analysis of the diversity of solutions during the optimization process. As proposed in this work, the genotypic diversity was analyzed using

the GDM index. With this index was possible to relate the genotypic diversity with the energy convergence, verifying that the versions in which maintained the diversity got better results in all three proteins.

Although the genotypic diversification strategies increased the population's diversity, the gaussian perturbation always got the best energy values in comparison with the standard DE and the generation gap version. All four algorithms were also compared to state-of-art algorithms found in literature that used CHARMM as energy function. The DE_{GP} version showed to be competitive in all three proteins, 1PLW, 1ZDD and 1CRN, even being a much simpler approach than the works found in the literature.

Also, it was verified that when the algorithm ends, the diversity for bigger proteins is about 40%. This indicates that exploitation routines, like local search algorithms could be used to explore this diversity aiming to reach better energy values. Another future research could be the use of GPUs for energy minimization, possible granting higher speed ups when comparing with CPU approaches, providing bigger amounts of function evaluations and longer convergences.

As our developed approaches showed to be much simpler than the literature ones and reached competitive results, it is possible to use a famous Occam's razor statement: when you have two competing theories that make exactly the same predictions, the simpler one is the better.

References

1. Benítez, C.M.V., Parpinelli, R.S., Lopes, H.S.: An ecologically-inspired parallel approach applied to the protein structure reconstruction from contact maps. In: Genetic and Evolutionary Computation Conference, GECCO 2016, Denver, CO, USA, pp. 1299–1306 July, 2016
2. Borguesan, B., e Silva, M.B., Grisci, B., Inostroza-Ponta, M., Dorn, M.: APL: an angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. Comput. Biol. Chem. **59**, 142–157 (2015)
3. Brooks, B.R., Brooks, C.L., Mackerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A.R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R.W., Post, C.B., Pu, J.Z., Schaefer, M., Tidor, B., Venable, R.M., Woodcock, H.L., Wu, X., Yang, W., York, D.M., Karplus, M.: CHARMM: the biomolecular simulation program. J. Comput. Chem. **30**(10), 1545–1614 (2009)
4. Corriveau, G., Guilbault, R., Tahan, A., Sabourin, R.: Review and study of genotypic diversity measures for real-coded representations. IEEE Trans. Evol. Comput. **16**(5), 695–710 (2012)
5. Corriveau, G., Guilbault, R., Tahan, A., Sabourin, R.: Review of phenotypic diversity formulations for diagnostic tool. Appl. Soft Comput. **13**(1), 9–26 (2013)
6. Custodio, F.L., Barbosa, H.J., Dardenne, L.E.: A multiple minima genetic algorithm for protein structure prediction. Appl. Soft Comput. **15**, 88–99 (2014)
7. Cutello, V., Narzisi, G., Nicosia, G.: A multi-objective evolutionary approach to the protein structure prediction problem. J. Roy. Soc. Inter. **3**(6), 139–151 (2006)

8. Cutello, V., Narzisi, G., Nicosia, G.: Computational studies of peptide and protein structure prediction problems via multiobjective evolutionary algorithms. In: Knowles, J., Corne, D., Deb, K., Chair, D.R. (eds.) Multiobjective Problem Solving from Nature. Natural Computing Series, pp. 93–114. Springer, Heidelberg (2008)
9. Do, O., Tragante, V., Tinos, R.: A self-organizing genetic algorithm for protein structure prediction. Learn. Nonlinear Models **8**(3), 135–147 (2010)
10. Dorn, M., e Silva, M.B., Buriol, L.S., Lamb, L.C.: Three-dimensional protein structure prediction: Methods and computational strategies. Comput. Biol. Chem. **53**, 251–276 (2014)
11. Guyeux, C., CoTe, N.M.L., Bahi, J.M., Bienia, W.: Is protein folding problem really a NP-Complete one? first investigations. J. Bioinform. Comput. Biol. **12**(01), 1350017 (2014)
12. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**(12), 2577–2637 (1983)
13. Pal, A.: Ab-initio protein structure prediction using bacterial foraging optimization algorithm. Ph.D. thesis, Jadavpur University KOLKATA (2014)
14. Price, K.V., Storn, R.M., Lampinen, J.A.: Differential Evolution: A Practical Approach to Global Optimization. Natural Computing Series. Springer, Berlin, New York (2005)
15. Romero, D.C.B.: A multi-objective Ab-initio model for protein folding prediction at an atomic conformation level. Ph.D. thesis, Universidad Nacional de Colombia. Facultad de Ingeniera. Departamento de Ingeniera de Sistemas y Computacin (2010)
16. Sarma, K.: Generation gaps revisited. Found. Genet. Algorithms (FOGA 2) **2**, 19 (1993)
17. Tantar, A.A., Melab, N., Talbi, E.G., Parent, B., Horvath, D.: A parallel hybrid genetic algorithm for protein structure prediction on the computational grid. Future Gener. Comput. Syst. **23**(3), 398–409 (2007)
18. Venske, S.M., Gonalves, R.A., Benelli, E.M., Delgado, M.R.: ADEMO/D: an adaptive differential evolution for protein structure prediction problem. Expert Syst. Appl. **56**, 209–226 (2016)

Using Cluster Barycenters for the Generalized Traveling Salesman Problem

Mehdi El Krari[1,2(✉)], Belaïd Ahiod[1,3], and Bouazza El Benani[1,2]

[1] Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco
mehdi@elkrari.com
[2] Computer Science Laboratory,
Mohammed V University in Rabat, Rabat, Morocco
[3] LRIT, Associated Unit to CNRST (URAC 29),
Mohammed V University in Rabat, Rabat, Morocco

Abstract. We propose in this paper a novel idea to handle a tour for the Generalized Traveling Salesman Problem (GTSP), which is an NP-hard optimization problem very solicited for its numerous applications. Knowing that for each instance, cities are grouped in clusters. The proposed method finds for each one its barycenter in order to get in a first phase a good order of visiting clusters. Then, it uses one of the well-known methods to choose a city from each cluster. The obtained solution can be a good starting tour that can be used as an input for improvement methods. Our work is validated with some practical tests on benchmark instances. Obtained results show that our method gives feasible solution instantly.

Keywords: Generalized Traveling Salesman Problem · Clustering · Initialization · Barycenter · Combinatorial optimization problem

1 Introduction

The Generalized Traveling Salesman Problem (GTSP) is a generalization of the famous Traveling Salesman Problem (TSP) where the traveler must move to all nodes of a given graph. For the GTSP, we have a set V of n nodes partitioned over m clusters $(C_1, C_2, ..., C_m)$, for each pair of nodes (x, y) of V a distance $c(x, y)$ separates them. The goal is to visit from each cluster C_i one node and go back at the end to the first cluster/node visited while having the shortest tour. Knowing that the TSP is an NP-hard problem, the GTSP is NP-hard too since we can consider the TSP as a special case of the GTSP where all clusters contain only one node, $|C_i| = 1, 1 \leqslant i \leqslant m$. Which makes GTSP even more challenging to solve.

The GTSP is a combinatorial optimization problem widely used for its numerous applications such as sequencing computer files, postal routing, warehouse order picking with multiple stock locations, carrier-vehicle systems [4], and many others.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_14

Ben-Arieh et al. [1], Laporte and Semet [10], Noon and Bean [12] and other researchers tried to convert GTSP instances into TSP ones since there are multiple efficient heuristics for the TSP. The idea is appealing but too much limited as an exact solution from the obtained TSP instances must be obtained, otherwise, it may lead to an infeasible GTSP solution for the converted instance.

By using dynamic programming, Henry-Labordere [7] and Srivastava et al. [20] were the first researchers to propose solving the GTSP. Fischetti et al. [3] developed later a Branch and Cut solution giving optimal solutions for all instances with up to 442 cities and 89 clusters.

Several tour construction heuristics inspired from the TSP have been adapted on the GTSP, Fischetti [3] proposed insertion based algorithms with farthest, nearest and cheapest insertions. While Noon [11] presented the nearest neighbor heuristic.

Several local search heuristics were studied in [2,5,8,9,15] and some others. Many researchers worked also on Genetic Algorithms [17–19] while others made hybrid algorithms with local search heuristics and Genetic Algorithms yielding memetic algorithms [2,5,6].

We present in this paper a new idea to handle clusters of a GTSP instance. The approach consists on finding the barycenter for each cluster and to consider it as the representative or a fictitious node of the cluster. These new nodes will enable us, among others, to initialize a tour by first determining a cluster visiting order.

The paper is outlined as follows. Section 2 describes the GTSP problem and how a solution is represented. In Sect. 3, we show how barycenters can be used to find a solution, either to initialize a tour or to improve it. Section 4 gives some practical works for tour construction using barycenters. Finally, we conclude our paper in Sect. 5.

2 GTSP Definition and Solution Representation

2.1 Problem Definition

The GTSP problem can be defined as follows. We consider an undirected weighted graph $G = (V, E, c)$ such that V is the set of nodes, E the set of edges and c the cost function $E \rightarrow \mathbb{R}^+$. The set V of nodes is distributed over m clusters $C_1, C_2, ..., C_m$ as the union of all clusters gives the set V ($\bigcup_{i=1}^{m} C_i = V$) and the intersection of two clusters is an empty set ($C_i \bigcap C_j = \emptyset, i, j = 1, ..., m, i \neq j$).

A solution of the GTSP is a subgraph $S(P, T)$ of G with $P = p_1, p_2, ..., p_m$ is the set of selected nodes as p_i belongs to C_i for $1 \leqslant i \leqslant m$, and T set of edges belonging to E and forming a Hamiltonian cycle. The cost of this cycle is the total cost of edges that make it up:

$$W(T) = c(p_m, p_1) + \sum_{i=1}^{m-1} c(p_i, p_{i+1}) \tag{1}$$

2.2 Cluster-Based Approach

Pourhassan and Neumann [13] proposed two approaches for constructing a GTSP solution, one consist of getting all spanning nodes and then finding the shortest tour from it, called Node-Based (NB) approach. The other approach aims at finding a permutation of clusters then selecting one node from each one, named Cluster-Based (CB) approach.

CB approach is also used in the generalized 2-Opt heuristic given by Renaud et al. [15]. Once the visiting order of clusters $(C_{\pi_1}, C_{\pi_2}, ..., C_{\pi_m})$ is defined, we build a graph containing edges between nodes of consecutive clusters obtained before. We add a copy of the first cluster attached to the last clusters we can see in Fig. 1.

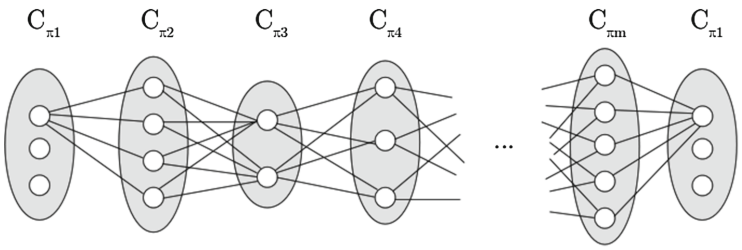


Fig. 1. Finding the shortest tour after defining the visiting order of clusters.

Instances of the GTSP are set on files in which we often find the coordinates of nodes and clusters of belonging, allowing us to get the distances between the different nodes of these instances. The idea in this work is to use these details to identify the position of each cluster by finding its corresponding barycenter, which will let us handle cluster differently.

3 Using Barycenters in the GTSP

3.1 Tour Initializing

Existing Methods. The majority of the existing methods are inspired by the greedy methods of the TSP, such as the farthest insertion or the nearest neighbor, which can be applied to all instances of the GTSP.

Adaptation is necessary to apply the nearest neighbor heuristic on the GTSP. Algorithm 1 below is a pseudo-code of this adaptation. We start by randomly selecting a node c of a cluster C chosen arbitrarily (line 1), we move iteratively to the nearest node but not belonging to any of the visited clusters (lines 5–7), the process is repeated as long as there are not yet visited cluster. Once completed, the tour is closed by returning to the first node c (line 9).

Noon [11] presented a similar approach by repeating the process described above on all nodes of the first chosen cluster (C) and selects the best of these

tours. Karapetyan and Gutin [9] show that Noon's approach does not give a great impact on the quality of the solution especially after running a local search, but only slows the heuristic.

Algorithm 1. Nearest Neighbor heuristic for the GTSP

Ensure: A tour T
1: Select randomly a cluster (C) then a city (c) from (C)
2: $c_1 \leftarrow c$
3: Add (c_1) to the tour (T)
4: **while** visited cities are less than m **do**
5: Find nearest city (c_n) to (c) from unvisited clusters
6: Add (c_n) to the tour (T)
7: $c \leftarrow c_n$
8: **end while**
9: Add (c_1) to the tour (T)
10: **return** the tour (T)

Many other adaptations have been used by Fischetti *et al.* [3], such as farthest insertion, nearest insertion, and cheapest insertion. Renaud and Boctor [15] proposed recently a heuristic called Generalized Initialization, Insertion and Improvement (GI^3), which is a generalization of the I^3 heuristic presented in [16] for the TSP.

We also find in the literature a pseudo-random construction method [9], which randomly defines a permutation of clusters and then applies the Cluster Optimization (CO) heuristics to choose the best node in each cluster.

Construction Based on Cluster Barycenters. Another way to put the clusters in order better than random is to consider their respective locations. Clusters are built in GTSP instances in order to group the nodes that are in the same neighborhood (Example on Fig. 2), so it is impossible to have an intersection between two clusters of neighborhoods, like the one shown on the Fig. 3.

Therefore, we propose to define for each cluster an imaginary point placed on its barycenter to build a permutation in a better order. This can only be done for instances that display the coordinates of the nodes in their files. The barycenter B of a cluster $C, |C| = k$ (or set of k points) is calculated as follows.

$$B(x, y) = \begin{cases} x = \sum_{i=1}^{k} (x_i)/k \\ y = \sum_{i=1}^{k} (y_i)/k \end{cases} \tag{2}$$

Once the barycenters of respective clusters are defined Fig. 4, any tour construction heuristic for the TSP may be applied, such as nearest neighbor heuristic, where a barycenter is chosen randomly and then moving iteratively to the nearest barycenter not yet visited.

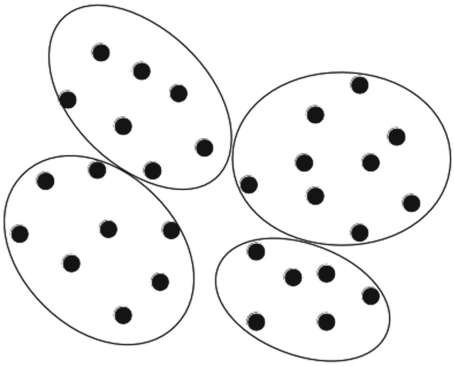


Fig. 2. GTSP instance example.

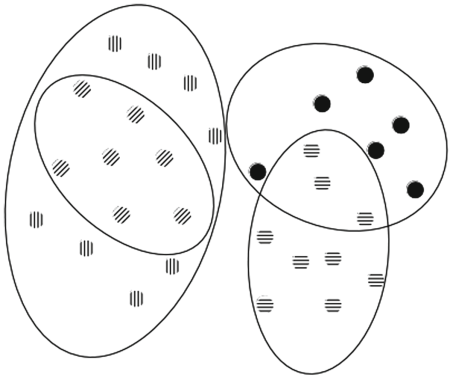


Fig. 3. Cluster intersection on a GTSP instance.

After obtaining a permutation of clusters $(C_1, C_2, ..., C_m)$, remains to choose a node from each one. The choice may be random or opting for a strategy as the nearest neighbor: We select for the first two clusters C_1 and C_2 the closest nodes to each other, then the closest node to C_2 and belonging to C_3 and so on. The selection of a new node from the next cluster is repeated until arriving at the cluster C_m. This process is detailed in Algorithm 2.

3.2 General Use of Barycenters

In general, the set of all barycenters of a GTSP instance can be considered as a fictive instance of the TSP. The Fig. 5 shows how a fictive TSP instance can be created from the GTSP instance shown in Fig. 4. When using the CB approach, we can build a permutation of clusters via an initialization heuristic as we did earlier. But also to improve an existing permutation through improvement methods for the TSP as running a local search heuristic on this permutation tour to improve its quality.

Algorithm 2. Nearest Cluster heuristic

Require: A cluster permutation $(C_1, C_2, ..., C_m)$
Ensure: A tour T
1: Find the node pair $(x, y), x \in C_1, y \in C_2$ giving the shortest distance between (C_1) and (C_2)
2: Add (x) and (y) to the tour (T)
3: $c \leftarrow y$
4: **for** i from 3 to m **do**
5: Find nearest city (c_n) to (c) from cluster (C_i)
6: Add (c_n) to the tour (T)
7: $c \leftarrow c_n$
8: **end for**
9: Add (x) to the tour (T)
10: **return** the tour (T)

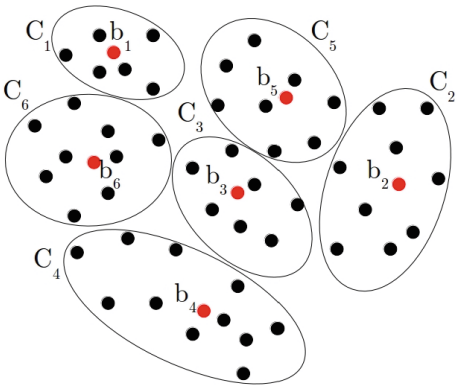


Fig. 4. Finding barycenters for each cluster.

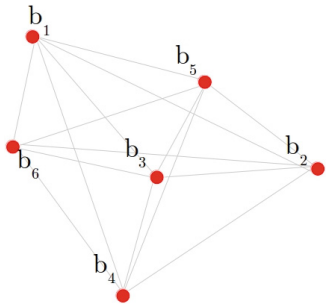


Fig. 5. Cluster barycenters making a TSP instance.

4 Experimental Results

We present in this section some experimental results of our new approach. The algorithms are implemented using Java on a machine with Pentium(R) Dual-Core CPU T4300 @ 2.10 GHz × 2 and 3.9 GB of memory. The results are the average of objective values obtained after 20 executions.

The tests relate to the use of barycenters for initializing a solution of a GTSP instance. We used the nearest neighbor heuristic on the barycenters in order to have a permutation of clusters, then used the method seen in the algorithm 2 (bar/nn). The tests were performed on 51 instances from the TSPLIB [14] converted to a GTSP instances by the standard clustering procedure [3]. In Table 1, we report some results for 20 instances in order not to clutter the paper. For each instance are given the number of *clusters* and *nodes*, the best-known solution (the *optimum*), and the average over 20 runs of both the objective

Table 1. Construction results using barycenters

Instance	Clusters	Nodes	Optimum	bar/nn	(AVG, CPU (ms))	Random	(AVG, CPU (ms))
11berlin52	11	52	4040	**7400.60**	0	8410.65	0
11eil51	11	51	174	**306.50**	0	416.5	0
20kroA100	20	100	9711	**32945.50**	0	36379.8	0
20kroB100	20	100	10328	**35523.60**	0	37536.9	0
20rd100	20	100	3650	**10516.70**	0	11525.25	0
21eil101	21	101	249	**710.20**	0	850	0
21lin105	21	105	8213	**27494.95**	0	28226.6	0
28pr136	28	136	42570	**160173.80**	0	177748.5	0
30ch150	30	150	2750	**9818**	0	11112.85	0
32u159	32	159	22664	**88754.15**	0.05	92740.35	0
39rat195	39	195	854	**4373**	0.35	4742.05	0
40kroa200	40	200	13406	**67180.6**	0.3	68981.35	0
53gil262	53	262	1013	**5265.55**	1.05	5863.2	0
56a280	56	280	1079	**6653.85**	1.15	7077.55	0
99d493	99	493	20023	**111992.2**	3.9	115085.1	0
115u574	115	574	16689	**133338.20**	4.85	139343.1	0
115rat575	115	575	2388	**23703.05**	5.05	24090.3	0
201pr1002	201	1002	114311	**1285180**	14.05	1307635	0
212u1060	212	1060	106007	**1325044**	16.95	1351463	0.15
217vm1084	217	1084	130704	**1540614**	16.45	1604300	0.25

value (AVG) and CPU time in ms for the proposed method and the random construction.

All tour constructions are done instantly (less than a tenth of a second) for all instances, 217vm1087 being the largest.

The obtained results in the previous tests are indicative and aim to validate the feasibility of the use of cluster barycenter. As the paper presents a new approach that may be applied to various uses.

5 Conclusion

Using cluster barycenters is a promising new idea which can give very good results in building a permutation of clusters. We presented in this paper a new approach for handling a GTSP instance. Whether for building a permutation of clusters in order to initiate a tour for this instance, or for improving an existing one. Considering all barycenters of each instance as a fictive instance of TSP, the cluster permutation can be enhanced by applying well-known improvement methods of this problem.

Future works will be devoted to implementing metaheuristics based on cluster barycenters' approach to present an extended use and confirm the benefits of using this method.

Acknowledgments. The authors of this paper sincerely thank the Agence Universitaire de la Francophonie (AUF) for the generous support and convey their gratitude to Mr. Mohamed El Yafrani for his precious remarks on this work.

References

1. Ben-Arieh, D., Gutin, G., Penn, M., Yeo, A., Zverovitch, A.: Transformations of generalized ATSP into ATSP. Oper. Res. Lett. **31**(5), 357–365 (2003)
2. Bontoux, B., Artigues, C., Feillet, D.: A memetic algorithm with a large neighborhood crossover operator for the generalized traveling salesman problem. Comput. Oper. Res. **37**(11), 1844–1852 (2010)
3. Fischetti, M., Salazar Gonzalez, J.J., Toth, P.: A branch-and-cut algorithm for the symmetric generalized traveling salesman problem. Oper. Res. **45**(3), 378–394 (1997)
4. Garone, E., Determe, J.F., Naldi, R.: Generalized traveling salesman problem for carrier-vehicle systems. J. Guidance Control Dyn. **37**(3), 766–774 (2014)
5. Gutin, G., Karapetyan, D.: A memetic algorithm for the generalized traveling salesman problem. Nat. Comput. **9**(1), 47–60 (2010)
6. Gutin, G., Karapetyan, D., Krasnogor, N.: Memetic algorithm for the generalized asymmetric traveling salesman problem. In: Krasnogor, N., Nicosia, G., Pavone, M., Pelta, D. (eds.) Nature Inspired Cooperative Strategies for Optimization (NICSO 2007), pp. 199–210. Springer, Heidelberg (2008)
7. Henry-Labordere, A.: The record balancing problem: a dynamic programming solution of a generalized travelling salesman problem. RIRO **B–2**, 43–49 (1969)
8. Hu, B., Raidl, G.R.: Effective neighborhood structures for the generalized traveling salesman problem. In: Hemert, J., Cotta, C. (eds.) EvoCOP 2008. LNCS, vol. 4972, pp. 36–47. Springer, Heidelberg (2008). doi:10.1007/978-3-540-78604-7_4
9. Karapetyan, D., Gutin, G.: Efficient local search algorithms for known and new neighborhoods for the generalized traveling salesman problem. Eur. J. Oper. Res. **219**(2), 234–251 (2012)
10. Laporte, G., Semet, F.: Computational evaluation of a transformation procedure for the symmetric generalized traveling salesman problem. INFOR Inf. Syst. Oper. Res. **37**(2), 114–120 (1999)
11. Noon, C.E.: The generalized traveling salesman problem. Ph.D. thesis, University of Michigan (1988)
12. Noon, C.E., Bean, J.C.: An efficient transformation of the generalized traveling salesman problem. INFOR Inf. Syst. Oper. Res. **31**(1), 39–44 (1993)
13. Pourhassan, M., Neumann, F.: On the impact of local search operators and variable neighbourhood search for the generalized travelling salesperson problem. In: Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation, pp. 465–472. ACM (2015)
14. Reinelt, G.: TSPLIB–a traveling salesman problem library. ORSA J. Comput. **3**(4), 376–384 (1991)
15. Renaud, J., Boctor, F.F.: An efficient composite heuristic for the symmetric generalized traveling salesman problem. Eur. J. Oper. Res. **108**(3), 571–584 (1998)

16. Renaud, J., Boctor, F.F., Laporte, G.: A fast composite heuristic for the symmetric traveling salesman problem. INFORMS J. Comput. **8**(2), 134–143 (1996)
17. Shi, X.H., Liang, Y.C., Lee, H.P., Lu, C., Wang, Q.: Particle swarm optimization-based algorithms for tsp and generalized tsp. Inf. Process. Lett. **103**(5), 169–176 (2007)
18. Silberholz, J., Golden, B.: The generalized traveling salesman problem: A new genetic algorithm approach. In: Baker, E.K., Joseph, A., Mehrotra, A., Trick, M.A. (eds.) Extending the Horizons: Advances In Computing, Optimization, and Decision Technologies, pp. 165–181. Springer, New York (2007)
19. Snyder, L.V., Daskin, M.S.: A random-key genetic algorithm for the generalized traveling salesman problem. Eur. J. Oper. Res. **174**(1), 38–53 (2006)
20. Srivastava, S., Kumar, S., Garg, R., Sen, P.: Generalized travelling salesman problem through n sets of nodes. CORS J. **7**, 97–101 (1969)

On Pollution Attacks in Fully Connected P2P Networks Using Trusted Peers

Cristóbal Medina-López[1(✉)], Ilshat Shakirov[2], L.G. Casado[1], and Vicente González-Ruiz[1]

[1] University of Almería (ceiA3), Almería, Spain
{cristobalmedina,leo,vruiz}@ual.es
[2] Perm State University, Perm, Russia
im.shakirov@gmail.com

Abstract. This work studies the impact of the *pollution attack* in a fully connected live video streaming P2P push based overlay network, as well as the effectiveness of using trusted peers as the only defense mechanism. Determining the best strategies for both: (i) trusted peers (TPs) with the aim to detect and expel malicious peers (MPs) and (ii) malicious peers in order to pollute without being detected, is a challenge. Preliminary experimental results for the studied strategies show that the use of TPs is not an enough solution to defeat a large number of MPs because the number of TPs should be at least equal to the number of MPs. Additionally, this study shows the ingredients of a possible game theory problem formulation where the proposed attacker strategy could be improved to detect TPs, which in turn could improve the defense strategy, and so on.

Keywords: P2PSP · Content integrity · Live streaming · Pollution attacks · Peer-to-peer networks · Network performance evaluation

1 Introduction

The scalability problem in live video streaming systems based on a client-server model is well known: the server is forced to upload a copy of the content per client. In a peer-to-peer (P2P) system, the number of copies uploaded by the server is usually orders of magnitude lower than the number of peers, which are in charge of sharing the content among them, using their upload bandwidth excess. For that reason, the P2P model is an attractive alternative to the client-server model, especially since the available upload bandwidth grows progressively at the client side.

In this context, the Peer-To-Peer Straightforward Protocol[1] (P2PSP) was developed as an application-layer protocol designed for the real-time broadcasting of media over a P2P overlay. P2PSP establishes a push-based fully-connected

[1] http://p2psp.org/en.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_15

mesh scheme where every peer sends data to each other. A basic P2PSP network (see Fig. 1) consists of a *splitter* (S) and a collection of *peers* (P_i) named *team*. The splitter receives a media stream from a *source* (O), divides the stream into chunks of data with the same size and sends them to the team following a Round-Robin scheme. Next, each peer forwards those chunks received from the splitter to the rest of peers (chunks received from other peers are not forwarded). Finally, each peer sends the reconstructed stream to a *listener* (L), which usually is a media player. A round consists in the splitter sends one chunk to every member of the team.

Because peers obtain most of the content from other peers rather than from the splitter, it is easy for a malicious peer (MP) to change the delivered content in some way, i.e. to produce a *pollution attack* [4]. Pollution attacks can be classified as *persistent attacks* [4], when the attacker always sends polluted chunks to their victims or as *on-off attacks*, when the attacker only poisons some chunks (for instance, 10% of the total number of chunks sent to their victims), in order to overcome reputation systems [6]. Free-riders or selfish peers (those that do not send data) affect in a similar way as polluters in our system and are treated in the same way.

Previous attacks can be modified by including *selective (non-selective)* and *collaborative (non-collaborative)* characteristics. In the selective attack, the chunks are poisoned for just one peer or a small subset of peers. In the collaborative attack, several attackers collaborate to perform the attack in order to increase the damage and to avoid their detection. In the presence of Trusted Peers (TPs), attackers could also collaborate to discover them.

An attacker can use a *hand-wash attack*: it leaves the team in case it thinks it was discovered and returns with another identity. This can be seen as a type of *sybil attack*, where the same attacker uses different identifications to deploy several peers in the team.[2] Additionally, an attacker can use a *bad-mouth attack* [7] under reputation systems by complaining about well-intended peers (WIPs), but we do not consider reputation systems in this study.

In this research, the defense Strategy is based on Trusted Peers and Digital Signatures (STrPe-DS) [10]. In STrPe-DS, each transmitted chunk is signed with the private key of the splitter to allow their verification by peers. In this study, peers do not handle other cryptographic material, and the content is not ciphered.

The main contributions in this study are:

- To define policies for TPs and MPs in a full connected overlay P2P network using a push model.
- To experimentally determine the necessary number of TPs to expel MPs using previous policies.
- To show future lines of study for policy improvement.

The rest of this paper is organized as follows. Related work is presented in Sect. 2. Section 3.1 shows a description of the studied strategies and the impact

[2] In the context of P2PSP, peers are identified by their IP addresses, so, attacker should use botnets to perform a sybil attack.

of different attacks. Results from the simulation are shown in Sect. 4. Finally, Sect. 5 shows the conclusions and the future work.

2 Related Work

Pollution attacks produced by MPs are some of the vulnerabilities of P2P overlays. Consequently, they have been studied extensively, specifically in the live streaming scenario [2,4,6,9]. Most of the defenses are based on trust management by gathering the reputation information from subsets of neighbors in the network. The main drawback of reputation systems are the false positives and false negatives, mainly due to bad-mouth attacks, resulting in MPs remaining in the team and WIPs being expelled.

The use of a set of TPs will reduce the number of MPs in the team. A framework with a set of TPs to monitor the bandwidth usage of untrusted peers in the system to prevent free-riding was studied in [3]. A detection algorithm for a set of TPs was considered in [5], where peers inform to the assigned TPs about the correct reception of a chunk. Then, the set of TPs can identify MPs by solving an inference problem. In P2PSP, TPs have their own information about the team because it is a complete connected overlay network. An attacker is expelled by the splitter based on the information received from TPs. Splitter expels a peer by removing it from its list of peers and therefore it will receive chunks anymore.

3 Description of the Model and Attacks

3.1 STrPe-DS Model

STrPe-DS has been designed to mitigate the selective attack and to identify poisoned chunks by using digital signatures. Any peer can know if a chunk was poisoned and/or relayed by a different peer than the one in charge of it, as shown below. The following rules define the STrPe-DS [10]:

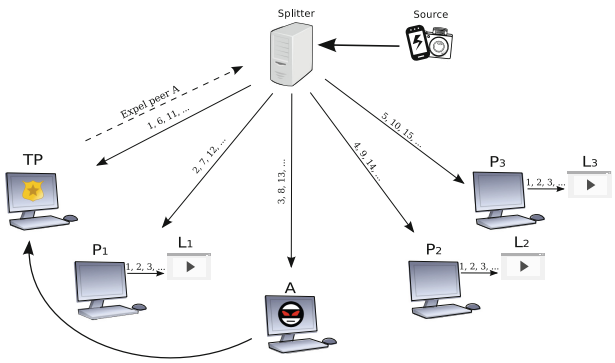


Fig. 1. A P2PSP team using STrPe-DS.

1. A peer receives the splitter's public key (S_{pub}) from the splitter, plus other necessary information, when a peer joins the team.
2. The splitter sends a different chunk and a message (m) to different peers. The message includes the chunk number (`cNumber`), the destination address (`dst`) and a digital signature $m = \{\texttt{cNumber}, \texttt{dst}, S_{priv}(H(\texttt{chunk}+\texttt{cNumber}+\texttt{dst}))\}$, where H is a hash function, and S_{priv} is the private key of the splitter.
3. When a peer receives a message from another peer, it performs the following steps:
   (a) Check whether `dst` matches the address of the sender. Notice that this action is vulnerable to the well-known *spoofing attack* [1].
   (b) If 3a was satisfied, it checks the correctness of the hash value in the message.

   The sender is removed from the list of peers of the current peer if 3a or 3b fails.
4. The splitter periodically requests, and the TPs serve, the list of removed peers since the previous request.
5. Peers removed by any TP will be expelled by the splitter after a random time, in order to disassociate the expulsion with the action taken by the attacker.

3.2 The Impact of Pollution Attacks

When an MP attacks a WIP or a TP, both know that they have been attacked (see Rule 3 on Sect. 3.1). For the sake of simplicity, the worse scenario is supposed: all MPs collaborate and attack anytime they want to, and the splitter does not use a reputation system. Some notation is defined in Table 1.

Table 1. Notation used for measuring the impact of attacks.

Symbol	Meaning
p	Probability of a TP detects an MP in a round
A_C	Total number of polluted chunks per round independently of their cNumber
C_M	Number polluted chunks with different cNumber per round and non-MPs
M	Number of MPs per round
T	Number of TPs per round
W	Number of WIPs per round
N	Number of peers in the team (T+W+M)

Non-selective Attacks. In a non-selective attack, MPs poison[3] chunks to the non-MPs. Non-selective attacks can be classified into persistent and on-off attacks. In the first case, an MP poisons all the chunks received from the splitter.

[3] Or refuse-to-send, which is equivalent for our analysis.

In the second case, MPs poison a chunk with a probability q. Because STrPe-DS is memory-less, in a persistent attack, $q = 1$, $A_C = N - M$ and $C_M = M$. For on-off attack, $p = q$, $A_C = q(N - M)M$ (each MP performs $q(N - M)$ attacks) and $C_M = qM$.

Non-selective attacks are not appealing for MPs in the presence of TPs because there exist a trade-off between the probability of being detected (qT) and the number of attacks per round (A_C).

Selective Attacks. $M = 1$ is not interesting to overthrowing the network because it can pollute a maximum of $N-1$ peers and an attacked peer will have just one polluted chunk out of N, per round. Using selective attack, trying to escape from TPs, will reduce the impact of the non-selective attack.

The probability to be discovered by a TP increases with the number of attacked peers. Therefore, a collaborative-selective attack is more interesting for attackers. Several collaborative-selective attacks can be performed:

Each one-to-one: Each MP attacks one different peer. Its impact is low because $C_M = 1$ and $A_C = M$. The only motivation of this type of attack is to scan the team looking for TPs. The expelled MP can determine who could be the TP(s).

Each one-to-many: Each MP attacks a set of n peers. The motivation is to increase the impact of each one-to-one attack. Different objectives arise:

Increase C_M: MPs attack the same set of non-MPs. So, $A_C = Mn, C_M = M$ for n peers and $C_M = 0$ for the rest. The aim of this attack is to hinder the reproduction of the stream in the set of attacked peers because degrading the streaming quality will cause that a WIP leaves the team.

Increase the scouting of TPs: MPs attack a different set of n non-MPs, minimizing the overlap among them. The altered chunks/round $A_C = Mn$. When the sets do not overlap, $C_M = 1$ for Mn peers. The aim is to narrow the TPs location when $M << N$. Otherwise, *each one-to-one* attack is preferable.

Many-to-one: All MPs attack just one peer. In this case, $A_C = M$ and $C_M = M$ for the attacked peer and $C_M = 0$ for the rest. Its aim is to increase the probability that the attacked peer leaves the team due to the low quality of the stream. MPs may reduce their exposure if a new prey is not selected while the current one is on the team.

Many-to-many: Each MP performs a *one-to-many* attack, but MPs collaborate to determine the targets.

MPs have a difficult multi-objective optimization problem: to (i) maximize the number of detected TPs, (ii) minimize the number of MPs (cost), (iii) minimize the number of expelled MPs, (iv) maximize C_M in WIPs and (v) perform previous tasks in a minimum number of rounds.

TPs have also a multi-objective optimization problem: to (i) minimize the number of MPs (this minimize C_M in WIPs), (ii) minimize the number of TPs (cost) and (iii) performs previous tasks in a minimum number of rounds.

The difficulty of both problems increases with the existence of churn. Several solutions can be devised, and each one will present a possible countermeasure. Following, we show the strategies studies here as possible non-optimal solutions.

Discovering and Hiding TPs. MPs must devise an algorithm to detect TPs. In a simple approach, each MP increases the set of attacked peers by just one per round. MPs collaborate to attack a peer that has not been attacked before. Under the assumption that a TP informs to the splitter as soon as it is attacked, the MPs team can collaborate to make out who is a TP when an MP is expelled.

For $T = 1$, the TP can be discovered in $\lceil \frac{N}{M} \rceil$ rounds. The expelled MP will inform to others MPs in the team and MPs will attack the discovered TP anymore. Then, the impact of the attack would be $p = 0$, $A_C = N - T - M$ (all WIPs), $C_M = M$ in the worst case.

To overcome the previous algorithm, the splitter will expel an MP in one of the next rounds (at random). Moreover, a TP should leave the team in next rounds (at random) when it informs to the splitter about an MP, and another TP with different identity should be included in the team. In this way, MPs have more difficulties to figure out who a TP is.

4 Experimentation

The experimentation is performed to measure the number of affected peers during the attack, the number of affected chunks (per peer and round), the average expulsion time for MPs, the average number of WIPs and MPs and the state of the buffer of WIPs. Time is expressed in rounds.

4.1 Test Bed

For all the experiments we used a simulator specifically coded for evaluating different aspects of the P2PSP protocol. Both, the protocol and the simulator are open source and they are available at [12,13]. They were run on a Unix machine (Ubuntu server) with 2.3 TB of RAM and 8 Intel Xeon E7 8860v3 (16 cores) processors.

Churn:

1. The initial team consists of a splitter and an initial number of WIPs (#IWIPs).
2. To avoid any tracking of TPs, they will be incorporated into the team at random, mixed with the arrival of the rest of peers (MPs and WIPs) up to reach a $\#\text{TP}_{max}$.
3. As $\#\text{TP}_{max}$, the values of the maximum number of WIPs ($\#\text{WIP}_{max}$) and MPs ($\#\text{MP}_{max}$) limit the number of entries of WIPs and MPs in the team along the simulation.
4. A new peer (MP or WIP) arrives with probability P_{in}. The peer will be WIP with probability P_{WIP} and MP with probability P_{MP}. Additionally, when a new TP is needed, it joins the team also with probability P_{in}.

5. Peers (all types) stay in the team a period of time modeled by a Weibull distribution with the shape parameter $a = 1$ [11].
6. Events in peers are triggered when a new chunk arrives at them. A peer that spends its Weibull time will leave the team with probability P_{out} in next event.

MPs Behavior:

1. MPs always attack (persistent, selective and collaborative attack).
2. An MP should not attack over 50% of the team to reduce the probability of being detected.
3. MPs should scout for TPs. We define NAY as the set of peers Not Attacked Yet. Only one MP attacks to a NAY peer. NAYs peers that were not attacked by expelled MPs are attacked before others.
4. Additionally, MPs should attack peers as much as possible. Then, every MP also attacks peers that have been attacked more than MPTR (Malicious Peer Test Round) by other MPs.
5. MPs share the list of attacked peers and how many rounds an MP was attacking each of them. Thus, they know the NAY and peers attacked more than MPTR rounds.

WIP Behavior Under Attacks:

1. They will leave the team if the Exponential Smoothing of Chunks Loss Rate (ESCLR) is greater than ESCLR_{max}. ESCLR defined as:

$$\text{ESCLR}_i = \alpha \text{CLR}_i + (1 - \alpha)\text{ESCLR}_{i-1} \tag{1}$$

where CLR_i is the chunks loss rate measurement in last round.
CLR is defined as:

$$\text{CLR} = \frac{\text{polluted/lost chunks in B}}{|\text{B}|} \tag{2}$$

where B is the buffer of the peer and |B| is the size of B.

TPs Behavior Under Attacks:

1. TPs inform to the splitter about attacks as soon as they are attacked.
2. TPs, as WIPs do, will leave the team if the ESCRL>ESCLR_{max}.
3. A TP will leave the team after it complains about an MP with probability P_{TPL} in order to avoid its detection by MPs.

Splitter Behavior Against Attacks:

1. A reported MP will be expelled in current round with probability P_{MPL}.

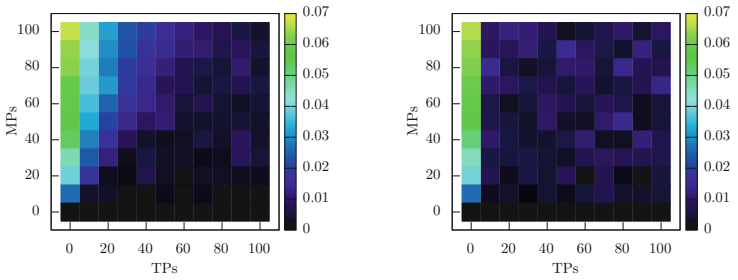
4.2 Experimental Results

The main aim of this study is to know if it is feasible to mitigate pollution attacks with previous strategies. We focus on the defense side and the tune of P_{MPL} and P_{TPL} values. For all experiments P_{in}=1, P_{out}=0.25, P_{MP}=P_{WIP}=1, MPTR=2 and ESCLR_{max}=1 (peers stay on the team even when the stream has a low quality). $\#\text{TP}_{max}$ and $\#\text{MP}_{max}$ values were taken from the 100 combinations of values in {0, 10, 20, ..., 90, 100}. The simulation stops 100 rounds

after the requested number of peers (#WIP$_{max}$+#TP$_{max}$+#MP$_{max}$) got in the team. Additionally, P_{MPL} and P_{TPL} values used in each of the previous combination are shown in Table 2. Therefore, the number of experiments is 1100. Five executions have been carried out for each experiment, and after discarding the outliers, the average of the remaining three values was returned.

Table 2. Parameter values for the different experiments.

Exp	1	2	3	4	5	6	7	8	9	10	11
P_{MPL}	100	100	100	100	100	100	80	60	40	20	0
P_{TPL}	100	80	60	40	20	0	100	100	100	100	100



(a) Exp 1. (b) Exp 6.

Fig. 2. Numerical results. Each small square is a combination of #TP$_{max}$ and MP$_{max}$ values. Right column shows ESCLR values. The darker the better.

All the instances in Table 2, except Exp. 6, show similar results to Exp. 1 instance, shown in Fig. 2a. Therefore, variation in P_{MPL} and P_{TPL} has not a great impact on the results. In general, when #TPs≥#MPs the ESCLR is low.

Figure 2b shows the results for Exp. 6 with P_{TPL}=0. Now, TPs do not leave the team during the entire simulation. The value of MPTR=2 and MPs sharing their "safe to attack" list including TPs produce that all MPs attack TPs and consequently they are expelled. A solution for MPs is to increase the number of attacked peers just by one in each round and a different one by each MP. In this way, the number of MPs attacking a TPs due to the "safe to attack" report of other MPs is much smaller than the entire MP set.

Figure 3 depicts the state of the team and the chunk-loss ratio for the Exp. 1 and two different combinations of #TP$_{max}$ and #MP$_{max}$. In order to have a stable state in the team before the attack, the initial rounds, where only WIPs are allowed, are not plotted). These results show that TPs can expel almost all MPs when their number is similar (all MPs if TPs are a majority), but in general TPs cannot do it when they are a minority. In such case, the remaining number of MPs is approximately #MP$_{max}$ − #TP_{max}. Therefore, other MPs' control

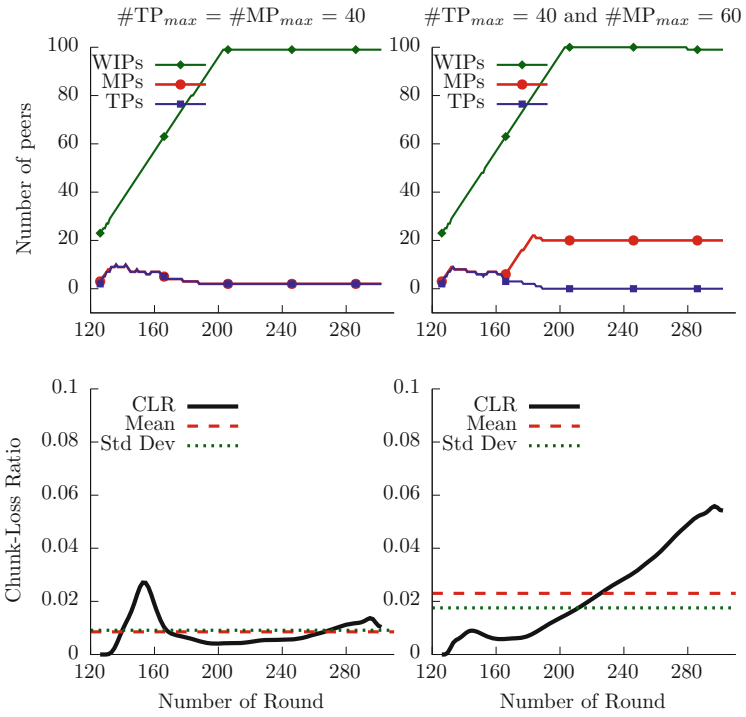


Fig. 3. Examples of changes in the composition of the team for Exp. 1 (up) and their consequences in the chunk-loss (down).

mechanisms, such as reputation systems, well behavior rewards, etc. should be added to the use of TPs. Notice that most of these systems have been studied for pull based P2P systems, but there exist few studies for push based systems, like the P2PSP [8].

5 Conclusions

Attackers and defenders have to solve a multi-objective optimization problem and each decision made by one part may have a countermeasure by the other part. A possible collaborative-selective algorithm performed by attackers is devised with two main objectives: (i) to discover trusted peers and (ii) to pollute as much as possible the buffer of well-intended peers without been expelled. At the defense side, the proposed algorithm tries to hide the trusted condition of a peer to attackers. Both algorithms are interesting because the following question could arise: Who are the good/bad guys? The answer depends on the actual scenario. In a normal one, the stream belongs to defenders. But good guys are the attackers when bad guys perform an illegal public free rebroadcast in a P2P network of their private stream, i.e., the stream belongs to attackers.

The experimental results show that using only trusted peers as a defense strategy is not appealing when the number of malicious peers can be large. Therefore, the improvement of the policies and additional techniques should be addressed. It will be done as a future work.

Acknowledgments. This paper has been supported by the Spanish Ministry (TIN2015-66680), in part financed by the European Regional Development Fund (ERDF). Cristóbal is supported by an FPU Fellowship (FPU14/00635) from the Spanish Ministry of Education (MECD). The work done by Ilshat was funded by Google Summer of Code (GSoC) 2015 initiative.

References

1. Bellovin, S.M.: Security problems in the TCP/IP protocol suite. SIGCOMM Comput. Commun. Rev. **19**(2), 32–48 (1989)
2. Borges, A., Almeida, J., Campos, S.: Fighting pollution in P2P live streaming systems. In: 2008 IEEE International Conference on Multimedia and Expo, pp. 481–484. IEEE (2008)
3. Conner, W., Nahrstedt, K.: Securing peer-to-peer media streaming systems from selfish and malicious behavior. In: Proceedings of the 4th on Middleware Doctoral Symposium, MDS 2007, pp. 13:1–13:6. ACM, New York (2007)
4. Dhungel, P., Hei, X., Ross, K.W., Saxena, N.: The pollution attack in P2P live video streaming: measurement results and defenses. In: Proceedings of the 2007 Workshop on Peer-to-Peer Streaming and IP-TV, pp. 323–328. ACM (2007)
5. Gaeta, R., Grangetto, M.: Identification of malicious nodes in peer-to-peer streaming: A belief propagation-based technique. IEEE Trans. Parallel Distrib. Syst. **24**(10), 1994–2003 (2013)
6. Hu, B., Zhao, H.V.: Pollution-resistant peer-to-peer live streaming using trust management. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 3057–3060, November 2009
7. Kang, X., Yongdong, W.: A trust-based pollution attack prevention scheme in peer-to-peer streaming networks. Comput. Netw. **72**, 62–73 (2014)
8. Lin, E., de Castro, D.M.N., Wang, M., Aycock, J.: SPoIm: A close look at pollution attacks in P2P live streaming. In: 2010 18th International Workshop on Quality of Service (IWQoS), pp. 1–9 (2010)
9. Marti, S., Garcia-Molina, H.: Identity crisis: anonymity vs reputation in P2P systems. In: 2003 Proceedings of the Third International Conference on Peer-to-Peer Computing (P2P 2003), pp. 134–141. IEEE (2003)
10. Medina-López, C., Casado, L.G., González-Ruiz, V.: Pollution attacks detection in the P2PSP live streaming system. In: Herrero, Á., Baruque, B., Sedano, J., Quintian, H., Corchado, E. (eds.) International Joint Conference. AISC, pp. 401–410. Springer, Heidelberg (2015)
11. Stutzbach, D., Rejaie, R.: Understanding churn in peer-to-peer networks. In: Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement, IMC 2006, pp. 189–202. ACM, New York (2006)
12. The P2PSP Team: P2PSP war-games repository. https://github.com/P2PSP/war-games
13. The P2PSP Team: Peer to Peer Straightforward Protocol Source Code. http://code.p2psp.org

Certification Under Uncertainties of Control Methods for Multisource Elevators

Chloé Desdouits[1,2,3(✉)], Mazen Alamir[1], Rodolphe Giroudeau[2], and Claude Le Pape[3]

[1] GIPSA-Lab, CNRS, Université de Grenoble-Alpes, Domaine Universitaire, 38000 Grenoble, France
{chloe.desdouits,mazen.alamir}@grenoble-inp.fr
[2] LIRMM, UMR 5506 - Université Montpellier 2, CC477, 161 rue Ada, 34095 Montpellier Cedex 5, France
rodolphe.giroudeau@lirmm.fr
[3] Schneider Electric Industries SAS, 38 TEC - Technopole, 37 quai Paul-Louis Merlin, 38050 Grenoble Cedex 09, France
claude.lepape@schneider-electric.com, chloe@desdouits.fr

Abstract. As the interest in regulating energy usage and participating in the demand-response market is growing, new energy management algorithms emerge. Energy consumption of elevators represents about 5% of the overall consumption of a building. Thus, improving energy performance of elevators becomes a relevant challenge. In this paper, a method is described to assist potential customers in choosing the most relevant storage units and control method for their multisource elevator. In order to obtain a certification on electricity cost savings and maximum power peaks, a randomized algorithm is designed. The best storage units design is found for a given use case, and a guarantee (with a fixed probability) is given that the elevator's consumption from the grid will not exceed a given power peak. Moreover, a similar guarantee is given on the minimum savings that could be achieved, choosing a given controller.

Keywords: Energy optimization · Robustness · Randomized algorithm

1 Introduction

Global energy consumption regulation is a major issue nowadays. The will to participate in the emerging demand-respond market encourages industries and building managers to look for energy management algorithms. Coupled with relevant control algorithms, the design of energy efficient systems is investigated.

One of the core questions we often encounter in the optimization of design, planning, and control, is the robustness of the proposed solution to uncertainty in the measured (or predicted) data. This includes robustness of the optimization

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_16

method to unusual data sets, even when the data are available or more-or-less accurately forecasted at the time at when the optimization method is applied. The robustness question concerns either the optimization criterion itself (i.e., what is the probability that the proposed solution effectively brings improvements over the current practice?) or the satisfaction of some constraints (i.e., what is the probability that the execution of the proposed solution becomes impossible after a given amount of time?). Let us note that, in some cases, an optimization criterion can be replaced by a more or less "soft" constraint which could be violated, but "not too often". For example, a contract may allow peak electrical power over a given day P_{day} to exceed a given limit P_{max}, but not more than a given number of days during the year; otherwise a significant penalty is incurred.

There are an infinite number of daily scenarios of elevator usage; thus we cannot simulate them all. A particular method, that can be used to offer guarantees, relies on the fact that an event (such as exceeding P_{max} on a given day) which is not observed over a representative set of N samples (with a large N) has low probability of occurrence (depending on N). In this paper, we investigate the use of such a method for the design and control of a multisource elevator.

Two use cases are considered. The first use case concerns peak power consumption: what kind of system design allows to satisfy the power peak constraint (with a given probability) and what is the daily extra cost induced? The second use case concerns the choice of relevant controller and electricity tariff in order to enable savings? These two use cases are managed as extensions of our previous work on the multisource elevator control problem [5].

2 The Multisource Elevator Problem

We call "*prosumers*", entities that consume and/or produce energy. An energy hub allows each prosumer to consume power produced by all other prosumers at the same time.

Definition 1. *Let* $\mathcal{P} = \{\pi_1, \ldots, \pi_6\}$ *be our set of* $n_p = 6$ *prosumers, all connected to the same energy hub* h.

The set $\mathcal{P}$ of prosumers is composed of: an elevator π_1, that can get energy from a battery π_2, a supercapacitor π_3, the grid π_4 and a solar panel π_5. The supercapacitor is here to absorb power peaks above the maximum power capability of the battery. But the former is more expensive than the latter, while usage age them both. Moreover, energy can be recovered from the elevator when the brakes are applied. Finally, energy can be dissipated in a resistor π_6 if there is too much.

A system composed of an energy hub h and its prosumers can be represented by a star oriented graph rooted in h. There is an arc (π_i, h) if Prosumer π_i can produce energy and the weight of the arc is the maximum power production. In the same way, there is an arc (h, π_i) if Prosumer π_i can consume energy and the weight of the arc is maximum power consumption.

Definition 2. *Let $\mathcal{G} = (\mathcal{P} \cup h, \mathcal{A})$ be the star oriented graph associated to the multisource elevator.*

We suppose that time can be sampled in a regular, uniform way.

Definition 3. *Let $\tau \in \mathbb{R}$ be the sampling period (expressed in hours), and $H \in \mathbb{N}$ be the number of periods considered in predictions and planning. Then time-steps are expressed in the following way: $t_l = t_{l-1} + \tau = l \times \tau, \forall l \in \{1, \ldots, H\}$.*

Finally, an electricity cost function $\text{cost}_{\pi_4} : [1, \ldots, H] \Rightarrow \mathbb{R}$ gives the price associated to purchasing electricity from (or reselling electricity to) the grid.

Then, we can define the multisource elevator problem:

Instance: a set of prosumers $\mathcal{P}$, a graph $\mathcal{G} = (\mathcal{P} \cup h, \mathcal{A})$, a period $\tau \in \mathbb{R}$, a time horizon $H \in \mathbb{N}$, An energy cost function cost_{π_4}
Solution: S, a $n_p \times H$ matrix of power values
Question: given data $p_{\mathbb{P}}^{max} \in \mathbb{R}$ and $cost_{hub}^{max} \in \mathbb{R}$, Can power peaks from and to the grid and the energy bill remain bounded, respectively by $p_{\mathbb{P}}^{max}$ and by $cost_{hub}^{max}$?

3 State of the Art

In order to regulate the energy consumption of an elevator, the energy consumption itself can be decreased, or energy can be recovered and stored into storage units to be reused later. In [7], smart ways to choose elevator physical components (motor, drive, etc.) are summarized, as well as appropriate sizing methods.

Coupling elevators with supercapacitors has been studied by many researchers and companies. Some of them investigate how to commute softly between the grid and a supercapacitor, like in [6,9]. In [6], a physical multisource system is designed to power an elevator. Rules are used to charge or discharge batteries depending on whether the electrical current is below or above a given reference. Likewise, [9] presents three rule-based methods to control a battery coupled with an elevator. This method takes into account peak/off-peak tariffs and reduces energy consumption cost by storing energy recovered from the elevator. These methods allow to control very reactively the system, but cannot take into account optimally external considerations such as the electricity tariff or battery state of health. Therefore, these control methods may not be efficient regarding economic objectives.

Works have also been conducted on how to take into account future energy consumption to minimize the energy bill, using storage units. In [2], a General Energy and Statistical Description (GESD) of the possible missions of an elevator is proposed, as well as a dynamic programming-based energy manager. The energy manager is inspired by stock management theory and minimizes the sum of energy (i) absorbed from the grid, (ii) dissipated in the braking resistor and (iii) not provided to the elevator. The optimization is done off-line.

This method allows to find an optimal solution regarding economical objectives from probabilities of consumption. But elevator usage is unpredictable by nature and what is to be done when the strategy is unfeasible is not investigated.

This work follows up our previous development on the subject where several energy sources are supposed to be available to be used by the elevator and the problem is to choose between them over time, as stated in Sect. 2. In [3], the real-life deployed application was described, including communication process, and web interface. In [4], two coupled controllers were proposed to solve the problem and the interactions between them are studied. The idea was to use a Strategic Optimizer (SO) to solve a linear program and get a sourcing strategy over a day. This sourcing strategy is then sent to a Local Controller (LC) that controls in real-time the energy hub, making a trade-off between the current situation and the strategic instructions. In [5], the sourcing problem and the multisource elevator problem were formalized. Several local controllers were proposed and deterministic experiments were conducted to give a first evaluation of the method potential savings.

The current paper extends the previous work on SO/LC based on a solution introduced in [4] by explicitly handling the effect of the uncertainty on the outcome of the closed-loop operation. A Certification Framework is used to guarantee customer savings on the energy bill and a limit on the power peak consumed from the grid.

4 Design of a Robust Solution

In this section, the goal is to design robust solutions regarding uncertainties. As we want to provide a guarantee that a certain condition will be satisfied at least a great percentage of the days, the uncertainties to be considered concern all elements that vary from a day to another. Let us note that these elements may be more or less known or predictable at time t_0: the state of charge of the battery is known at t_0; a weather forecast enables to estimate solar production precisely enough for our needs; and a forecast of the elevator's energy needs can also be available but often much less precise. The control method takes such forecasts into account. To certify a property on a given percentage of the days, we have to consider these forecasts as part of the uncertainty set, even though the controller take them as data.

For that purpose, we draw samples of elevator calls according to a statistical model. This model distinguishes multiple types of travels: morning and afternoon arrivals and departures, lunch breaks, inter-floor travels, arrivals and departures of external visitors. Statistical laws are identified based on historical data. For each travel type and relevant pair of floors, these laws provide information on the number of people moving during the day, the distribution of their weights, the distribution of the events, etc. The attendance of the building is tuned depending on the day-of-week and week-of-year. A random generator is used on this basis to generate scenarios. A forecast of elevator energy consumption is built by averaging energy production or consumption on several daily scenarios of the same kind. Thus, there is one forecast per kind of day.

Definition 4. *A drawn scenario of daily elevator consumption, for the kind of day D, is denoted* $w^D_{cons} \in W^D_{cons}$. *The associated forecast is denoted* $\mathbf{f}^D_{cons}$.

The second kind of uncertainties is solar radiation, uniformly drawn between two bounds: a typical cloudy day and a typical sunny day. The predicted solar production, known at the beginning of each day, is drawn between the daily profile minus ten percent and the daily profile plus ten percent. Note that solar production samples are uniformly drawn, regardless of the kind of day.

Definition 5. *A drawn scenario of daily solar production, is denoted* $w_{prod} \in W_{prod}$. *The associated forecast is denoted* $\mathbf{f}_{prod}$.

Thus, let $\mathbf{f}$ be the forecast that feeds SO. It is deterministic and depends only on the kind of day considered.

Definition 6. $\mathbf{f} = [\mathbf{f}^D_{cons}, \mathbf{f}_{prod}]$, *a* $2 \times H$ *matrix, with* H *the number of periods considered for the predictions.*

The last kind of uncertainty is the initial state of charge of storage units at the beginning of a simulated day. These states of charge are uniformly drawn between 0% and 100%.

Definition 7. *Drawn initial state of charge of storage units are denoted* $w_{initSOC} \in [0, 100]^2$.

Finally the uncertainty set considered contains the three sources of uncertainties stated before, as well as the forecasts (even if they depend only on the kind of day).

Definition 8. *Let* $W = \bigcup_D (W^D_{cons}) \times W_{prod} \times [0, 100]^2$ *be the uncertainty set considered and* $w^D \in W$ *be a scenario drawn for a given day D such that:* $w^D = (w^D_{cons}, w_{prod}, w_{initSOC}, \mathbf{f})$.

Now, let us consider two different use cases where a certification has to be given on the control method achievements. In the first one, a customer wants to respect a limit of power consumption enforced by law or by his electricity tariff (that can be much more expensive above a given power limit). In this use case, the aim is to determine the optimal sizing of a battery and a supercapacitor in order to certify that the customer will never exceed the given power limit. In the second use case, a customer already has storage units and wants to use them to participate in the demand-response market. We want to certify that, if he buys our control solution, he will gain X% on his daily electricity bill with a prescribed high probability.

In order to solve these optimization problems, a randomized algorithm is used. The principles of the chosen method are the following. Some key design parameters, relevant for the given problem, are chosen. Then, for all possible values of the design parameters, control algorithms are simulated with several uncertainty scenarios. The design vector, that gives admissible solutions and the best objective value on simulated scenarios, is kept.

Because there is an infinite number of possible uncertainty scenarios, we cannot simulate them all. Thus, in order to get a certification that simulation results are representative, a statistical model of the elevator travels (based on Gaussian laws) has been developed and used in a randomized algorithm.

To obtain this certification, we use results on randomized algorithms shown in [1]. We briefly recall them, and apply them to our context below. Let:

- Θ be the set of possible values of design parameters we want to optimize (i.e.: storage unit possible sizes, different electricity tariffs, etc.). The cardinality of Θ is denoted n_C.
- $\theta \in \Theta$ be a vector with n_θ components.
- W be the set of uncertainties (Definition 8), composed of possible elevator consumption curves during a day, solar production curves, and initial state of charge.
- $w^D \in W$ an uncertainty scenario drawn for a given day D, as stated in Definition 8.
- $u^*(\theta, w^D)$ be a vector of target state of charge for the battery and target energy purchase from the grid over the day, given by SO.
- $J(\theta)$ be an objective function to minimize.
- C be a set of constraints on design parameters.
- $g : \Theta \times W \rightarrow \{0, 1\}$ be a function that returns zero if all constraints $c \in C$ are satisfied and one otherwise.

Then, we use the following result: if we generate N i.i.d. samples $\{w^{(1)}, \ldots, w^{(N)}\}$ from W according to the probability Pr_W and then solve the following sampled optimization problem:

$$\min_{\theta \in \Theta} J(\theta) \text{ subject to } \sum_{l=1}^{N} g(\theta, w^{(l)}) \leq 0 \tag{1}$$

with

$$N \geq \frac{1}{\eta}\left(\frac{e}{e-1}\right)\left(ln\frac{n_C}{\delta}\right) \tag{2}$$

then, the probability for the optimal solution $\hat{\theta}_N$ to the optimization problem, to violate the design constraint is smaller than η, with a confidence probability no smaller than $1-\delta$. In other terms, we can certify, with a probability $1-\delta$, that the control strategy will satisfy the design constraint in at least $100 \times (1-\eta)$ percent of the cases. In order to solve this optimization problem, we use Algorithm 1.

5 Experiments

These experiments were conducted in Matlab [8]. The randomized algorithm is applied to the use cases presented above and the results are discussed.

A solution value is evaluated according to the daily cost c^{daily} Key Performance Indicator (KPI). Let: c^{ebill} be the energy bill of the whole day and c^{aging} be the aging cost associated to the storage units usage that has been done during

Algorithm 1. Randomized algorithm inspired of [1]

$S \leftarrow \emptyset$ ▷ N-set i.i.d. samples to be drawn from W
for each kind of day D **do**
  let n^D be the number of D-days in year
  draw a set S^D of $\lceil \frac{n^D \times N}{365} \rceil$ samples of D-days
  $S \leftarrow S \cup S^D$
end for ▷ at this point $S = \{w^{(1)}, \ldots\}$ is a set of at least N samples, representative of every kind of day
for each $\theta \in \Theta$ **do**
  for each $w^{(l)}$ **do**
    compute the optimal sourcing strategy $u^*(\theta, w^{(l)})$
    simulate a day $w^{(l)}$ using $u^*(\theta, w^{(l)})$
    check if the constraints C are respected: $g(\theta, w^{(l)})$
  end for
  if $\sum_{l=1}^{N} g(\theta, w^{(l)}) = 0$ **then** ▷ θ is feasible
    compute $J(\theta)$ ▷ the objective function value for θ
  end if
end for
keep the feasible θ for which $J(\theta)$ is minimal

the day. But depending on the controller, final states of charge can be different. Then we note c^{soc} the cost associated to refill (or to empty) storage units to match their initial state of charge. We consider that corresponding energy is purchased (or sold) at the cheapest tariff.

Definition 9. *The daily cost is defined as* $c^{daily} = c^{ebill} + c^{aging} + c^{soc}$.

5.1 Which Storage Units to Avoid Power Peaks?

In this sub-section, suppose that the objective is to sell our control solution to a customer admitting a very big annual penalty if he purchases electricity from the grid over a given power value p_{max} more than a given number of days during the year. Thus, we want to ensure (with a confidence of $1 - \delta$) that there is probability lesser than η to exceed the power value p_{max}. Among solutions satisfying this condition, the most profitable solution should be chosen.

The theory presented in Sect. 4 is used to decide which storage units the customer should buy to be certified to achieve his goal. In this experiment, we consider that the customer subscribes to a peak/off-peak tariff and uses our LC that reduces power peaks coupled with our SO. Now let us define:

- a function $f(\theta, w, u^*)$ that gives the daily cost c^{daily} obtained by the local controller at the end of the day, depending on: θ the considered design parameter, w the daily scenario, and u^* the strategy pre-computed by SO.
- a set $\mathcal{B} = \{3000, 6000\}$ of possible battery energy capacities (in Wh).
- a set $\mathcal{S} = \{60, 120, 180\}$ of possible supercapacitor energy capacities (in Wh).

- a set $\mathcal{F}$ of maximal daily costs certified to the customer (in €).
- the resulting set of design parameters $\Theta = \mathcal{B} \times \mathcal{S} \times \mathcal{F}$, and $\theta \in \Theta$ a vector with $n_\theta = 3$ components; the first component is the battery capacity, the second is the supercapacitor capacity, and the third is the corresponding certified daily cost.
- the design constraint c_1, that states the maximum power peak from (or to) the grid should be below p_{max},

$$c_1 : \max_{t \in \{0,\ldots,H\}} (|p_4(t)|) \leq p_{max}$$

- the design constraint c_2, that states the net daily gain should be above the third component of the design parameter:

$$c_2 : f(\theta, w, u^*) \leq \theta_3$$

- the resulting set of design constraints $C = \{c_1, c_2\}$
- $J(\theta) = \theta_3$, a maximum daily cost allowed
- the maximum power peak allowed is $p_{max} = 6000$ W
- representativeness and confidence probabilities are $\eta = 0.05$, $\delta = 0.05$, and the corresponding number of drawn samples is $N = \lceil \frac{1}{\eta}(\frac{e}{e-1})(ln\frac{n_C}{\delta}) \rceil = 152$

These parameters given, the optimization problem (1) is solved. Let us note that, in practice, the third component of the design parameters does not have to be introduced. Indeed, θ_3 can be chosen as $\max_{w^{(l)} \in S} f(\theta, w^{(l)}, u^*)$.

In the context studied (chosen elevator, storage units yield, etc.), the following report could be done to the customer.

"The smallest supercapacitor (60 Wh) does not permit to avoid all consumption peaks above 6 kW. Thus, the best choice is the 3 kWh battery and the 120 Wh supercapacitor, because they have the smallest investment cost, avoid peaks, and have the same daily cost as every other case: 0.49 €. For the sake of the comparison, the mean daily cost of a business as usual controller in the same case is 0.33 €. Thus, we can certify (with probabilities given above) that avoiding peaks above 6 kW will cost at most 16 euro cents more per day than with a classical controller."

Note that: (1) the mean daily cost computed on the 152 samples is 0.32 €, far better than the certified value (money is saved actually), (2) these results were obtained without considering reselling energy to the grid, reselling will be considered in the next subsection, (3) as the tariff gap grows, the daily cost induced by the MinPeaks LC decreases.

5.2 Which Tactic and Tariff to Get Savings?

Now, let us consider a case where a customer already has storage units (required for safety reasons) and wants to use them to improve his energy bill. Several control solutions, and electricity tariffs, must be compared to tell this customer which ones are the best for him. The customer wants to show that over period of a given length, gains have been made. For that purpose, let us define:

- a set $\mathcal{T}$ of possible electricity tariffs: $\mathcal{T}$ = {flat (0.00013), peak/off-peak (0.00015/0.00010), indexed on spot with the same coefficient for both purchase and sale of energy (between 0.0002991 and 0.0009386)} (€/Wh). These tariffs are representative of those can be found in France. The flat tariff is cheaper than the others but does not allow a customer to take advantage of his storage units by shifting consumption or inject into the grid. The peak/off-peak tariff is cheaper than the flat one during off-peak hours (18:00 - 9:00) and more expensive otherwise. The indexed on spot tariff is more expensive than the others but varies significantly every hour and remunerates the injection of energy at a high value.
- a set $\mathcal{C}$ of possible controllers (introduced in [5]): $\mathcal{C}$ = {MinPeaks LC that reduces power peaks following strategic instructions from SO, Opportunistic LC that minimizes dissipated energy, Secure LC that does not use storage units at all}. The MinPeaks LC is coupled with the strategic optimizer to take advantage of the storage units to smooth power peaks. The other controllers are on their own to reduce the installation cost of the solution, while achieving savings, or reducing costs.
- a set $\mathcal{M}$ of possible maximum power values from and to the grid: $\mathcal{M} = \{[-50000, 0], [-50000, 20000]\}$ (W). The first possibility deny re-selling electricity to the grid, while the second one allows re-selling.
- a set $\mathcal{F}$ of maximal daily costs that could be certified to the customer (in €)
- the resulting set of design parameters $\Theta = \mathcal{T} \times \mathcal{C} \times \mathcal{M} \times \mathcal{F}$, and $\theta \in \Theta$ a vector with $n_\theta = 4$ components
- the design constraint c', that states the daily cost should stay below the 4th component of the design parameter:

$$c' : \ f(\theta, w, u^*) \leq \theta_4$$

- the resulting set of design constraints $C' = \{c'\}$
- $J(\theta) = \theta_4$, a maximum daily cost allowed
- Two probabilities $\eta = 0.05$, $\delta = 0.05$, and $N = \lceil \frac{1}{\eta}(\frac{e}{e-1})(ln\frac{n_C}{\delta}) \rceil = 187$.

In this context, we would give the following report to the customer.

"A business as usual controller that does not use storage units achieves a daily cost of 0.58 €. with the flat tariff; while a daily cost of 0.25 € is achieved when re-selling is allowed, with the peak/off-peak tariff.

If re-selling is not allowed, the Opportunistic LC coupled with the peak/off-peak tariff should be chosen. The associated certified cost is 0.27 € per day while the average cost obtained is −0.03 € (which is actually a gain).

If re-selling energy to the grid is allowed, the indexed on spot tariff should be chosen, along with the MinPeak LC coupled with SO. The corresponding certified daily cost is −0.09 € per day. That means that the elevator operational cost and the storage units investment cost are compensated by the gain on electricity bill. Moreover the mean daily cost obtained on those samples is −1.24 €."

6 Conclusion

In this paper, a way to certify, under uncertainties, savings achieved by a control method for a multisource elevator, is proposed. This Certification Framework is based on a randomized algorithm and its associated bound on the number of needed drawn samples, proved in [1].

Using this algorithm, a certified bound on the maximum power peak purchased from the grid by a multisource elevator can be given, as well as a lower bound on the savings achieved. Such a solution could be used by a building manager to participate in the demand-response market. In the same way, a customer can be told the best control solution he should buy for his multisource elevator, and be certified of the minimum economical gain induced.

The experiments show that power peaks can be avoided and a multisource elevator can be made energy cost-free, though this is strongly dependent on the context. The key point of the method presented is the possibility to find the best design in a given context, and the ability to certify the associated savings. This allows a manager to arbitrate if the chosen design worth to be installed or not.

Future work will be focused on the influence of the kind of building and elevator on the potential savings. A real-life experiment would have to be conducted to enforce computation results.

Acknowledgment. This work has been conducted as part of the Arrowhead European project and has been partially funded by the Artemis/Ecsel Joint Undertaking, supported by the European Commission and French Public Authorities, under grant agreement number 332987.

References

1. Alamo, T., Tempo, R., Luque, A., Ramirez, D.R.: Randomized methods for design of uncertain systems: Sample complexity and sequential algorithms. Automatica **52**, 160–172 (2015)
2. Bilbao, E., Barrade, P.: Optimal energy management of an improved elevator with energy storage capacity based on dynamic programming. In: Energy Conversion Congress and Exposition (ECCE), pp. 3479–3484. IEEE, Raleigh (2012)
3. Boutin, V., Desdouits, C., Louvel, M., Pacull, F., Vergara-Gallego, M.I., Yaakoubi, O., Chomel, C., Crignon, Q., Duhoux, C., Genon-Catalot, D., Lefevre, L., Pham, T.H., Pham, V.T.: Energy optimisation using analytics and coordination, the example of lifts. In: Emerging Technology and Factory Automation (ETFA), pp. 1–8. IEEE, Barcelona (2014)
4. Desdouits, C., Alamir, M., Boutin, V., Le Pape, C.: Multi-source elevator energy optimization and control. In: European Control Conference (ECC). EUCA, Linz (2015)
5. Desdouits, C., Alamir, M., Giroudeau, R., Le Pape, C.: The sourcing problem, energy optimization of a multisource elevator. In: 13th International Conference on INformatics in COntrol, Automation and Robotics (ICINCO 2016), vol. 1, pp. 19–30. Lisbon, Portugal (2016)

6. Paire, D., Simoes, M.G., Lagorse, J., Miraoui, A.: A real-time sharing reference voltage for hybrid generation power system. In: Industry Applications Society Annual Meeting (IAS), pp. 1–8. IEEE, Houston (2010)
7. Sachs, H.M.: Opportunities for elevator energy efficiency improvements. Technical report, American Council for an Energy-Efficient Economy, Washington, DC (2005)
8. The MathWorks Inc.: Matlab 2015b. The MathWorks Inc., Natick (2015)
9. Tominaga, S., Suga, I., Araki, H., Ikejima, H., Kusuma, M., Kobayashi, K.: Development of energy-saving elevator using regenerated power storage system. In: Proceedings of the Power Conversion Conference (PCC), vol. 2, pp. 890–895. IEEE, Osaka (2002)

Robust and Reliable Bionic Optimization of Nonlinear Control Problems

Lukas Haas and Rolf Steinbuch(✉)

Reutlingen University, Reutlingen, Germany
rolf.steinbuch@reutlingen-university.de

Abstract. For nonlinear control problems the use of bionic optimization methods has been tested and qualified. The well accepted ideas to perform robust and reliable optimization in a given space of scattering parameters and input data is expanded to control problems by re-interpreting them as minimization problems. So a wide field of solutions and the influences of the ever present scatter might be handled without too much introduction of new tools but by application of qualified approaches to a new class of tasks. The control of cranes, which defines a well-known range of examples to the applicability of control problems is used to demonstrate the methods used and might help to apply the bionic methods to even more elaborated problems.

Keywords: Bionic optimization · Robustness · Reliability · Control · Crane · Metamodeling · Scatter · Acceleration

1 Introduction

Control problems are often looked at as optimization problems. The difference between a given value or state of a control value and its realization at one point in time or over a time interval might be looked at as the goal of the optimization. To find the control parameters at that moment or during this time interval defines the search of the parameter space while the physical and technical restrictions limit this parameter space, thus defining constraints that are to be taken into account. The possibility of many different feasible solutions requires an optimization strategy that is able not to stick to local optima. Therefore bionic optimization methods could be preferable. Furthermore the inevitable scatter of the input, let it be the physical data or the control functions definition yield uncertainties that require to check the demanding problem of reliability and robustness.

2 Terms and Definitions

In the field of optimization the following terms and definitions are well accepted in the community of researchers [1, 2, 8]:

- Goal or objective: the value we want to maximize or minimize.
- Free parameter: all data we might modify to improve the goal.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_17

- Constraints, restrictions: define inacceptable parameter combinations.
- Local optima: isolated extreme values of the goal in the parameter space.

For robustness and reliability we often introduce the terms (Fig. 1):

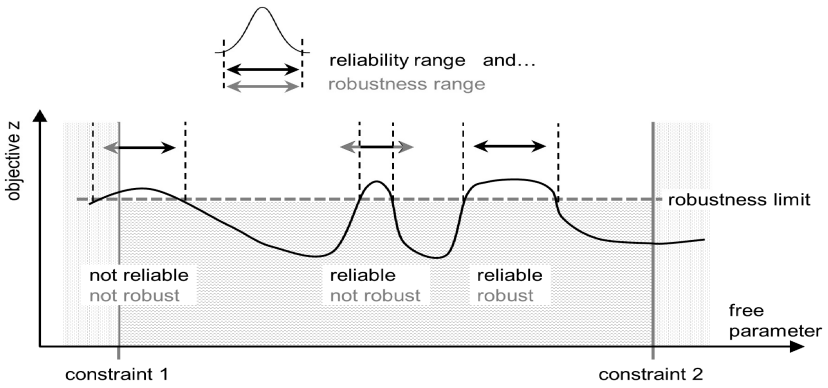


Fig. 1. Reliability and robustness in a 1D Problem

- Scatter: inevitable variation of the problem determining data.
- Reliability: the chance of not failing due to scatter.
- Robustness: the well performance under scatter.

We realize that an optimum that is sensitive to small parameter changes is not robust, while a violation of given constraints defeats a robust solution as it is not reliable. So reliability, robustness and optimality define a chain of qualities that need to be satisfied.

Bionic optimization in contrary to deterministic optimization is a set of strategies to search given parameter spaces for optima with a high chance to find best or very good solutions even in the presence of some or many local maxima. The most important bionic optimization strategies for continuous problems are:

- Evolutionary strategies (ES) [1, 2], which copy the evolutionary process of reproducing natural populations including the sexual crossing of parents' properties and mutation of the kids' properties. The parameters used here are listed in Table 1.
- Particle swarm strategies (PSO) [3, 4] which try to find optima like swarms of birds or fish during their wandering in their living environment.

Both methods are used successfully at many engineering applications [5, 6].

3 The Crane Control Problem

One of the classical control problems is given by a crane that has to place a given payload at a given position [7] (Fig. 2). The payload must not oscillate at the end position so it might be handled without any further difficulties or danger to the people

Table 1. Input data for the evolutionary optimization process

```
nparent =  40;              % # of parents
nkids   =  2*nparent;       % # of kids
span    =  1;               % range of variables
radmut  = .15*span;         % rad of mutation % of span
ngen    =  40;              % # of generations
isurv   =  1;               % parents survive
nloop   =  15;              % # of loops, repeated studies
```

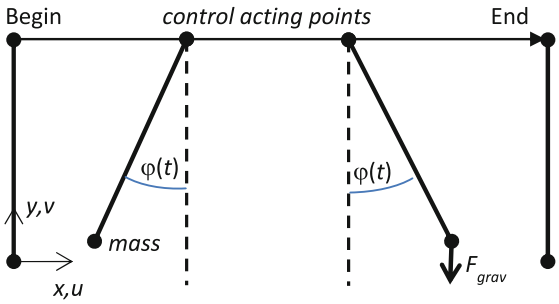


Fig. 2. 1D crane control problem

dealing with it. The control might be done by the positioning of the upper end of the crane, the sheave. Here we use a longitudinal rope stiffness k = 100,000 N/mm, a rope length l = 1,000 mm and a payload mass m = .001 t and a small damping c. The well-known equations of motion could be written as

$$m\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}\ddot{u}\\ \ddot{v}\end{pmatrix}+c\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}\dot{u}\\ \dot{v}\end{pmatrix}+\begin{pmatrix}\sin^2\varphi & \sin\varphi\cos\varphi\\ \sin\varphi\cos\varphi & \cos^2\varphi\end{pmatrix}\begin{pmatrix}u\\ v\end{pmatrix}=F_{grav}\begin{pmatrix}0\\ 1\end{pmatrix} \tag{1}$$

The goal is a weighted sum of the square of the offset at the end of the transport and the square of the final velocity of the load. We accept solutions as soon as their goal is below a given value (using here weights of 0.5 for both components of the goal):

$$goal = w_1 * distance^2 + w_2 * velocity^2 \tag{2}$$

Here *distance* stands for the error in final position, *velocity* for the unwanted end velocity. The parameters of the optimization process are the positions of the cranes sheave over time. We assume a constant base velocity for most of the process time and at the end a standstill. A certain overlay is added to this positioning at the end of the transport (Fig. 3).

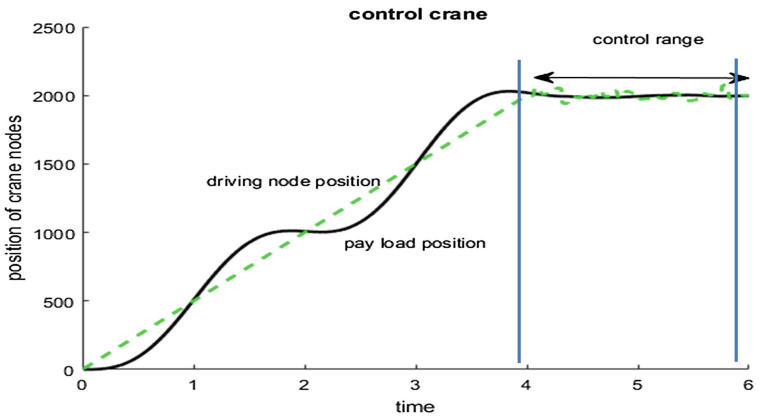


Fig. 3. Crane positions in time

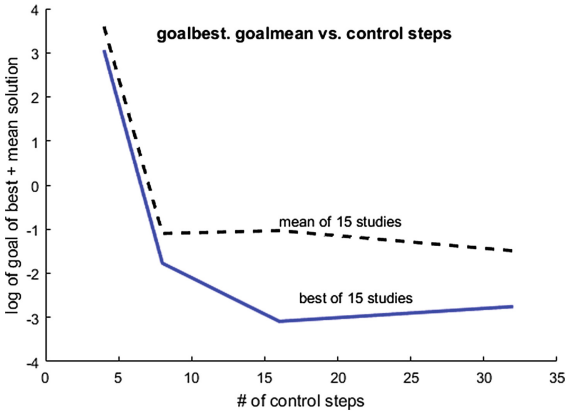
The optimization now has to be done by a qualified selection of this overlay positions. As constraints for the reliability studies we use the maximum axial deflection of the payload at the end position and the maximum velocity of the payload.

4 Bionic Optimization of the Crane Travel History

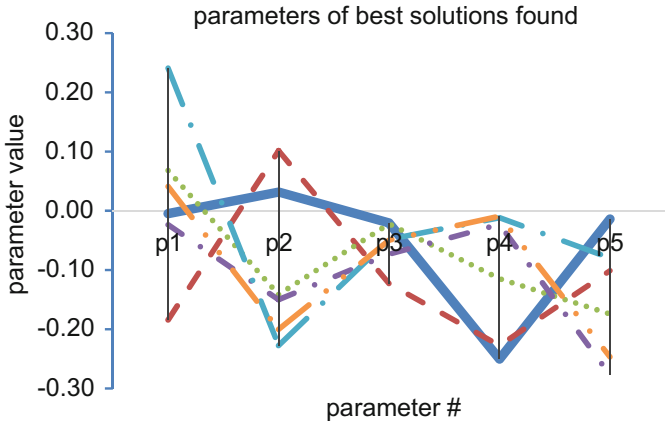
The well-known problem might easily be solved by using classical control theory and Laplace transformations [7]. Here we transform it into a bionic optimization task by using the overlay crane position as free parameters to minimize the goal e.g. given by Eq. (2). We assume the cranes sheave to move from one position to the next without any time delay. To come up with an idea of the performance dependent on the number of control steps we modify this number of control steps during the time called 'control range' in Fig. 3. For the current study and the results displayed we used an Evolutionary Strategy. Its data are listed in Table 1 [1]. By some preliminary studies we found that a population of 40 parents that produces a total of 80 kids is able to model a nicely converging process. To produce a kid two parents are selected randomly, their properties are averaged and then distorted within a given mutation radius. The best performing elements from the total set of the kids and the old parents are selected to be parents in the next generation. After 20 and 30 generations the mutation radius is reduced by a factor of 2 to increase convergence velocity.

From Fig. 4a we learn that a number of 8 control steps seems to be able to produce satisfactory values for the final position and velocity. The plot indicates that even if there are less well performing solutions, 15 loops or repetitions should provide acceptable proposals. We assume that the sheave is at the final position for the last 3 control steps so only 5 active control parameters have to be defined. Nevertheless, these 5 control steps produce results that are within a range of sufficiently small deflection and velocity.

Figure 4b indicates by some lines for the control parameters values that there is not a unique best solution but different ones able to come up with good values of the goal.



a) Quality of the control the # of control steps



b) Parameters of the control process at different good solutions

Fig. 4. Performance of optimizer depending on # of control steps

So we have to decide which of the solutions is to be preferred. To classify the different solutions we interpret the lists of control positions as a 5-dimensional vectors and calculate the cosine of the angle between them by the dot-product. If we agree on a limit to distinguish vectors by a certain angle, we identify all pairs of vectors who differ by a smaller angle.

5 Reliability and Robustness

Table 2 indicates the values of the goal and the constraints for 6 well performing variants of solutions from Fig. 4b. We realize, that the chosen constraints of maximum velocity (*r_velo*) and maximum longitudinal deflection (*r_end*) do not depend on the control parameters. Up to now we have been dealing within a perfect world where all data have been precisely known and all control activities are perfectly done. In real cases none of this is true so we have to deal with the scatter of the

- length of the rope,
- mass of the payload,
- values of the control parameters.

Table 2. Results for some well performing solutions

Vari.	p1	p2	p3	p4	p5	Goal	r_velo	r_end
1	−0.00463	0.03155	−0.02057	−0.25052	−0.01398	0.00005	1078.0	30.1
2	−0.18422	0.10168	−0.12312	−0.22780	−0.10138	0.00084	1078.0	30.1
3	0.06809	−0.14158	−0.02204	−0.11472	−0.17420	0.00087	1078.0	30.1
4	−0.02365	−0.15016	−0.07317	−0.02300	−0.27739	0.00130	1078.0	30.1
5	0.24110	−0.22782	−0.04809	−0.01164	−0.07953	0.00165	1078.0	30.1
6	0.04107	−0.20060	−0.04952	−0.00920	−0.24688	0.00169	1078.0	30.1

Assuming that we know the mean and the scatter of all of these data we could start doing a reliability and robustness study. But to get the goal and restriction values we need a number of additional runs taking into account these scatters. From [8] we know that the number of additional n_{run} runs should be about (if n_p denotes the number of free parameters)

$$n_{run} > 2[1 + n_p + \frac{n_p(n_p+1)}{2}] \approx n_p^2 \tag{3}$$

For 5 active control parameters and 3 scattering problem data we are in the range of about 60–70 additional jobs which is not too much for such a small problem. For more complicated control problems where the number of control steps is essentially higher, this number might easily exceed any acceptable value. From Table 3 we learn about the dependencies (first and second derivatives *f'* and *f''*) of goal and restrictions if the input data change by 1%. This information helps us to compensate changes of e.g. the length by a correspondent change of the control parameters. On the other hand we realize, that changes in mass would not affect much the restrictions but the goal, while changes in length yield essential restriction deviations.

For the further studies we select one of the good control proposals and assume that all data entries (control overlay, mass, length, end position) scatter by 1%. The values achieved by the scatter of the input data are now used to build up Response Surfaces (RS). We assume the RS to be given e.g. by low order polynomials. Often we use second order functions in the scattering data:

Table 3. Sensitivity of goal and restrictions to parameter scatter

Param	Init. value	Goal f'	Goal f"	r_velo f'	r_velo f"	r_end f'	r_end f"
Mass	0.0010	−2124.18	2793910.00	0.00	0.00	−0.02	0.00
Length	1000.0000	0.30	31.00	45.00	1000.00	−974.24	−474.00
p1	0.2743	−9.99	11445.00	0.00	0.00	0.00	0.00
p2	−0.1683	2.17	1937.00	0.00	0.00	0.00	0.00
p3	−0.0847	−0.38	4546.00	0.00	0.00	0.00	0.00
p4	−0.0050	−0.12	11.00	0.00	0.00	0.00	0.00
p5	−0.0015	0.03	0.00	0.00	0.00	0.00	0.00

$$RS = a_0 + a_1 p_1 + a_2 p_2 + \ldots + a_n p_n + a_{11}^2 p_1^2 + a_{12} p_1 p_{12} + \ldots + a_{n-1,n} p_{n-1} p_n + a_{nn}^2 p_n^2 \quad (4)$$

The coefficients of this RS might be found by a least square method or any other appropriate approach. Such a RS might now help to interpolate the function values without additional repetition of the time consuming integration of the dynamic system shown in Eq. (1). So once a problem is handled and a sufficient set of solutions near an interesting point are known, further function evaluations are available fast and, dependent on the quality of the approximation, with a small error.

6 Interpretation

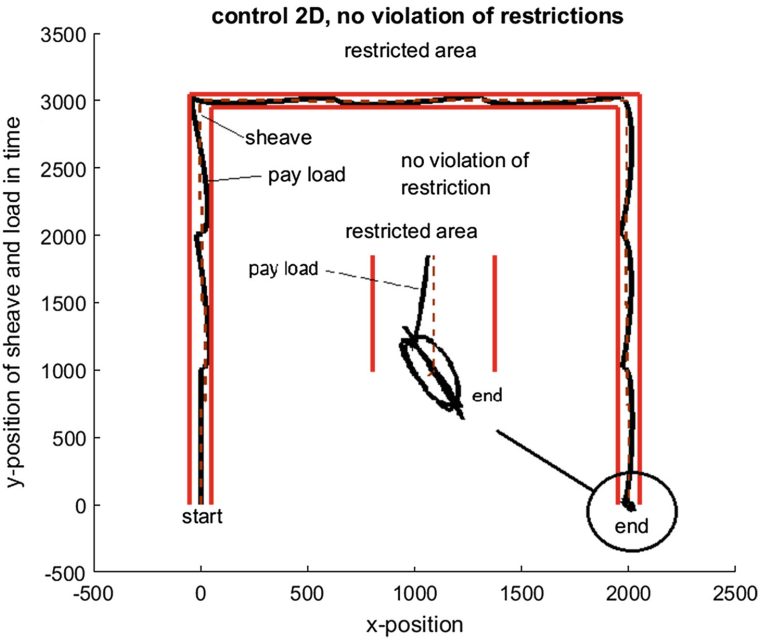
From the results and some additional studies we learn from the simple 1D-crane problem that any scatter yields results that are less or more deviating from the original one. On the other hand we might use the RS or another appropriate approximation to find parameter combinations that produce good results even in the presence of this inevitable scatter. We just have to solve Eq. (4) for those values of the control parameters that compensate the deviation due to the non-perfect values of some input data. As long as there is a higher number of free and adjustable entries p_i in Eq. (4) than there are goals and restrictions to be satisfied, there is a good chance to find such a compensation of the defective input.

7 Expansion to More Complex Problems

The ideas outlined are not only useful for the simple 1D-crane problem. Due to the use of bionic optimization we are not even restricted to control processes where the goal or the constraints are given by continuous or even smooth functions. We only need to access the results for some (n_{run}, Eq. (4)) parameter combinations within reasonable time. From Eq. (4) we know that the total number of studies to be performed is growing with the square of the number of free parameters n_p. Often this is going to be a central problem. Nevertheless any control problem which could be interpreted as a minimization task could be solved by using the ideas outlined.



a) U-shaped travel path of payload, violations of restrictions



b) Add more control steps and generations: no violation of restrictions, but still oscillations at end of travel

Fig. 5. 2D crane running between limiting borders

Example: We expand the crane control problem to 2D using the same values for mass, length and stiffness like before. The payload has to be delivered along a U-shaped path from the initial position at (0, 0) to the end position at (2000, 0) within 18 s (see Fig. 5a). First traveling a transition of 3000 mm along the *y*-axis (0, 3000), then 2000 mm along the *x*-axis to (2000, 3000), then back down to the end point (2000, 0).

Following the transport time there are 6 s to remove the remaining oscillations. The payload must not deviate from the design path by more than 50 mm, indicated by the restricted area in Fig. 5a. We assume large deformations and rotations of the rope, so non-linear expansions of Eq. (1) have to be used.

The initial study to derive the control steps is done using 24 parents, 48 kids and 10 generations, the initial mutation radius being 15% of the parameter range (±50 mm) for the control node, the sheave. The mutation radius is reduced during the study like in Sect. 4. The dotted lines in Fig. 5a and b depict the positions of the sheave as it travels through the given range. In this study 12 control steps are activated on each side of the path. The thick continuous line represents the position of the payload. It indicates that there are some violations of the restrictions (Fig. 5a). In consequence this solution is not acceptable.

If we increase the number of generations to 20, there are no more violations of the restrictions (Fig. 5b). But the remaining oscillations at the end position are still too large. Further increasing of the number of generations and the number of control points along the travel distance allows to reduce these oscillations to very small values. So our approach proves to be able to solve the given problem. Expansions to other and more realistic problems will follow in the further steps of our current research program.

8 Conclusions

Bionic Optimization has been shown to be able to solve nonlinear and complex control problems. The general applicability makes it an efficient tool whose performance does not depend much on the question posed. In some applications dealing with difficult excitations it might be a powerful approach. Non continuous or irregular boundary conditions, large jumps and changes of the responses of loading and structure are managed successfully as long as there is a chance to find well performing solutions. But it must not be disregarded that in most control problems, the classical methods of control theory prove to be essentially faster.

The use of meta-models like the Response Surfaces allows for fast repetitive studies. A limiting factor for the application of the ideas presented is the total number of individual studies necessary in the case of large numbers of free parameters. So it is restricted to applications where all the possible constellations might be analyzed a priori and in reasonable time. Then the approximations by the Response Surfaces could be produce fast answers for given situations.

In real world applications preliminary studies using reduced parameter sets allow to limit the dimension of the parameter space to be searched. This could help to accelerate the setup of the responses. But every reduction of the solution space inhibits the danger of excluding promising solutions.

References

1. Gekeler, S., Pandtle, J., Steinbuch, R., Widmann, C.: Remarks on the efficiency of bionic optimisation strategies. J. Math. Syst. Sci. **4**, 139–154 (2014)
2. Rechenberg, I.: Evolutionsstrategie ’94. Frommann-Holzboog, Stuttgart (1994)
3. Kennedy, J., Eberhardt, R.: A discrete binary version of the particle swarm algorithm. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, pp. 4104–4108 (1997)
4. Plevris, V., Papadrakakis, M.: A hybrid particle swarm—gradient algorithm for global structural optimization. Comput.-Aided Civ. Infrastruct. Eng. **26**, 48–68 (2011). Blackwell Publishing Inc
5. Gen, M., Cheng, R.: Genetic Algorithms & Engineering Optimization. Wiley, New York (2000)
6. Steinbuch, R.: Successful application of evolutionary algorithms in engineering design. J. Bionic Eng. **7**(Suppl.), 199–211 (2010)
7. Geering, H.P.: Regelungstechnik, mathematische Grundlagen, Entwurfsmethoden, Beispiele, 4th edn. Springer, Berlin (1996). (in German)
8. Gekeler, S., Steinbuch, R.: Bionic Optimization in Structural Design, pp. 125–146. Springer, Berlin (2016)

Human Detection Using Biological Signals in Camera Images with Privacy Aware

Toshihiro Kitajima[1,2(✉)], Edwardo Arata Y. Murakami[1], Shunsuke Yoshimoto[2], Yoshihiro Kuroda[2], and Osamu Oshiro[2]

[1] Samsung R&D Institute Japan, ET Center ET-2 Lab, Minoh Semba Center Bldg. 2-1-11 Semba Nishi, Minoh City, Osaka, Japan
{t.kitajima,arata}@samsung.com
[2] Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka, Japan
{yoshimoto,ykuroda,oshiro}@bpe.es.osaka-u.ac.jp,
http://www.samsung-srj.co.jp
http://oshiro.bpe.es.osaka-u.ac.jp

Abstract. In the Internet of Things (IoT), in which people sensing occurs everywhere, cameras are being increasingly used because they provide inexpensive and effective sensing devices. However, while cameras can provide significant amounts of information, much of that information is personal and there are significant concerns that individual privacy could be compromised. Furthermore, since home appliances are increasingly being connected to the Internet via the IoT, it has become possible for user images to leak out unintentionally. With these concerns in mind, we propose a human detection method that protects user privacy by using intentionally blurred images. In this method, the presence of a human being is determined by dividing an image into several regions and then calculating the heart rate detected in each region. In our performance evaluation, the proposed method showed dominant performance results when compared with an OpenCV-based face detection method, and was confirmed to be an effective method for detecting human beings in both normal and blurred images.

Keywords: Human detection · Heart rate · Privacy · Camera images · IoT

1 Introduction

In the current Internet of Things (IoT) era, vehicles such as trains, airplanes, and automobiles, along with appliances such as televisions and refrigerators, are increasingly being connected to the Internet, which they automatically send to and receive from a wide variety of data. Additionally, various sensors are being attached to consumer electronics and the data they collect are increasingly being sent from and received by products in order to improve the services they provide.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_18

Since users can be expected to select home appliance services that support their conditions, it would be helpful if those services could sense the state of the humans present in their residences.

Meanwhile, as the cost of cameras has declined drastically in recent years, their usage as human sensing devices has increased. However, even though the amount of information that can be gleaned from camera imagery is quite large, which makes cameras very effective as sensing devices, personal information such as faces, clothes, etc., is also acquired, and there are concerns that personal privacy will be compromised. Furthermore, since home appliances are increasingly being connected to the Internet via the IoT, the possibility exists that user image data may leak out unintentionally (Fig. 1). Since such concerns have already convinced numerous consumers not to purchase Internet-connected or camera-equipped products, the development of a sensor device that safeguards the user's privacy is becoming increasingly important.

There are currently a number of devices, such as pyroelectric [1] and temperature sensors [2], that can perform human detection without the use of camera images. These devices can detect the presence of a human within their sensing range, but they cannot detect his or her direction and position. Additionally, while thermopile array sensors [3] can be used to measure the presence of a human, their resolution is very rough, which makes it difficult for them to determine exact locations. Furthermore, individual thermopile sensors are expensive, and their cost increases significantly when they are used as part of a thermopile array.

With these points in mind, this study targets the development of a method whereby a camera can be used to measure the presence, direction, and position of a human being while simultaneously protecting that individual's privacy. Some methods that aim to protect the privacy of camera image data have already been reported. These include software methods to blur human faces after an image is acquired [4,5], a method for replacing the region of the human face with a computer graphics (CG) generated image [6], and a method that requires a framework authentication key to view the images [7]. However, in all of those methods, there is a possibility that the original camera image could be hacked, so it is believed that privacy would be better protected if the image could be made more secure at the acquirement stage.

There are a number of methods for detecting humans using a camera image. One is the face detection method using Haar features and AdaBoost [8], and another is a human detection method using histogram of oriented gradients (HOG) features [9]. These methods are effective in clear images, but they do not work in images that lack sharpness, because contrast and edges characteristics cannot be extracted. The method used in this study applies a different concept that does not consider image appearance features. Instead, our method performs human detection using biological signals obtained from those images. More specifically, before images are captured with the camera, either a frosted glass is placed in front of the camera lens or the focal point of the lens is shifted until the image is sufficiently blurred to protect the user's privacy. Acquired

images are then divided into several regions and fast Fourier transform (FFT) is used to determine whether something is present in a region that has a value corresponding to the heart rate of a human being. In order to verify the effectiveness of this proposed method, a performance evaluation in which it is compared with an existing human detection method is carried out.

The heart rate-based human detection method is described in Sect. 2. A performance comparison in which our method is compared with an existing human detection method using both blurred and normal images is carried out in Sect. 3. A summary is provided and future works are described in Sect. 4.



Fig. 1. Leakage of user privacy via Internet and camera.

2 Method

Currently used human detection methods draw on significant amounts of appearance information, and thus require clear images. However, when it is necessary to protect the user's privacy, clear images are contraindicated, so we propose a method that does not require their use. Reports on a method of measuring heart rates from camera images [10,11] have been published previously. Since this method detects a face region and measures the heart rate in that area, we determined that it would be possible to use this heart rate measurement method to detect the presence of human beings. Specifically, the whole camera image is divided into regions, after which heart rate measurements are taken for each region. If a measured value that corresponds to a human heart rate is detected in a specific region, it is possible that a human face or skin is present in that region.

2.1 Principle of Heart Rate Measurement

In this subsection, the principles of obtaining heart rate measurements from human skin are explained. When used in the context of human blood circulation, pulse refers to the pulsation of blood arteries, which contain large amounts of oxygenated hemoglobin. Figure 2 [13] shows the absorption spectrum of human biological tissue material. Here, it can be seen that oxygenated

hemoglobin absorbs significant amounts of light in the wavelengths from 500 to 580 nm. Figure 3 [14] shows the corresponding relationship between wavelength and color. In this figure, it can be seen that the wavelengths from 500 to 580 nm correspond to green and yellow. More specifically, yellow is from 560 to 580 nm and green is from 490 to 560 nm. Thus, the green wavelength range is wider than the yellow range, and more green light is absorbed into the oxygenated hemoglobin. Commercially available finger-type pulse wave [15] and wristwatch-type [16] sensors measure heart rates by utilizing the property of the color green that is absorbed by the oxygenated hemoglobin. Figure 4 [17] shows the heart rate measurement principle that uses green light. The apparatus is provided with a green light-emitting diode (LED) and a photo-detector. Human skin is irradiated with green light by the LED, and the photo-detector receives the reflected light. As the volume of the blood vessel changes due to the pulse, the value received by the photo-detector varies as a waveform, as can be seen in the right graph of Fig. 4. The heart rate is detected from this waveform.

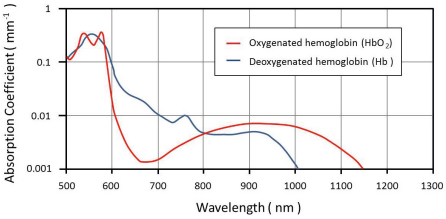


Fig. 2. Absorption spectrum of main substances in living tissue [13].

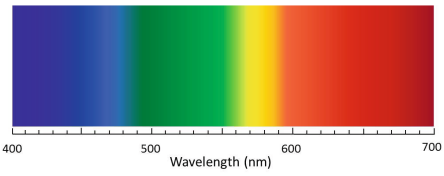


Fig. 3. Wavelength and color of the corresponding relationship [14].

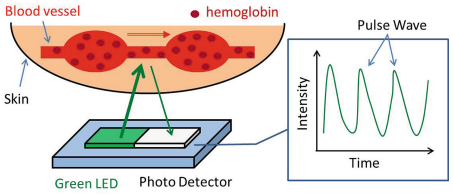


Fig. 4. Pulse wave sensor principle [17].

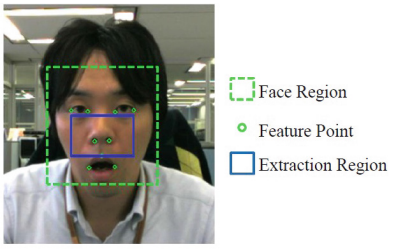


Fig. 5. Face detection and selected extraction region.

When this mechanism is applied to photographs, faces are detected using a commercially available face detection library, as shown in Fig. 5. Eye and mouth movements, as well as the area between the eyes and mouth, are acquired from the face feature points. Since the light in the room and sunlight contains the color green, a red-green-blue (RGB) camera extracts green (G) light as the

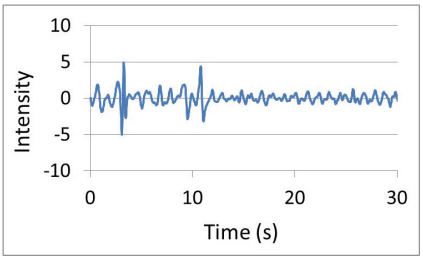


Fig. 6. Heart rate signal extracted from the face region time series data.

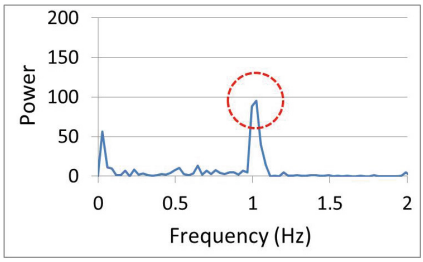


Fig. 7. FFT calculation of heart rate signal showing the detected peak frequency.

reflected wave. Figure 6 shows time series G data in this area. These data include a pulse component, which is equal to the heart rate. Figure 7 shows the result of the data frequency analysis contained in Fig. 6. When the peak surrounded by the red dotted line is detected, it is possible to calculate the human heart rate.

2.2 Human Detection Method

First, before images are captured with the camera, frosted glass is placed in front of the lens or the lens focal point is shifted to reduce image sharpness until human faces cannot be identified. The image in Fig. 8 was obtained by a normal camera, but by shifting the focal point of the lens, the image shown in Fig. 9 was acquired. Next, the entire image is divided into several regions, as shown in Fig. 10. Then, by using the heart rate measuring method discussed in [10], it is possible to calculate the heart rate in each region.

Regions A and B in Fig. 11 show background and human face areas, respectively. Figures 12 and 13 show 30 s of data for the average green intensity values in each region obtained via an RGB camera. Since Fig. 12 shows a background region, there are no substantial changes in green intensity values. In contrast, Fig. 13 shows a human face region, and it can be seen that waveforms indicating a human pulse have been obtained. Next, FFT is performed as input data for these green intensity values. The window function is the rectangular window. The time range window shows the data from current time extending back to the past. Furthermore, the time window moves successively, while overlapping the range processed immediately before. Figures 14 and 15 show the FFT result in each region of the background and human areas. Based on the power spectrum results, it is possible to determine whether data of these regions match the pulse of human being. Since measured values that do not correspond to human heart rates are outside the search range of this method, a heart rate range of 50 to 120 beats per minute (bpm) is set. Strenuous exercise or illness is excluded because the daily life of a human at home is the target of this method. The light gray areas in Figs. 14 and 15 shows areas where the heart rate is less than 50 bpm, and thus outside the scope of the search. Furthermore, when a detected power spectrum is weak, it is unlikely to be a human pulse. Therefore, a threshold value was set, and all areas outside the power spectrum of a human pulse

were excluded. The frequency of the maximum value is selected if multiple power readings exceeding heart rate threshold power are detected. In Figs. 14 and 15, the red dotted line represents the threshold value. Based on this criterion, there is no power value exceeding the threshold in Fig. 14. This indicates that there is no human presence in that area. In Fig. 15, the detected 0.95 Hz frequency power exceeds the threshold. $60 \times 0.95 = 57$, thus indicating a heart rate of 57, so it can be determined that there is a human in this area.

Figure 16 shows the results of this process for all regions. Regions where the power does not exceed the threshold value show a heart rate value of zero. In contrast, the human face region in Fig. 16 shows a heart rate of 57 bpm, so it can be concluded that there is a human in this region. By performing such processing, it is possible to detect the presence of a human being.

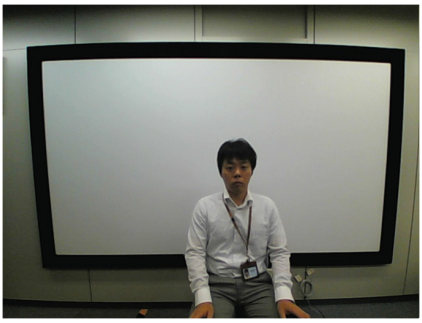


Fig. 8. Normal camera image.

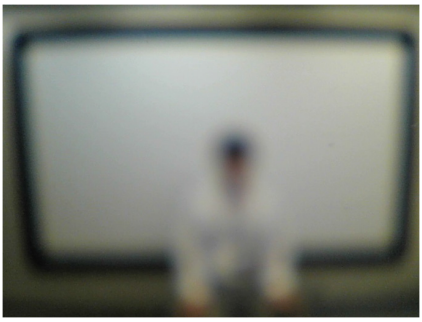


Fig. 9. Out-of-focus image.

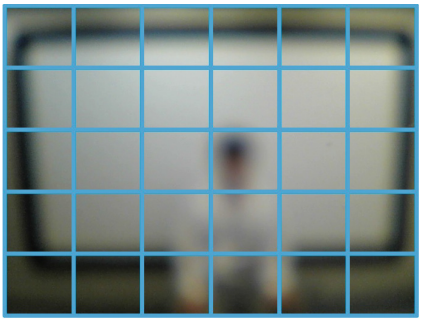


Fig. 10. Divided image to several small regions.

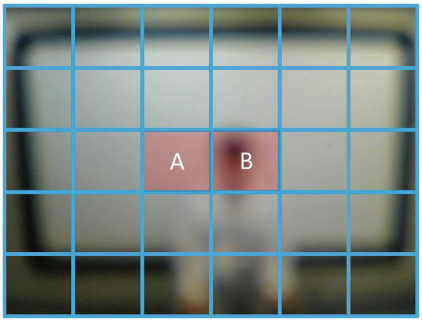


Fig. 11. Background region (A) and human region (B).

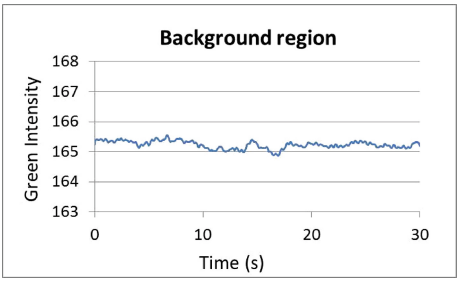


Fig. 12. Green light intensity average of pixels inside the background region (A).

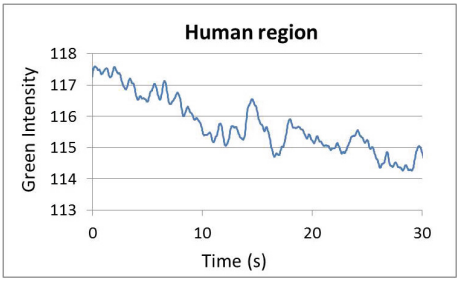


Fig. 13. Green light intensity average of pixels inside the human region (B).

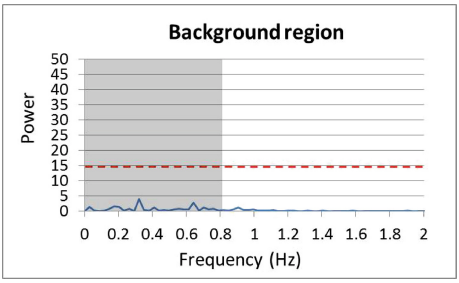


Fig. 14. FFT calculation of the signal in the background region (A). The red dotted line represents threshold values.

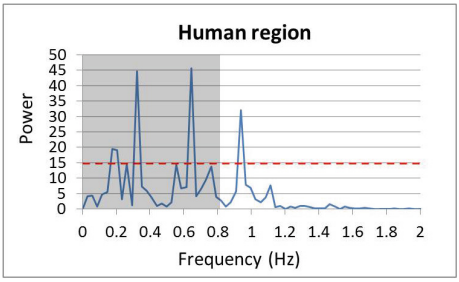


Fig. 15. FFT calculation of the signal in the human region (B). The light gray area is outside the scope of the search.

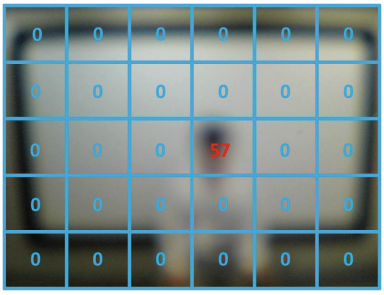


Fig. 16. Heart rate detection results for each region.

3 Performance

3.1 Experiment Environment

To ascertain the utility of our method, performance evaluation of its human detection capability was conducted. In these experiments, a web camera [12] type that is commonly used for video chats, as shown in Fig. 18, was employed. The evaluation using this camera was performed at 30 fps and VGA resolution. The camera specifications are shown in Fig. 17.

Type	RGB
Resolution (pixels)	640 ×480
Frame rate (fps)	30
Gradation (bits)	8 (1 ch)

Fig. 17. RGB camera specifications.



Fig. 18. RGB camera.

Since this report is focused on detecting the heart rate of human beings in stationary states, experiments involving the moving state were not conducted. Subjects sat 1 m from the camera, and stayed in a rest position. The proposed method was compared with an existing face detection method that was implemented in OpenCV Ver. 2.4.3. The minimum search window size of OpenCV is 40 × 40 (pixels), and the scaling factor of the search window was set to 1.2. For our proposed method, the divided region was also 40 × 40 (pixels). The FFT time window of the proposed method was set at 30 s and recording was performed for one minute and 30 s. Since the detection rate of the first 30 s was excluded from the evaluation, the FFT result for the first 30 s was not output. 1800 images remaining one minute were defined as evaluation image. The OpenCV method was also used to evaluate the same 1800 images. Five test subjects participated in this experiment.

The following detection criteria were used. The reference area, which represents the region in which there is human presence, was set manually for all images. If the output area of our human detection method overlaps the reference area, it is determined that human detection is successful. If the areas are not duplicates of each other, a false detection is recorded.

3.2 Experiment Results

While the comparative evaluation of our proposed method and OpenCV was performed using intentionally blurred images, the proposed method is also capable of human detection using normal images. Therefore, an additional evaluation was performed using normal images.

Intentionally Blurred Images. Three types of intentionally blurred images were used. The first were out-of-focus images, as shown in Fig. 9. The second and third were images taken by mounting optical filters in front of the lens.



Fig. 19. Image taken with cyclone filter.



Fig. 20. Image taken with radial filter.

Figures 19 and 20 respectively show photos taken with the cyclone [18] and radial filters [19] used in our experiment.

The receiver operating characteristic (ROC) curve was used in the comparison of the experimental results. The ROC curve shows a representation of detection error rates (false positives) on the horizontal axis and the detection rate (true positives) on the vertical axis. It is possible to create different curves by varying the detector threshold. As the curve is close to the upper left corner, it can be said that the detection performance is good. 27000 images were used as the evaluation images (1800 [images] × three [filters] × five [subjects] = 27000 [images]).

One example of the human detection results generated by the proposed method is shown in Fig. 21. The sole detection region is displayed in the green rectangles. The numbers in the rectangle reflect the detected heart rate.

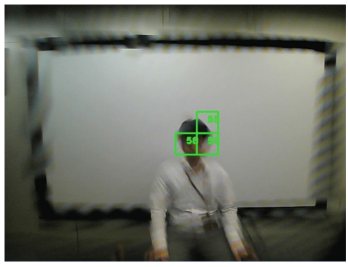


Fig. 21. Human detection result image generated by our proposed method (cyclone filter image). The detection region is displayed in the green squares. The numbers in the squares reflect the detected heart rates.

As shown in Fig. 22, the results from our proposed method are closer to the upper left than those produced by OpenCV. This indicates that the proposed method provides better performance. When the false detection rate was 2.7%, the detection rate was, respectively: proposed method 99.5%, OpenCV 4.1%,

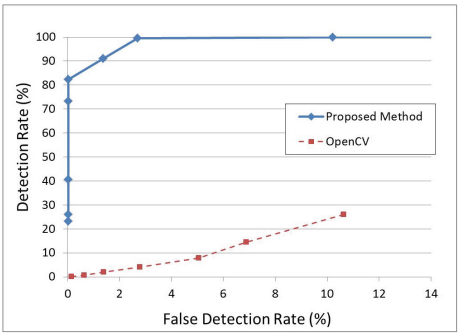


Fig. 22. ROC curves of human detection results with intentionally blurred images. The closer the curve is to the upper left corner, the better the detection performance level is.

which indicates a significant difference. As a whole, OpenCV had low detection rates due to the lack of contrast and edge necessary for facial feature detection. In contrast, the proposed method showed both a high rate of detection and a low rate of false positives. Based on these results, we can state that our proposed method provides an effective human detection method in circumstances where image privacy protection is an important consideration.

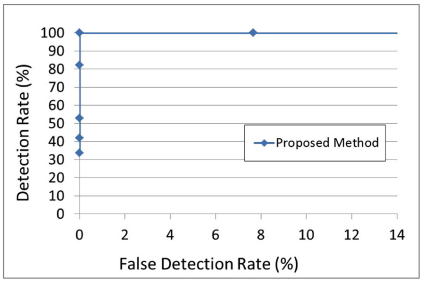


Fig. 23. ROC curve result of the proposed method using normal images.

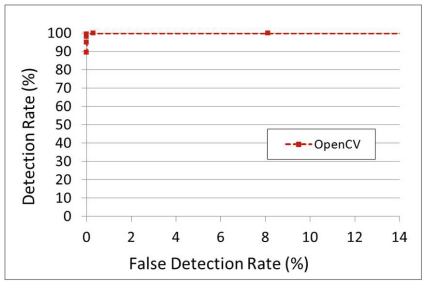


Fig. 24. ROC curve result of OpenCV face detection method using normal images.

Normal Image. Next, a comparison evaluation of the proposed method and OpenCV was performed using normal images, such as that shown in Fig. 8. A total of 9000 images were used in this stage of the experiment (1800 [images] $\times$ five [subjects] = 9000 [images]). The experimental results comparison used the ROC curve in the same manner as discussed above. Overall, the proposed method and OpenCV showed similar results, and it was difficult to comprehend the overlapping curves. Accordingly, the results were shown separately in Figs. 23 and 24 using the ROC curve.

As can be seen in Fig. 24, OpenCV is very close to the upper left, which means it can detect humans almost 100% of the time, with few false positives. However, the proposed method is also very close to the upper left, showing that it too can detect humans almost 100% of the time, with few false positives. From this result, it can be said that the proposed method provides detection performance levels that equal the performance of the existing method, even when normal images are used.

4 Conclusion

Herein, we proposed a novel human detection method that makes use of intentionally blurred images to prevent a subject from being identified, thereby protecting individual privacy. In our method, collected images are divided into regions and the presence or absence of a human being is determined by calculating human heart rates detected in each region. In a performance evaluation using intentionally blurred images, our proposed method showed a higher level of performance than an existing OpenCV-based face detection method. As a result, the proposed method was confirmed to be an effective technique for detecting humans using intentionally blurred images. Additionally, in a separate performance evaluation using sharp normal images, our proposed method showed a level of performance equivalent to the existing OpenCV-based method. This result confirmed that the proposed method can be effectively used with normal images as well. In its current state of development, the proposed method has only considered humans in a rest position. However, since performing detection on moving subjects is often necessary, it will be also be necessary to consider how the inter-frame difference method can be integrated into our proposed method in order to facilitate detection of moving humans. Furthermore, this paper limited detection to a single person. However, if more than one person present, it will eventually be necessary to count people via use of an algorithm for grouping the detection area. Accordingly, it is clear that there is still a need for improvements to our proposed method, and these areas will be the subjects of our future work.

References

1. muRata Infrared Sensors. http://www.murata.com/en-global/products/sensor/infrared. Accessed 2 Aug 2016
2. Thermopile Sensors. http://www.heimannsensor.com/products.php. Accessed 2 Aug 2016
3. Thermopile Array Sensors. http://www.omron.co.jp/ecb/products/sensor/special/humandetectingsensor/. Accessed 2 Aug 2016
4. Dufaux, F., Ebrahimi, T.: Scrambling for video surveillance with privacy. In: Proceedings of IEEE Computer Vision and Pattern Recognition Workshop on Privacy Research in Vision (2006)
5. Yu, X., Babaguchi, N.: Privacy preserving: hiding a face in a face. In: Proceedings of the 8th Asian Conference on Computer Vision, pp. 651–661 (2007)

6. Mitsugami, I., Mukunoki, M., Kawanishi, Y., Hattori, H., Minoh, M.: Privacy-protected camera for the sensing web. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems (2010)
7. Zhang, W., Cheung, S.S., Chen, M.: Hiding privacy information in video surveillance system. In: Proceedings of International Conference on Image Processing (2005)
8. Viola, P., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: IEEE CVPR, pp. 511–518 (2001)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE CVPR, pp. 886–893 (2005)
10. Kitajima, T., Choi, S., Murakami, E.A.Y.: Heart rate estimation based on camera image. In: 14th International Conference on Intelligent Systems Design and Applications (2014)
11. Kitajima, T., Choi, S., Murakami, E.A.Y., Yoshimoto, S., Oshiro, O.: Contactless measurement of heart rate using single channel camera images. In: 12th International Conference on Ubiquitous Healthcare (2015)
12. Web Camera Buffalo BSW20KM11BK. http://buffalo.jp/product/multimedia/web-camera/bsw20km11bk/. Accessed 2 Aug 2016
13. Absorption spectrum of main substances in living tissue. http://www.hamamatsu.com/jp/ja/technology/innovation/trs/index.html. Accessed 2 Aug 2016
14. Wavelength, color of the corresponding relationship. http://www.ieij.or.jp/what/yougo.html. Accessed 2 Aug 2016
15. iHealth Wireless Pulse Oximeter. http://www.ihealthlabs.com/health-and-fitness-products-wireless-wireless-pulse-oximeter_80.htm. Accessed 2 Aug 2016
16. Mio ALPHA Heart Rate Sport Watch. http://www.mioglobal.com/product_info.htm. Accessed 2 Aug 2016
17. Pulse wave sensor principle. http://www.rohm.com/web/global/pulse-wave-sensor. Accessed 2 Aug 2016
18. Kenko ZS-CYCLONE. http://www.kenko-tokina.co.jp/discontinued/filters/ef_around/4961607349403.html. Accessed 2 Aug 2016
19. Kenko ZS-RADIAL. http://www.kenko-tokina.co.jp/discontinued/filters/ef_around/4961607349397.html. Accessed 2 Aug 2016

Nuclei Malignancy Analysis Based on an Adaptive Bottom-Hat Filter

Tiia Ikonen[1], Keijo Haataja[1], Pekka Toivanen[1(✉)], Teemu Tolonen[2], and Jorma Isola[2]

[1] School of Computing, Kuopio Campus, University of Eastern Finland,
P.O. Box 1627 70211 Kuopio, Finland
{Tiia.Ikonen,Keijo.Haataja,Pekka.Toivanen}@uef.fi
[2] Institute of Biomedical Technology, University of Tampere,
Biokatu 8, 33520 Tampere, Finland
{Teemu.Tolonen,Jorma.Isola}@uta.fi

Abstract. In this paper, we study the concept of adaptive thresholding and present a new adaptive bottom-hat filter for the extraction of nuclei and analysis of malignancy in an image. The adaptiveness of the used structure element enables a more accurate and location-wise processing of areas in order to extract nuclei clearly for further processing such as classification of morphological features. In addition to the adaptive method we will present a color channel manipulation as a preprocessing step that adduces real benefit for the extraction phase. Some preliminary results are presented to illustrate the true potential of our method in the analysis of breast cancer. These results show that the color channel manipulation together with adaptive filtering extract the image in such a way that mainly only the nuclei is present.

Keywords: Breast cancer · Nucleus · Adaptive bottom-hat filter

1 Introduction

A pathologist distinguishes the cancerousness of a tissue based on the shape and size of nuclei inter alia. Performing this act manually is a time-consuming task which in addition to speed could be executed more objectively with a computer-assisted system. Creating this kind of system would require the use of state-of-the-art methods for the extraction of nuclei as well as the classification of these nuclei to either cancerous or healthy as well as benign and malign depending on the type of the cancer.

A graylevel image consists of bright points and dark regions which considering our breast cancer images means that bright points are background and dark regions are the nuclei that are under investigation. Top-hat filtering and bottom-hat filtering are basic methods of mathematical morphology for small target detection and the use of the methods in the nuclei extraction is justified due to their characteristics. However, the basic operations as such are not sufficient enough to detect nuclei effectively which has resulted in the creation of adaptive filters.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_19

Our results: In this paper, the idea of adaptive thresholding is studied and a novel adaptive bottom-hat filter is presented for the extraction of nuclei. In addition, based on our previous work, a color channel manipulation is combined with the adaptive filter to increase the extraction rate of nuclei.

The paper is organized as follows. Section 2 provides a brief introduction to the morphology of nucleus and Sect. 3 presents the idea and preliminary results of adaptive bottom-hat filtering. The results of adaptive bottom-hat filter combined with color channel manipulation are presented and discussed in Sect. 4. Finally, Sect. 5 concludes the paper and sketches future work.

2 Morphology of Nucleus

Cancerous nuclei possesses many changes compared to healthy nuclei. The size, shape, margin, chromatin pattern, nucleoli and perinucleolar space alterations are a group of these changes that occur during the carcinogenesis. Image analysis techniques along with morphometry are an aid in the characterization of the size and shape of organelles and structures. The categorization of different carcinomas is based on morphometric features [1].

According to Thiran and Macq, the malignancy of a tissue is defined based on four factors: nucleocytoplasmic ratio, anisonucleosis, nuclear deformity and hyperchromasia. The nucleocytoplasmic ratio is calculated by counting all nonwhite pixels from an image to separate the cytoplasmic area. Anisonucleosis and nuclear deformity are calculated by finding the center of each nucleus and defining the size of the nucleus by calculating the distance from the edge of the nucleus to the center. The hyperchromasia is defined by using a Top-hat filter to define the mean value of the spikes over the nuclear area [2].

3 Adaptive Bottom-Hat Filter

Bottom-hat filtering is defined as the difference between the original image and its closing. As top-hat filter preserves the sharp peaks or brightest points in an image, bottom-hat filter does the opposite and preserves sharp bottoms also known as the darkest regions. The nuclei in our images are small dark spots which obliges us to use a method of extraction that considers this fact.

Bottom-hat filtering consists of dilation (see Eq. 1) followed by erosion (see Eq. 2): the dilation adds pixels to the objects' boundaries while erosion removes them. The number of pixels modified in these operations depend on the size and shape of used structure element [3].

$$A \oplus \mathrm{B} = \{\mathrm{x}{:}\hat{B}_x \cap \mathrm{A} \neq \emptyset\} \tag{1}$$

$$A \ominus \mathrm{B} = \{\mathrm{x}{:}B_x \subseteq \mathrm{A}\} \tag{2}$$

Where A is the object and B is the structure element.

Thresholding an image aims to classify each pixel either as dark or light pixel. Performing the thresholding adaptively results in taking automatically the variations in illumination into account. Bradley and Roth suggest using a method that computes a different threshold value for each pixel in an integral image. The idea is that an average for a x a window of pixels is calculated centered around each pixel in two passes: in the first pass the computation is performed in linear time by using the integral image and in the second pass the calculation is performed by using the integral image for each pixel in constant time after which a comparison is performed [4]. In our solution, we will utilize mainly the same idea.

The adaptivity of our method lies in the definition of the used structure element. The structure element values are generated by first dividing an image into windows of certain size after which the calculation is processed inside each window. The value that defines the structure element is average graylevel value (see Eq. 3).

$$A = B/(m * n) \tag{3}$$

The value A after that used as the structure element for adaptive bottom-hat filter and applied on an image as presented. The used image (Fig. 1) is an image of breast tissue presenting heavy nuclear area. To compare, the basic bottom-hat filter in Matlab is not capable of extracting nuclear area from other areas effectively as seen in Fig. 2. The nuclei in the image is somewhat visible but the background is too strongly highlighted to make a full separation between nuclei and other cell areas.

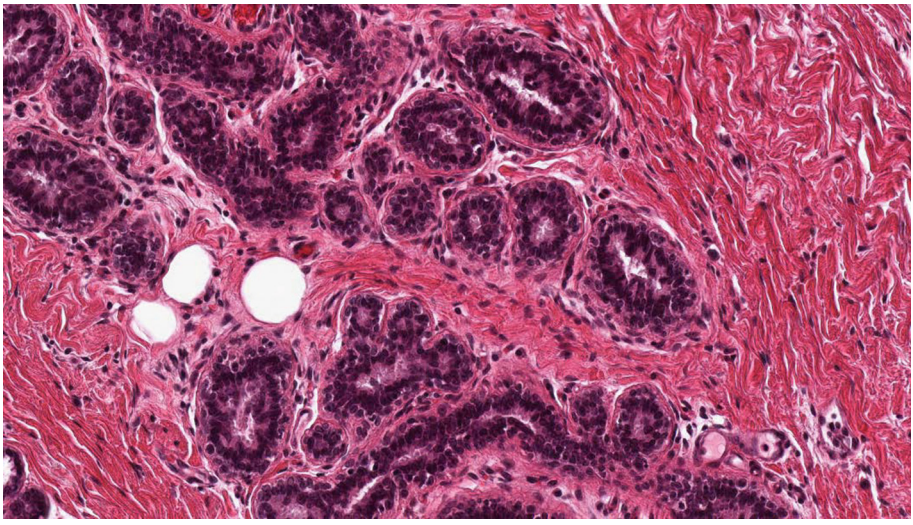


Fig. 1. Original microscopic image of breast tissue

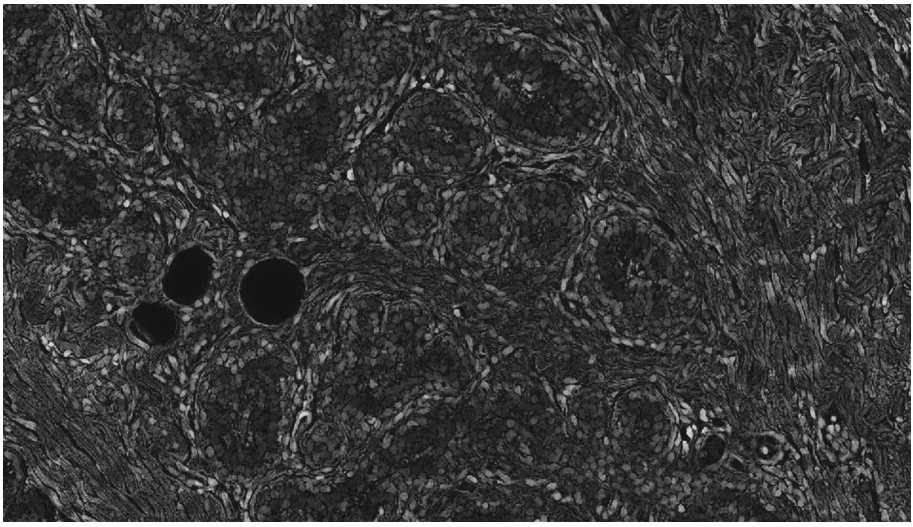


Fig. 2. Resulting image of basic bottom-hat filter with a structure element of disk with radius 7

The application of adaptive bottom-hat filter results in an image (Fig. 3) where the pixels representing nuclei are classified as dark and other areas bright. This serves as an excellent starting point for further analysis of nuclei malignancy based on the four criteria presented in an earlier chapter. The adaptiveness of the method results in a higher extraction rate of the nuclear area from other areas. The nuclei is much more easily distinguished even by human eye as it is the only area in the image marked with completely white pixels.

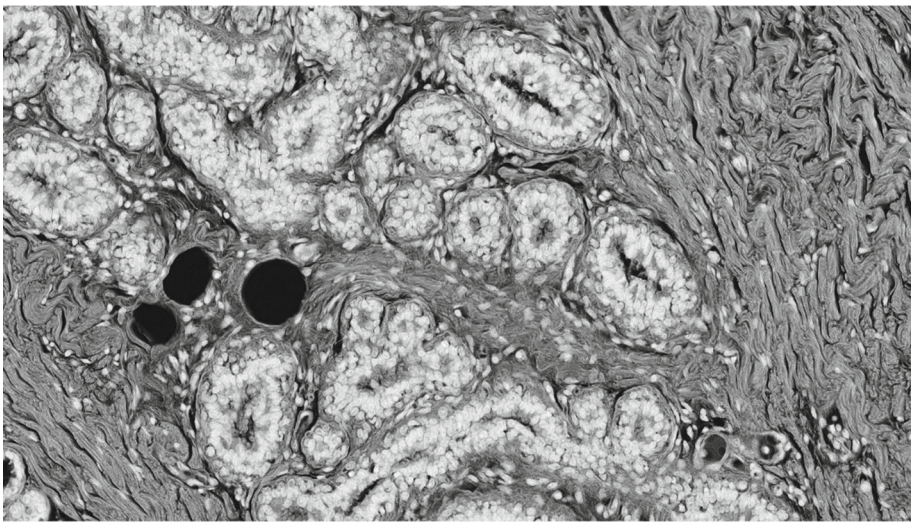


Fig. 3. Resulting image of adaptive bottom-hat filter

4 Adaptiveness Combined with Color Channel Manipulation

We have studied the idea of color channel manipulation in our previous work [5] and utilizing this idea we combined the two approaches. To enhance the histopathologic image, the RGB color channel manipulation is applied according to our formula (see Eq. 4):

$$R-(G+B)/2 \tag{4}$$

This work is based on Ali et al. [6] that suggested the use of a reference image formed from averaging the R and G channels after which the average was subtracted from the B channel. Resulting image (Fig. 4) for this computation demonstrates the usefulness of this method as a preprocessing step before the extraction of nuclei.

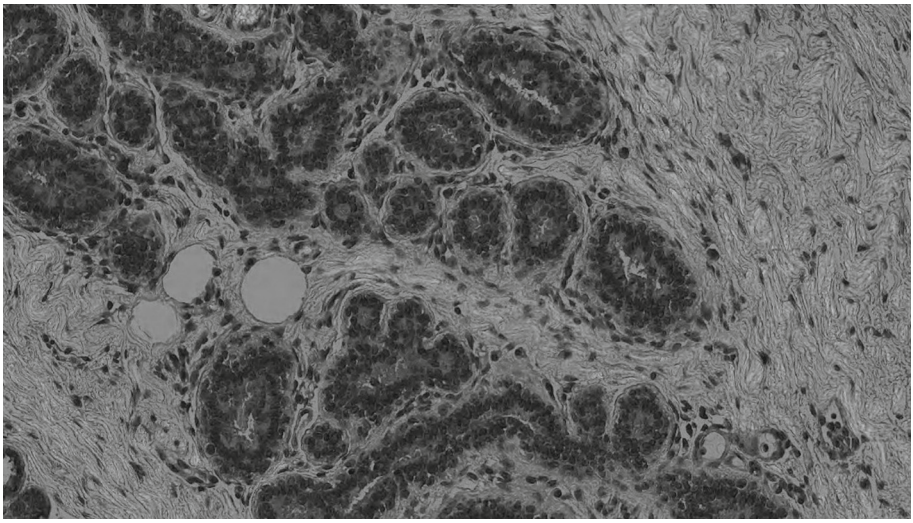


Fig. 4. Resulting difference image of the color channel manipulation on original image of the breast tissue on Fig. 1

The resulting image presents a clear contrast between the nuclear area and other areas as the nuclei is set to black and other areas gray. After the color channel manipulation step, the adaptive bottom-hat filter is applied and the resulting image (Fig. 5) shows the final result of nuclei extraction.

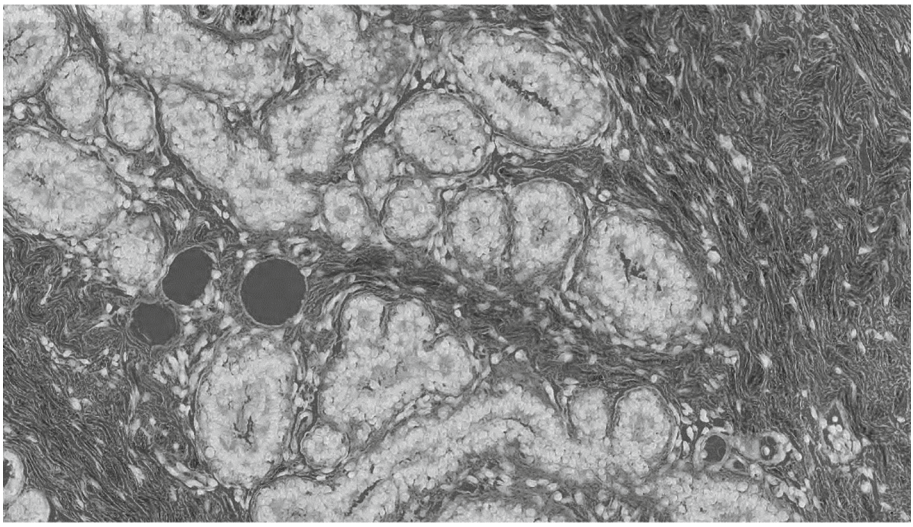


Fig. 5. Result of the application of adaptive bottom-hat filtering after color channel manipulation

The nuclei is highlighted as white and other areas gray while all the black pixels have vanished. To compare the results of our methods, another image (Fig. 6) was processed the same way as previous case.

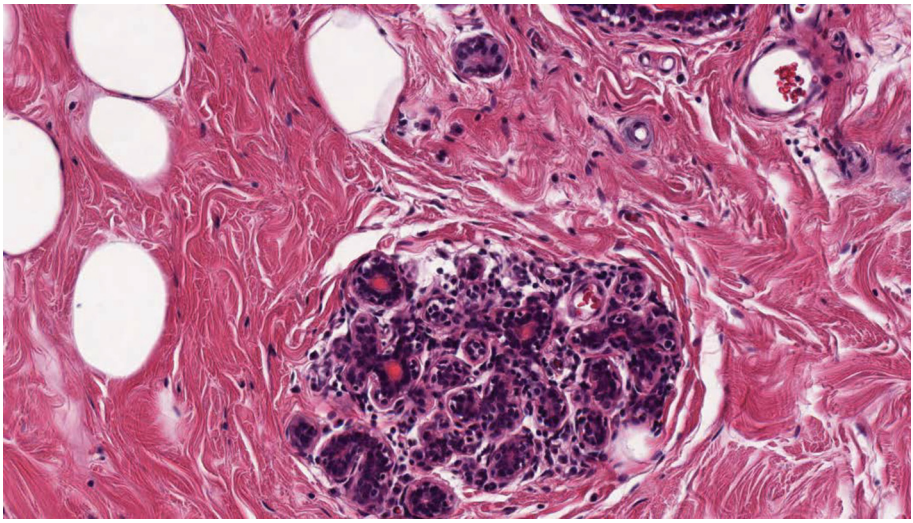


Fig. 6. Original image of breast tissue

The resulting image (Fig. 7) of adaptive bottom-hat filter without the color channel manipulation presents the nuclei as white, stroma as grey and fat as black pixels. In the combination of both adaptive bottom-hat filter and color channel manipulation, the

pixels representing nuclei are the only pixels highlighted as white and other cells areas as different shades of grey which is the goal of our nuclei extraction.

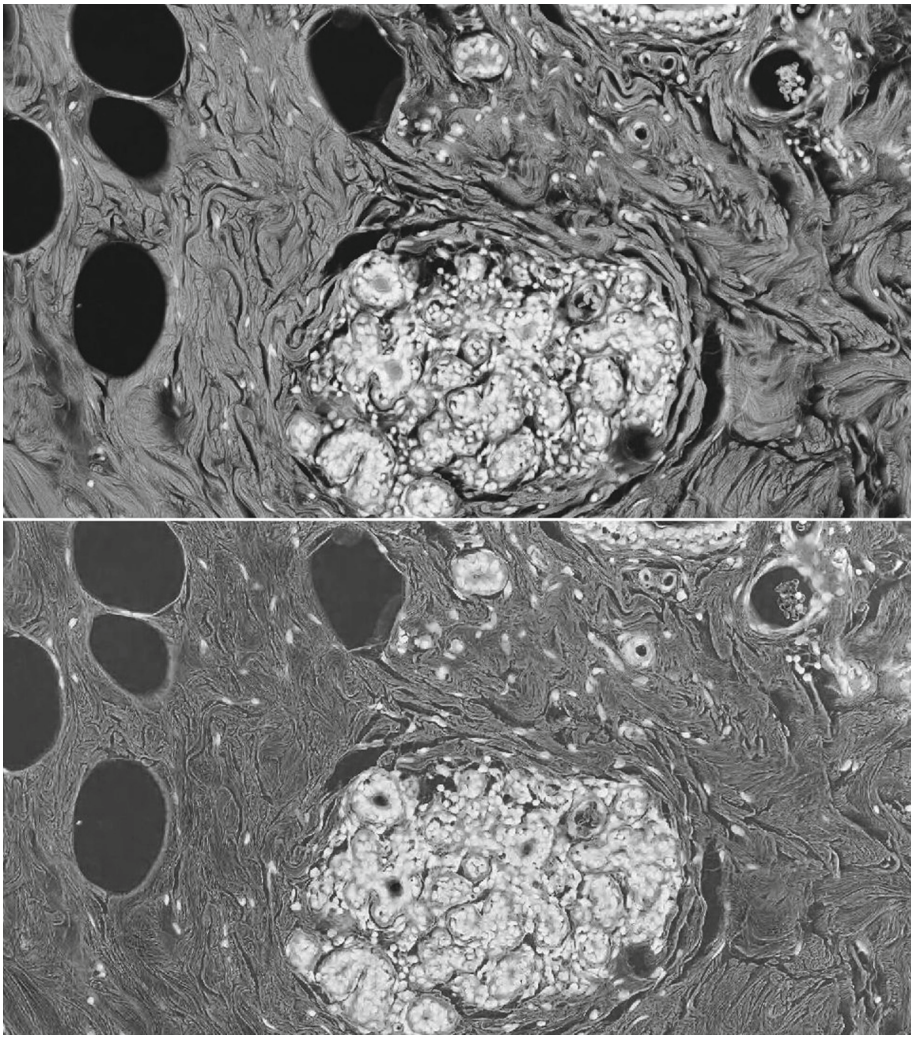


Fig. 7. Result of the application of adaptive bottom-hat filtering before (above) and after color channel manipulation (down)

To compare the results of our method to the method of Bradley, we use the same image as in our first case and show the resulting images of Bradley's method itself before and after our color channel manipulation in Fig. 8. As can be seen, also the basic Bradley method benefits from the preprocessing step. The original Bradley image has some unnecessary tissue structures still present in the image whereas the manipulated image

has succeeded in disposing them, which in turn bares the nuclei more explicitly for further analysis.

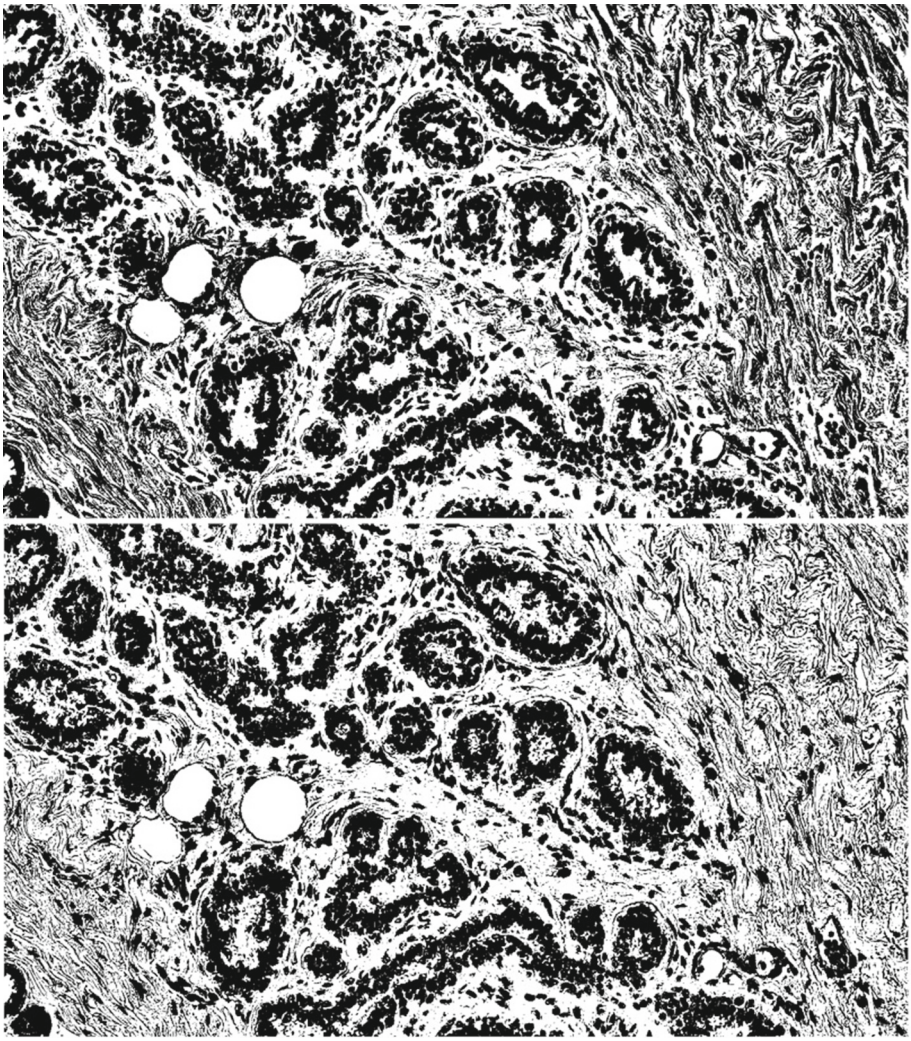


Fig. 8. Basic method of Bradley and Roth (above) versus the Bradley method computed after our RGB manipulation (down)

5 Conclusion and Future Work

The basic nuclei morphology was presented to give out an understanding of how malign nuclear area compares to healthy area and how the analysis of cancerousness proceeds. As the extraction and analysis of nuclei is in the spotlight of cancer research, new faster

and more accurate algorithms are needed. The parallelization of methods would be also something that needs more attention as the size of the images produced by different imaging modalities are larger than ever and containing more information than what was before. Our methods are all designed keeping a clear focus on this fact and therefore can be parallelized.

The idea of adaptive thresholding in bottom-hat Filtering in the use of breast cancer diagnostics provides a new approach to the problem pathologists' encounter. The basic bottom-hat filtering is not capable of filtering the breast cancer image in such a way that is needed. Meaning basically the fact that the produced image does not provide a sufficient enough contrast between nuclei and stroma, blood vessels and other tissue components. Therefore, the adaptiveness of the method provides a solution for the addressing of the variation in illumination and providing us the images needed for further nuclei malignancy analysis.

The use of color channel manipulation as a preprocessing method before adaptive bottom-hat filtering results in the elimination of such tissue details that are not needed in the feature extraction. The outcome of this process provides an excellent base for further analysis for the pathologist, as our method was originally aimed to be a helping hand in the automation process of cancer analysis.

As a conclusion for the results of our two-phased process of nuclei extraction, the resulting images provide us a base where the nuclei, that is needed in the analysis of cancer stage and development, is highlighted from the background. Background in this case means all other components and structures of cells. The benefit of this process is high to the pathologists facing numerous breast cancer images, as is provides a clear sight onto the nuclei. Also the automation of the whole cancer analysis process is easier after this step.

In our future work, we will continue improving our adaptive method and introduce the skewness factor to enable more detailed analysis of the shapes of the top- or bottom-hat filtered tissue elements. In addition, we will combine this method with our previous work [7, 8] with DTOCS and 3D-DTOCS.

Acknowledgement. This research work was supported by ARTEMIS Joint Undertaking under Grant Agreement no 621439 (ALMARVI). The utilized histopathological breast cancer images are acquired from BioMediTech, University of Tampere.

References

1. Dey, P.: Cancer nucleus: morphology and beyond. Diagn. Cytopathol. **38**, 382–390 (2010)
2. Thiran, J.P., Macq, B.: Morphological feature extraction for the classification of digital images of cancerous tissues. IEEE Trans. Biomed. Eng. **43**, 1011–1020 (1996)
3. Mathworks Documentation, Morphological dilation and erosion, 17 August 2016. http://se.mathworks.com/help/images/morphological-dilation-and-erosion.html
4. Bradley, D., Roth, G.: Adaptive thresholding using the integral image. J. Graph. **12**, 13–21 (2007)

5. Ikonen, T., Pöllänen, I., Braithwaite, B., Haataja, K., Toivanen, P., Tolonen, T., Isola, J.: Morphological extraction of cancerous nucleus in the diagnostics of breast cancer. In: 15th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 148–153 (2015)
6. Ali, H.R., Irwin, M., Morris, L., Dawson, S.-J., Blows, F.M., Provenzano, E., Mahler-Araujo, B., Pharoah, P.D., Walton, N.A., Brenton, J.D., Caldas, C.: Astronomical algorithms for automated analysis of tissue protein expression in breast cancer. Br. J. Cancer **108**, 602–612 (2013)
7. Ikonen, T., Haataja, K., Toivanen, P.: 3D imaging of human brain in the diagnostics of insomnia and depression: a comparative analysis, a novel insomnia/depression diagnostics approach, and lessons learned. In: Proceedings of the 2013 International Conference on Intelligent Systems Design and Applications (ISDA), pp. 230–235 (2013)
8. Ikonen, T., Niska, H., Braithwaite, B., Pöllänen, I., Haataja, K., Toivanen, P., Isola, J., Tolonen, T.: Computer-assisted image analysis of histopathological breast cancer images using step-DTOCS, In: 14th International Conference on Hybrid Intelligent Systems (HIS). pp. 187–192 (2014)

Self-Organizing Maps and Fuzzy C-Means Algorithms on Gait Analysis Based on Inertial Sensors Data

Rafael Caldas[1(✉)], Yabing Hu[1], Fernando Buarque de Lima Neto[2], and Bernd Markert[1]

[1] Institute of General Mechanics, RWTH Aachen University, Aachen, NRW 52062, Germany
rafael.caldas@iam.rwth-aachen.de
[2] School of Engineering, University of Pernambuco, 50.720-001, Recife, PE, Brazil
fbln@ecomp.poli.br

Abstract. Human gait corresponds to the physiological way of locomotion, which can be affected by several injuries. Thus, gait analysis plays an important role in observing kinematic and kinetic parameters of the joints involved with such movement pattern. Due to the complexity of such analysis, this paper explores the performance of two adaptive methods, Fuzzy c-means (FCM) and Self-organizing maps (SOM), to simplify the interpretation of gait data, provided by a secondary dataset of 90 subjects, subdivided into six groups. Based on inertial measurement units (IMU) data, two kinematic features, average cycle time and cadence, were used as inputs to the adaptive algorithms. Considering the similarities among the subjects of such database, our experiments show that FCM presented a better performance than SOM. Despite the misplacement of subjects into unexpected clusters, this outcome implies that FCM is rather sensitive to slight differences in gait analysis. Nonetheless, further trials with the aforementioned methods are necessary, since more gait parameters and a greater sample could reveal an undercover variation within the proper walking pattern.

Keywords: Computational intelligence · Self-organizing maps algorithm · Fuzzy logic · Gait analysis · Inertial measurement unit

1 Introduction

Human gait, or walking, refers to a person's way of locomotion [6]. This movement pattern has a cyclical nature with sequential movements of both legs and is described as a series of states with transitions, being divided into stance and swing phases [16]. Such pattern is usually affected by many injuries, thus an efficient gait assessment plays an important role in rehabilitation.

Gait analysis is the systematic study of human walking pattern [9]. Such analysis involves the assessment of several biomechanical features, which allows

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_20

physicians to identify causes of various injuries, enabling the observation of treatment effectiveness [1,15,19]. Conventional gait analysis uses laboratory-based systems and questionnaires, but these methods present several disadvantages [24]. The optoelectronic system, which corresponds to the gold standard method, is rather complicated and expensive, requiring a large space to perform the assessment [18]. On the other hand, the results obtained with specific questionnaires lack sufficient accuracy, implying a subjective conclusion [5].

In this context, over the past years, inertial measurement units (IMU) have been widely used in gait analysis, because of their easy application and low cost [8,11]. An IMU is an electronic device that consists of accelerometers for linear acceleration and gyroscopes for angular velocity measurements in three axes [20]. IMU could dynamically acquire the signals of linear acceleration and angular velocity during the whole walking process and those quantitative signals clearly reflect motion information of human gait [4].

After acquisition and preprocessing, the sensor data is still represented in a complex way, that might difficult healthcare professionals' interpretation of the results. Considering this, adaptive algorithms, such as Self-organizing maps (SOM) and Fuzzy c-means (FCM), are efficient methods for classifying data. SOM is particularly useful for analyzing data due to its ability of projecting high-dimensional data into a simpler low-dimensional representation [13]. On this nonlinear projection, each individual is mapped to one point, where the distances among the subjects reflect their similarities [21].

Additionally, Fuzzy logic is considered an advanced method for identifying features of periodic signals [22]. Such approach uses approximate rather than exact measures, represented by the membership function. This function denotes the degree of membership, characterizing a set that extends continuously from 0 to 1 [7]. As an adaptive tool, one can use Fuzzy logic for data analysis, considering its likely smooth and continuous detection of gait phases [14]. This paper is organized so that the performance of FCM and SOM algorithms through clustering of human gait data acquired by inertial sensors can be established. Section 2 presents the methods used for data acquisition and processing. In Sect. 3, the results are commented upon.

2 Methods

Data Acquisition. In this study, we used the OU-ISIR gait database, which was established to assist research efforts in general area of developing and testing algorithms for human identification based on gait data [17]. This dataset uses four sensors: (i) 3 IMUs, each one includes a triaxial accelerometer and a triaxial gyroscope; (ii) a triaxial KXTF9 accelerometer embedded in a smartphone ME860. They set the IMU dynamic ranges at $\pm$4g for accelerometers and $\pm$500 deg/s for gyroscopes to capture human gait signal, all the four sensors worked at 100 Hz.

Sensors were attached to a belt, pointing their x-, y- and z-axes to subject's right, upwards and backward, respectively. The centered IMU and smartphone

are located at the center back waist of subjects; the left and right IMU are located at the left and right waist. During flat ground walking, 6D signal sequences were captured with each IMU. The subjects were divided into 6 groups according to their ages: Group A includes subjects under 10 years old; Group B, 10–19 years old; Group C, 20–29 years old; Group D, 30–39 years old; Group E, 40–49 years old and Group F, over 50 years old.

Gait Temporal Features Extraction. Based on previous works and preliminary experiments, Average cycle time (ACT) and Cadence were the kinematic features extracted [3]. The number of cycles or steps per second represents the cadence, which indicates an relevant parameter regarding the subjects' performance [10]. The minimum value of each valley was set as the start of a cycle as well as the end of the previous cycle. Thus, the average cycle time is easily obtained, providing possibly information on the symmetry of subject.

Data Processing. Fuzzy c-means is a clustering method that allows each data point to belong to multiple clusters with varying degrees of membership. FCM is based on the minimization of the following objective function [2]:

$$J_m = \sum_{i=1}^{D} \sum_{j=1}^{N} \mu_{ij}^m \|x_i - c_j\|^2 \tag{1}$$

herein, D is the number of data points; N is the number of clusters; m is the fuzzy partition matrix exponent for controlling the degree of fuzzy overlap, with $m > 1$. This overlap refers to the boundaries between clusters and denotes the number of data points that have significant membership in more than one cluster; x_i is the *ith* data point; c_j is the center of the *jth* cluster; μ_{ij} is the degree of membership of x_i in the *jth* cluster. For a given data point, x_i, the sum of the membership values for all clusters is 1.

FCM performs the following steps during clustering: (i) Randomly initialize the cluster membership values, μ_{ij}; (ii) Calculate the cluster centers, as shown in formula (2); (iii) Update μ_{ij} according to the following, as shown in formula (3); (iv) Calculate the objective function, J_m.; (v) Repeat steps (ii)–(iv) until J_m improves by less than a specified minimum threshold or until after a specified maximum number of iterations:

$$c_j = \frac{\sum_{i=1}^{D} \mu_{ij}^m x_i}{\sum_{i=1}^{D} \mu_{ij}^m} \tag{2}$$

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{N} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right) \frac{2}{m-1}} \tag{3}$$

Self-Organizing Maps (SOM) classify input vectors according to the pairing in the input space. At the first level, a 2D feature map is formed, where the number of output neurons is significantly larger than the desired number of clusters. Then, at the second level the output neurons, representing the subjects, are

clustered such that the neurons on the map are divided into as many different regions as the desired number of clusters. During the training process, each input data point can be assigned into a cluster according to the best match unit (BMU), or its nearest output neuron [23].

In the original proposition of SOM, the Euclidean distance corresponds to the similarity criterion and the winner neuron is the one with the smallest distance [12]. Considering this process, the self-organization aims to minimize the distance, adjusting the BMU weights and its neighborhood toward the input vector using the following rule:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)h_{ci}(t)[x(t) - w_{ij}(t)] \tag{4}$$

herein, w_{ij} is the weight j of neuron i, $\alpha(t)$ is the learning rate and $h_{ci}(t)$ is the neighborhood radius function centered on winner neuron c with respect to neuron i. Typically, both learning rate and neighborhood radius are functions, which decrease with time, in order to facilitate system convergence.

Experiments Setup. Initially, we set the features of different axes as inputs to the FCM algorithm, intending to explore the potential differences of the results between them after clustering. In the first experiment, two groups, A and D, were chosen. Each group comprises the kinematic features (ACT and cadence) of 30 randomly selected subjects, acquired by the three IMUs. We disregard the smartphone information because it is redundant data due to the same position of the central sensor. In the second part of FCM experiments, 15 subjects of each of the six groups (A–F) were randomly selected and their kinematic features were used as inputs to the system.

Regarding the SOM, we intended to explore the influence on changing inherent variables of the algorithm, applied to the same dataset utilized in FCM trials. Based on a previous work from our group, the y-axis was also chosen as input for the SOM [3]. In the first experiment, Ordering-phase (OP) steps or the number of iterations were set as 500, 1000, and 2000. While the following parameters were fixed as following: OP learning rate = 0.9; Training phase (TP) learning rate = 0.02; TP Neighborhood distance = 1; and Map size = 20×20. In the second experiment, the OP steps were fixed in 500 iterations and the TP Neighborhood distances were set as 1, 10 and 20, while the other variables were not modified.

3 Results and Discussion

The performed experiments revealed the behavior of the two selected methods on grouping subjects due to their kinematic parameters. Figure 1(a), (b) and (c) show the results of Fuzzy c-means according to the inputs of x, y and z axes, respectively. The subjects are grouped according to their similarities on the graph, where the solid lines connecting the samples represent the distance between them. The clusters are clearly separated and there is no relevant difference between these three cases.

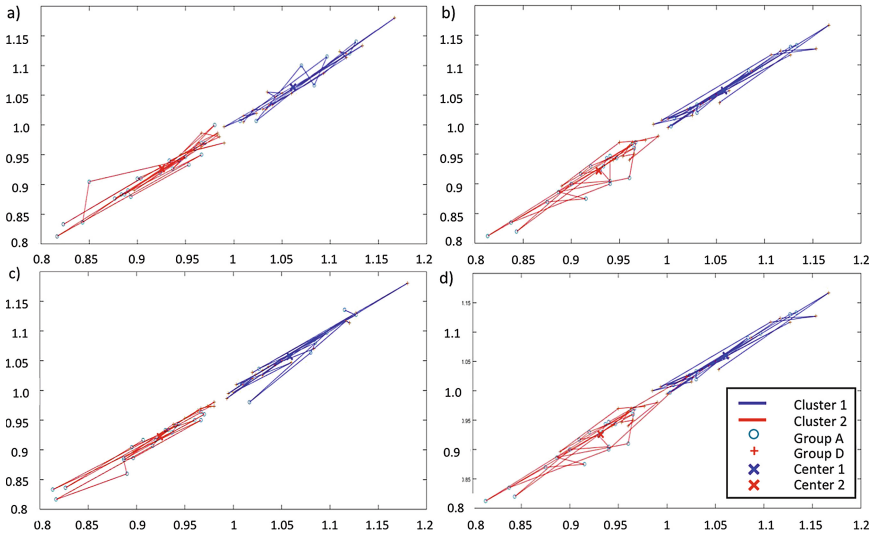


Fig. 1. Fuzzy c-means clustering of Groups A (circle) and D (plus): (a) Features of x-axis as inputs; (b) Features of y-axis as inputs; (c) Features of z-axis as inputs; (d) Features of the three combined axes as inputs.

As shown in Fig. 1(d), FCM outcomes, when using all axes as input, present only one subject from Group A clustered on the other group. This might indicate a small difference between such subject and Group D. Nonetheless, such graph does not present significant difference when compared to the outputs of previous experiments, considering the placement of most subjects of the Group A at left lower part of the graph. The application of more than one axis, thus, does not seem to be relevant to the algorithm. Thereupon, we used features of only one axis as input in the following experiments, conveniently y-axis was chosen.

In the second experiment, 15 subjects were randomly chosen from each of the 6 groups (A–F) and grouped into 6 clusters (Fig. 2). Since some of the clusters do not present 15 elements, the outcome seems not to corroborate with our hypothesis that the algorithm could differentiate the subjects, despite their slightly different kinematic features. For instance, only 4 individuals form cluster 6, what implies an unusual gait pattern of the misplaced individuals. In this case, either such subjects presented an unexpected performance to their age group or FCM was not sensitive enough to recognize the differences. However, we could not ensure the reason since there is no further anthropometric information, except for their age groups.

Based on the clustering outcomes, we assume that the pointed out subjects might present an abnormal gait pattern (Fig. 2). The left lower one, which belongs to Group E, lies among the Group A, while the upper right subject, from Group B, was placed distant from the other subjects within cluster 6. Aside these two highlighted individuals, the subjects were generally clustered as expected,

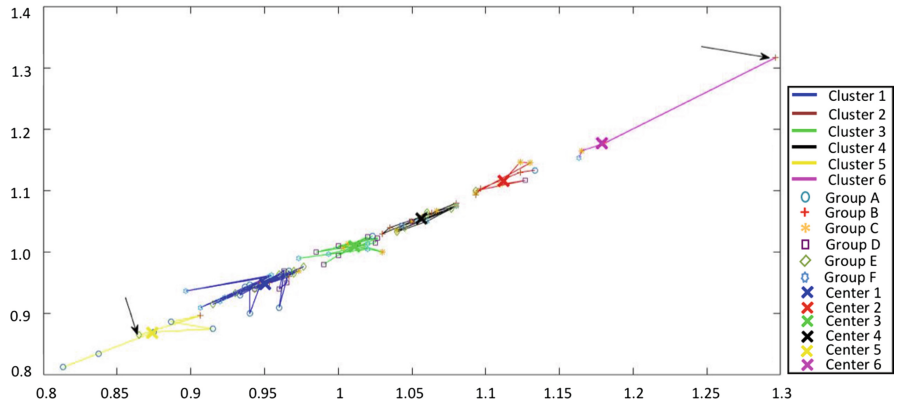


Fig. 2. FCM clustering of Group A to F, kinematic features of y-axis set as inputs. Black arrows indicate two individuals commented upon later.

the younger subjects from Group A on the left part of the graph, in contrast to the older ones placed on the top right part of the graph.

Considering the experiments with SOM, 15 subjects were selected for each age group (A–F) and kinematic features acquired by y-axis set as input. The graphs in Fig. 3 represent SOM neighborhood weight distances, which indicate the distances between neighboring neurons. The blue hexagons represent such neurons with red lines connecting neighboring ones. The colors in the areas containing the red lines indicate the distances between neurons. For instance, darker and lighter colors represent larger and smaller distances, respectively. By varying the number of iterations, one can notice that the algorithm groups the sample only on small areas of the top corners. Even though, these clusters include few subjects, while most of the subjects are not clustered very clearly. The algorithm spread them on the map even with the increased number of iterations (Fig. 3.c).

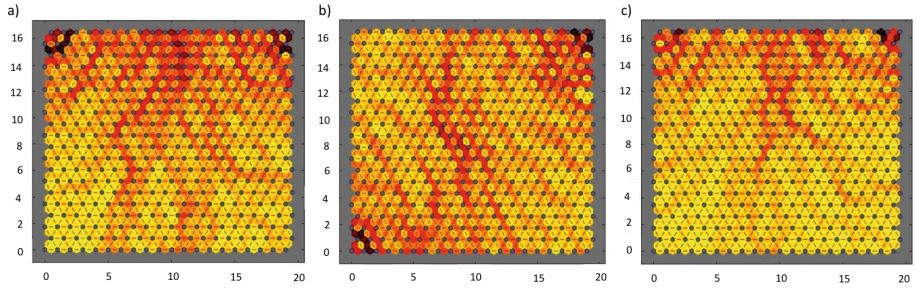


Fig. 3. SOM Neighbor weight distance of (a) 500, (b) 1000, and (c) 2000 Ordering-phase steps or iterations.

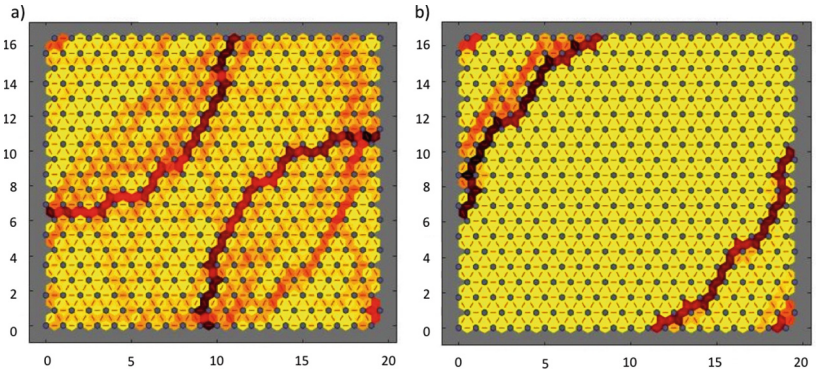


Fig. 4. SOM Neighbor weight distance of training phase - Neighborhood distances (a) 10 and (b) 20.

As shown in Fig. 4, the cluster becomes clearer and more compact with the TP neighborhood distance increasing, indicated by the dark red lines. The representation of sample hits of different TP neighborhood distance provides information about data points that are associated with each neuron (Fig. 5), representing the placement of subjects due to their own gait features. These graphs show straightforward two main clusters in the upper-left and lower-right corner, where subjects with similar features are placed. Despite individuals in the middle, who have less similar features, were mapped apart from the clusters. As the neighborhood distance increases, these points in the middle are placed closer to the two main clusters (Fig. 5b).

Previous research has proved that SOM could group gait data in the comparison of proper and pathological patterns [3]. FCM, nevertheless, seems to perform better when facing slight variations of healthy persons' gait pattern,

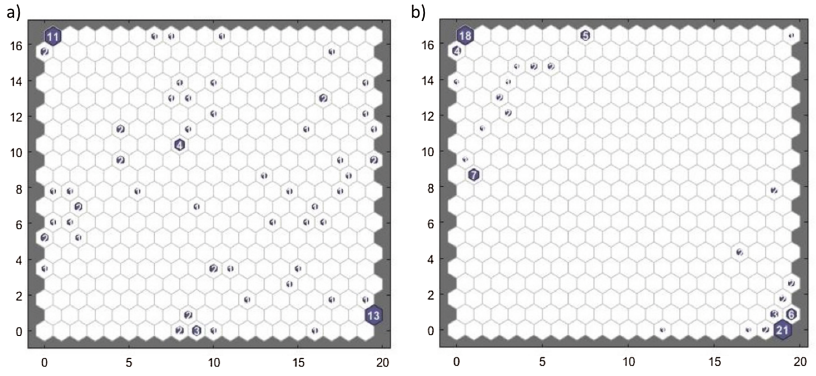


Fig. 5. SOM Sample hits of training phase - Neighborhood distances (a) 10 and (b) 20.

inherent to different age groups. Since such algorithm allows a subject to belong to different clusters, what might be the case of healthy subjects with similar gait features.

4 Conclusions

In the current work, we processed human gait data acquired by inertial sensors with two adaptive algorithms, Fuzzy c-means and Self-organizing maps. The results pointed out that FCM outperformed SOM in the task of grouping healthy subjects, divided into age-related groups. Subjects in this database were healthy, implying small variations in their gait patterns, which could not be easily distinguished. In this way, Fuzzy c-means presented better results, considering that a subject could belong to several clusters. SOM, on the other hand, has the advantage of not requiring the number of clusters in advance, interesting for when no further information about the dataset is available.

Therefore, further trials with the aforementioned methods are necessary, since more gait parameters and a greater sample could reveal an undercover variation within the proper walking pattern. Thus, we intend to obtain more walking features data, provided by a system with more IMUs, enabling a global gait analysis. Moreover, the sensors might provide less information about such pattern when placed on the waist than on the lower limbs.

Acknowledgment. This work was supported by the Brazilian National Council for Scientific and Technological Development (CNPQ 207445/2014-1) through a doctoral scholarship granted to R.C.

References

1. Alaqtash, M., Yu, H., Brower, R., Abdelgawad, A., Sarkodie-Gyan, T.: Application of wearable sensors for human gait analysis using fuzzy computational algorithm. Eng. Appl. Artif. Intell. **24**(6), 1018–1025 (2011)
2. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell (1981)
3. Caldas, R., Rativa, D., Lima Neto, F.B.: An application of self-organizing maps clustering to support kinematic gait analysis and evaluation of gait symmetry. J. Biomed. Health Inf. (Submitted for publication) (2016)
4. Chen, Y., Hu, W., Yang, Y., Hou, J., Wang, Z.: A method to calibrate installation orientation errors of inertial sensors for gait analysis. In: 2014 IEEE International Conference on Information and Automation (ICIA), pp. 598–603 (2014)
5. Clarke, J.E., Eccleston, C.: Assessing the quality of walking in adults with chronic pain: the development and preliminary psychometric evaluation of the bath assessment of walking inventory. Eur. J. Pain **13**(3), 305–311 (2009)
6. De Cann, B., Ross, A., Culp, M.: On clustering human gait patterns. In: 22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24–28, 2014, pp. 1794–1799 (2014)

7. Dote, Y.: Introduction to fuzzy logic. In: Proceedings of the 1995 IEEE IECON 21st International Conference on Industrial Electronics, Control, and Instrumentation, vol. 1, pp. 50–56 (1995)
8. Godfrey, A., Conway, R., Meagher, D., OLaighin, G.: Direct measurement of human movement by accelerometry. Med. Eng. Phys. **30**(10), 1364–1386 (2008)
9. Guha, A.R., Perera, A.M.: Calcaneal osteotomy in the treatment of adult acquired flatfoot deformity. Foot Ankle Clin. **17**(2), 247–258 (2012)
10. Hussain, S., Xie, S.Q., Jamwal, P.K.: Effect of cadence regulation on muscle activation patterns during robot-assisted gait: a dynamic simulation study. IEEE J. Biomed. Health Inf. **17**(2), 442–451 (2013)
11. Kavanagh, J.J., Morrison, S., James, D.A., Barrett, R.: Reliability of segmental accelerations measured using a new wireless gait analysis system. J. Biomech. **39**(15), 2863–2872 (2006)
12. Kohonen, T.: The self-organizing map. Proc. IEEE **78**(9), 1464–1480 (1990). doi:10.1109/5.58325
13. Kohonen, T., Schroeder, M.R., Huang, T.S. (eds.): Self-Organizing Maps, 3rd edn. Springer-Verlag New York Inc, Secaucus (2001)
14. Kong, K., Tomizuka, M.: Smooth and continuous human gait phase detection based on foot pressure patterns. In: IEEE International Conference on Robotics and Automation, ICRA 2008, pp. 3678–3683 (2008)
15. Köse, A., Cereatti, A., Della Croce, U.: Bilateral step length estimation using a single inertial measurement unit attached to the pelvis. J. Neuro Eng. Rehabil. **9**(9), 1–10 (2012)
16. Ma, H., Liao, W.: Human level walking gait modeling and analysis based on semi-markov process. In: 2014 IEEE International Conference on Robotics and Automation, ICRA 2014, Hong Kong, China, 31 May–7 June, 2014, pp. 240–245 (2014)
17. Ngo, T.T., Makihara, Y., Nagahara, H., Mukaigawa, Y., Yagi, Y.: The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication. Pattern Recogn. **47**(1), 228–237 (2014)
18. Pohl, M.B., Patel, C., Wiley, J.P., Ferber, R.: Gait biomechanics and hip muscular strength in patients with patellofemoral osteoarthritis. Gait Posture **37**(3), 440–444 (2013)
19. Senanayake, C., Senanayake, S.M.: A computational method for reliable gait event detection and abnormality detection for feedback in rehabilitation. Comput. Methods Biomech. Biomed. Eng. **14**(10) (2010)
20. Tessendorf, B., Gravenhorst, F., Arnrich, B., Tröster, G.: An IMU-based sensor network to continuously monitor rowing technique on the water. In: Proceedings of the Seventh International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 2011). IEEE press (2011)
21. Vesanto, J.: Som-based data visualization methods. Intell. Data Anal. **3**(2), 111–126 (1999)
22. Vicerra, R.R.P., David, K.K.A., dela Cruz, A.R., Roxas, E.A., Simbulan, K.B.C., Bandala, A.A., Dadios, E.P.: A multiple level mimo fuzzy logic based intelligence for multiple agent cooperative robot system. In: TENCON 2015–2015 IEEE Region 10 Conference, pp. 1–7 (2015)
23. Wu, S., Chow, T.W.S.: Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. Pattern Recogn. **37**(2), 175–188 (2004)
24. Yang, M., Zheng, H., Wang, H., McClean, S., Hall, J., Harris, N.: A machine learning approach to assessing gait patterns for complex regional pain syndrome. Med. Eng. Phys. **34**, 740–746 (2012)

Age, Gender, Race and Smile Prediction Based on Social Textual and Visual Data Analyzing

Onsa Lazzez(✉), Wael Ouarda, and Adel M. Alimi

REGIM-Lab.: Research Groups in Intelligent Machines, University of Sfax
National School of Engineers (ENIS), BP 1173, 3038 Sfax, Tunisia
{onsa.lazzez.tn,wael.ouarda,adel.alimi}@ieee.org

Abstract. In a relatively short period of time, we have observed the explosion of social network platforms which have acquired a prominent role in the people's daily life. In this context, each Facebook user can easily access and control the data shared on his/her corresponding profile. However, despite the large amount of information contained in Facebook user profile, this information is unstructured and not organized. In this paper, we propose a user ontology profiling in social networks by using a framework containing a Facebook application written in PHP, which accumulates all the equivalent shared data and stocks it in a not only SQL database Cassandra used for further analysis.

Keywords: Social networks · Facebook application · Not only SQL database · User profile data · User profile modeling

1 Introduction

In order to provide people with a good place to live, a new idea of a city is presented: Smart City. A Smart City is a novel-generation planning that allows enhancing performance of a city using modern information communication technologies. In a Smart City, various services and solutions are proposed to improve the daily lives of citizens such as enhancing quality service in various topics, according to their preferences. Indeed, social networks are very famed for several days and are the important communication channels in the internet today. Hence, social networks have a vital role in the daily lives of citizens. Allowing to [1], the users' number in social networks is currently increasing as in 2016, Facebook has 1.55 billion users, Instagram has 400 million users, and Twitter has 307 million users, etc.

Facebook is the most popular social network. It is a social tool used by citizens to share their hobbies, events, personal data, likes, pictures, etc. and to communicate well with friends. Each citizen can access Facebook by creating his/her own profile in which he/she provides a large amount of data. We remark that the current increase of the users' number in Facebook and the large amount of data posted on users profiles encourage researches to collect and analyze the user profile data to discover the interests and preferences of users and classify their priorities.

Therefore, the primary contribution of this paper is to provide a framework which allows to collect all data available on Facebook users' profiles by obtaining an authentification with an existing Facebook account, and storing it in a not only SQL

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_21

database. As a second contribution, data collected based on our novel framework are alimented to our new proposed user profile ontology.

This paper is structured as follows. Section 2 illustrates various related works about user profile modeling and social network data collection. Section 3 details our framework and its different modules. Section 4 shows the attained results. Finally, Sect. 5 recapitulates the paper and presents our outlines future works.

2 Materials and Methods

2.1 User Profile

According to [2, 3], the user profile is defined as information specifying each user such as personal data, preferences, interests, contact information, etc. In [4] the user profile is described as a series of structured information identifying the communication between user and system. In the domain of internet research engines [5], the user profile is an ontology utilized in order to have structured representation of user's interests. The definition proposed in [6], define the user profile as data when it used by systems wishing to respond to specific needs or as knowledge when it used by the behavioral analysis systems.

2.2 User Profile Modeling

Modeling user profile allows defining a structure for storing the data which identifies the user and indicating his/her centers of interests, his/her personal data and his/her activities. User profile modeling is a process composed of two phases: (i) user profile data collect; (ii) user profile representation [7].

- *Facebook application to collect user profile data*

As social networks databases are not publically accessible, the task of data collection from social networks attracted many researchers. In this sub-section, we present a brief review on data collection from Facebook. The work presented by [8] proposed "Netvizz Application". It is a data collection Facebook application for exporting data via the official Application Programming Interface from Facebook users' profiles. The authors have developed it in 2009. It allows collecting data from different sections from Facebook such as friendship networks, groups and pages. It required user to log into by a current Facebook account and to accept a list of permission to limit the access at her/his profile data." According to [9], the authors described "RUM extractor"; a data extraction tool which allows extracting different forms of data that are provided on Facebook pages, such as posts, comments on posts, shares of post, etc. using the official Application Programming Interface API. The work reported in [10], developed a Facebook application based on the open graph API which by an authentification token permits collecting all the information provided in user profile and storing it in MYSQL database.

Table 1 shows some Facebook applications to extract Facebook user profile data using Facebook API.

Table 1. Related work on Facebook user profile data extraction using Facebook API: "Application Programming Interface"

Application	Environement	Data source
"Netvizz Application"	Php/Mysql	Personal network, groups, pages
"Rum extractor"	Php/Rum server	Facebook pages
"Facebook event collector"	Php/Mysql	Information posted on the user profile

- *User profile representation*

Subsequently, we achieve the representation step of user profile. In the literature we can find three forms of representation: vector, semantic and multidimensional representation [3, 7]. The vector user profile representation is founded on the vector model suggested via Salton in 1971 [3, 11]. The user profile is identified by a vectors described in a space terms. The authors in [12] used the vector model to represent the user preferences on different items. The semantic user profile representation is based on the hierarchy of concepts or ontology. In reference [13], the authors presented a semantic user profile model using the Resource Description Framework (RDF) graph which connects entities and topics that interest the user. In [14], the authors presented an ontology based on Web Ontology Language (OWL) which model user profile concepts. The multidimensional user profile representation consists of captures and organizes all information specifying user profile. All form of information denotes a dimension in the model [15]. The authors in [16] defined a multidimensional user profile model which contains Identity data, Activity data and Safety data. The work in [17] proposed a user profile model composed on several dimensions such as Profile information, Workplace information, Academic information and Contact information. Table 2 shows the advantages and disadvantages of different approaches on user profile representation.

Table 2. User profile representation technique

Representation technique	Advantages	Disadvantages
Vector representation	Takes into accounts the different interest of users	Data unstructured
Semantic representation	Semantic relation between the interest of users profiles	Difficulty linked by the diversity and heterogeneity of users' profiles information
Multidimensional representation	Better implementation	Ambiguity in interpreting the roles of each dimension

In this paper, we chose to combine the semantic and multidimensional representation to represent data collected from Facebook user profile.

3 Realized Work

Social networks encouraged people to publish and share contents with their friends. Although such information is helpful for the personalization of the user profile, services providers do not exploit it. Note that, in our work, we focus on the social network Facebook due to its large use by people of all ages and backgrounds. Facebook is supportive for developing and keeping relations and has a sort of optimistic psychological results [18]. Figure 1 illustrates our proposed framework.

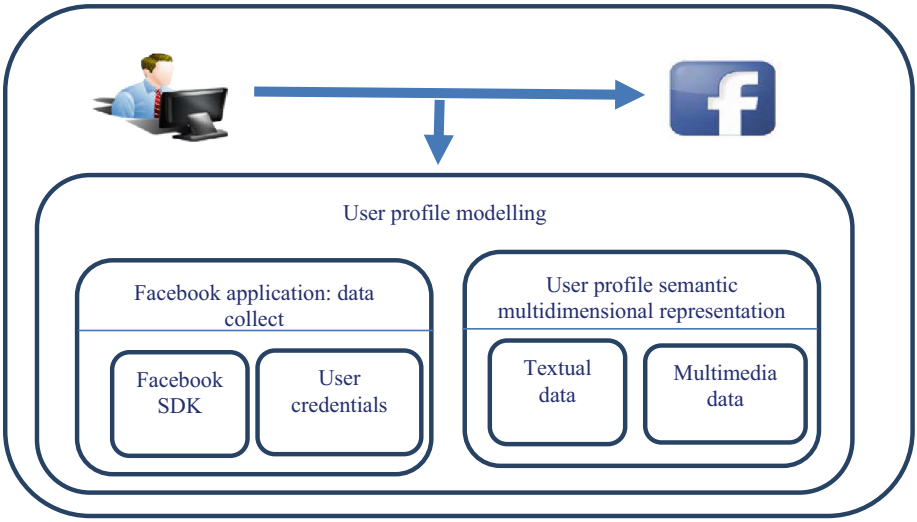


Fig. 1. Proposed Framework architecture

The user profile modeling is a structure that aims modeling and storing the user data [19].

3.1 User Profile Data Collect

Data on Facebook can be very diverse, unstructured and even unsuitable. Therefore, Facebook users voluntarily deliver every moment a variety of data on their profiles. Indeed, Multimedia data sharing in Facebook becomes very important and more popular [20]. Huge volume of multimedia data such as pictures, audios and videos are being generated and shared on Facebook daily. Our Facebook application was developed to facilitate data extraction from Facebook. It requires prior authentification. It respects the privacy of users, since we only collect data that users accept to be accessed using a list of required permissions.

3.2 User Profile Representation

To represent the user profile data, we combine semantic and multidimensional model. Firstly, we use multidimensional model by the event that the user profile is considered via different dimensions of information. Moreover, this structure organized along two dimensions: (i) "Textual data" which contains all textual data provided by the user; (ii) "Multimedia data" such as pictures collect, video collect. Secondly, we use semantic model due to the heterogeneity and diversity of user profile information and to provide semantic relations between users profiles dimensions defined by multidimensional model. Employing ontologies to model user profile has been suggested by diverse applications in the field of information systems in general and personalized web system in particular [23]. According to [24], using a multidimensional ontology allows developing ontology as a set of modules that can be amended, reduced and conserved by their developers. An important advantage of using multidimensional ontology is the expediency of keeping each ontology and the feasibility of switching parts of the ontology [25].

In this paper, we describe a semantic multidimensional ontology model. Figure 2 shows two dimensions: user profile textual data and user profile multimedia data. Our ontology based on user specifications was designed by the standard Web Ontology Language (OWL) and developed by the Protégé ontology editor.

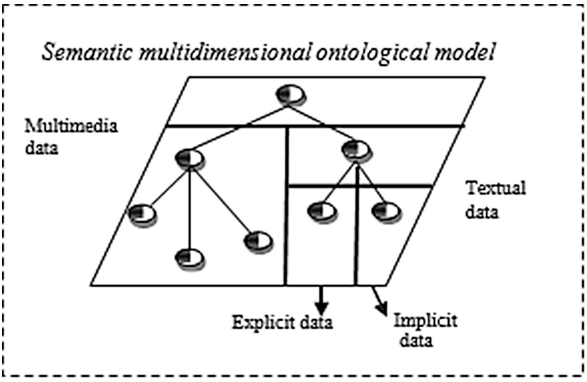


Fig. 2. Our Ontology architecture

To develop our user profile ontology, we start with describing the user profile by identifying the corresponding ontological classes as illustrated in Table 3. The class "REGIM-User" defines the central concept within the ontology. Within this, there are four key user classes, which are illustrated in Figs. 3, 4, 5 and 6.

3.3 Not only SQL Database Specification

To analyze Facebook profiles data, it is important to store it on a database for a future analysis. Due to the huge amount of data provided by the user profiles, it is wise to

Table 3. User profile ontology classes and sub classes

Class name	Sub classes name
Regim_User	Dynamic_data
Dynamic_data	Textual_data, Multimedia_data
Textual_data	Explicit_data, implicit_data
Explicit_data	Explcit_feedback, Personal_data
Implicit_data	Implicit_feedback
Personal_data	Employement, Education, Family_situation, Language, Religious_belief
Explicit_feedback	Likes, Shares
Likes	Favorite_book, Favorite_movies, Favorite_music, Favorite_sport_team, Favorite_TV, Favorite_restaurant
Implicit_feedback	Comments, Status
Multimedia data	Pictures collect, Video collect, vocal collect
Picture collect	Cover picture, Profile picture, Shared picture
Video collect	Shared video
Vocal collect	Shared vocal

register these data in a database that provides high accessibility and scalability as Not Only SQL database. Actually, relational databases work poorly in availability and reliability while processing large amounts of data spread out across many servers [23]. Our database named Social@Regim database, was created using the Cassandra Query Language (CQL). According to [26], CQL is the primary language for communicating with Cassandra database.

3.4 Social RegimLab Database

Each user can connect at our Facebook application via a website page named "Social@RegimLab" with the domain-name: "Regim_lab.ddns.net", illustrated in "Fig. 7". For an evaluation surpose, we collectes about 50 Facebook and Google+ users' accounts. Figure 8 illustrates our constructed database by some samples of the users' profile pictures collected.



Fig. 3. The basic concepts of ontology

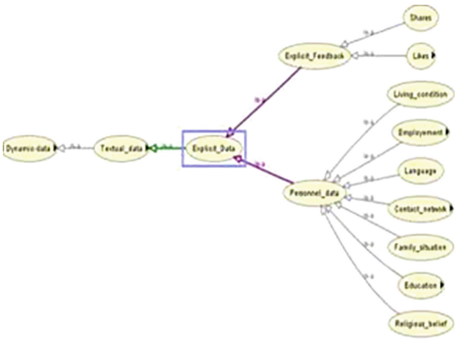


Fig. 4. The explicit textual concepts of ontology

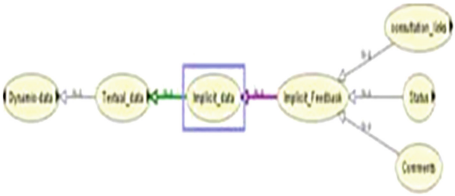


Fig. 5. The implicit textual concepts of ontology



Fig. 6. The multimedia concepts of ontology



Fig. 7. Website page "Regim_Lab.ddns.net

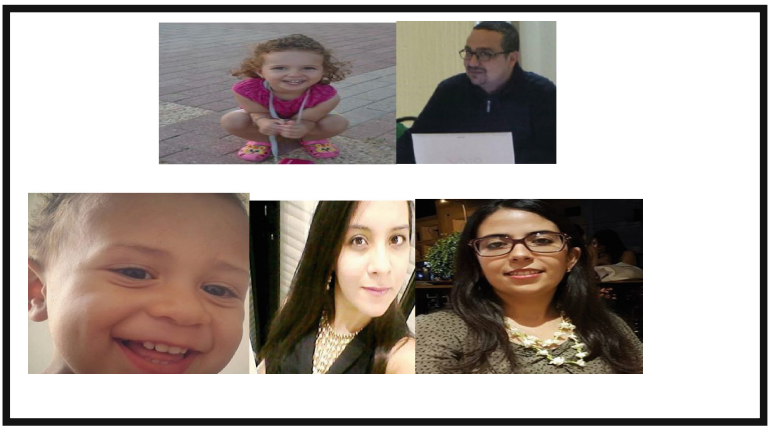


Fig. 8. Some profile pictures collected to construct our Social RegimLab database

3.5 Experimental Results

From the user' profile picture, we estimates gender, age, race and smiling as given in Fig. 9.

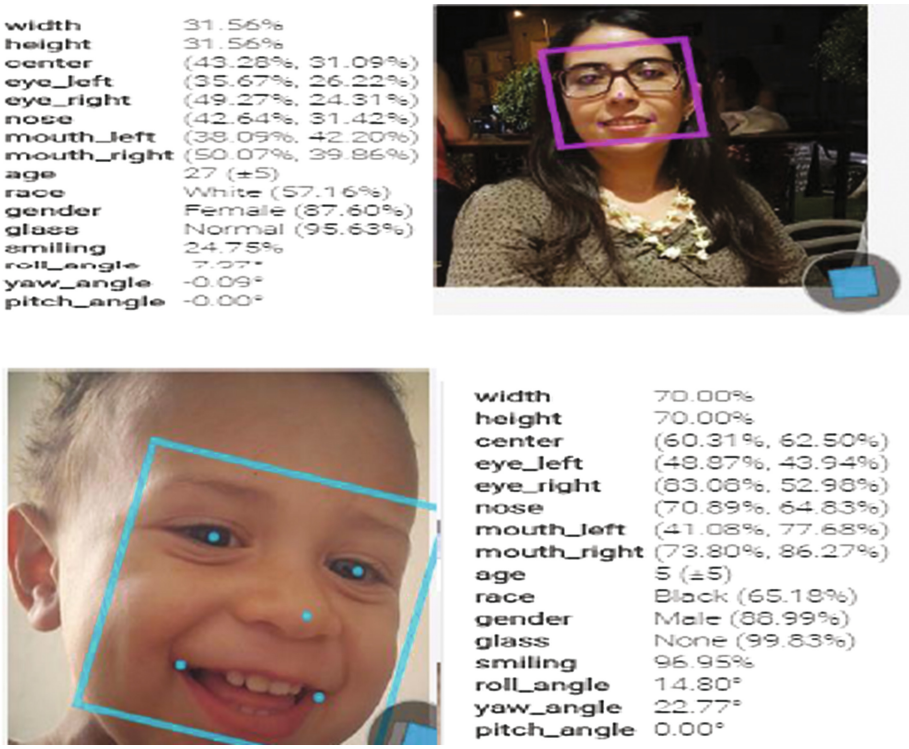


Fig. 9. The User' attributes prediction

4 Conclusion

Information on the internet is growing exponentially and in recent years, social networks are being widely used. The extraction of user profile data in social network such as Facebook is an attractive task for many researchers. Certainly, the extraction of user interests requires a global knowledge of their shared data in their profiles within social network. This paper has introduced a novel Facebook application, specialized in extracting data from Facebook, which aims to collect user profile data depending on the authorizations provided by the user. These data include comments, shares and likes, etc. Due to huge amount of data in the user profile, we use the Not Only SQL database Cassandra to store the collected data. To prove the usability of our application, i asked my colleagues in our research laboratory REGIM to access the website "Regim-lab.ddns.net", log into with their existing Facebook accounts and choose the data that can be collected by our application using the list of permissions. In the future we will try to aliment our proposed user profile ontology directory from *Social Regim* database.

References

1. Chaffty, D.: Global Social Media Research Summary 2016. National Archives, Office of Innovation, Social Media Team (2016)
2. Gao, Y., Zhen, Y., Li, H.: Filtering of brand-related microblogs using social-smooth multiview embedding. IEEE Trans. Multimedia **18**, 2115–2126 (2016)
3. Batool, R., Khattak, A., Maqbool, J., Lee, S.: Precise Tweet classification and sentiment analysis. In: IEEE/ACIS 12th International Conference Computer and Information Science (ICIS) (2013)
4. Bouhnin, C., Géry, M., Legon, C.: Personalized information retrieval models integrating the user's profile. In: Research Challenges in Information Science (RCIS) (2016). http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7541913
5. Kaur, K.: Development of a framework for analyzing terrorism actions via Twitter lists. In: IEEE International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT) (2016)
6. Tedmori, S., Al-Lahaseh, R.: Towards a selfie social network with automatically generated sentiment-bearing hashtags. In: 7th IEEE International Conference Computer Science and Information Technology (2016)
7. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P.: Practical extraction of disaster-relevant information from social media. In: Proceedings of International ACM World Wide Web Conference Committee (IW3C2), Rio de Janeiro, Brazil (2013)
8. Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., Gummadi, K.P.: Cognos: crowdsourcing search for topic experts in microblogs. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA (2012)
9. Merler, M., Cao, L., Smith, J.R.: You are what you Tweet…PIC! Gender prediction based on semantic analysis of social media images. In: IEEE International Conference on Multimedia and Expo (ICME) (2015)
10. Bonneau, J., Anderson, J., Danezis, G.: Prying data out of a social network. In: Advances in Social Network Analysis and Mining (2009)

11. Borth, D., Ji, R., Chen, T., Breuel, T., Chang, S.-F.: Large-scale visual sentiment ontology and detectars using adjective noun pairs. In: Proceedings of the 21st ACM International Conference on Multimedia (2013)
12. Ntalianis, K., Doulamis, N.: An automatic event-complementing human life summarization scheme based on a social computing method over social media content. Multimedia Tools Appl. **75**, 15123–15149 (2015)
13. Ma, X., Tsuboshita, Y., Kato, N.: Gender estimation for SNS user profiling using automatic image annotation. In: ICME Workshop on Cross-Media Analysis from Social Multimedia (CASM) (2014)
14. Fan, H., Yang, M., Cao, Z., Jiang, Y., Yin, Q.: Learning compact face representation: packing a face into an int32. In: ACM Multimedia (2014)
15. Zhou, E., Fan, H., Cao, Z., Yin, Q.: Extensive facial landmark localization with coarse-to-fine convolutional neural network. In: ICCV Workshop (2013)
16. Ouaftouh, S., Zellou, A., Idri, A.: User profile model: a user dimension based classification. In: 10th International Conference on Intelligent Systems: Theories and Application: (SITA) (2015)
17. Ouarda, W., Trichili, H., Alimi, A.M., Solaiman, B.: Towards a novel biometric system for smart riding club. J. Inf. Assur. Secur. **11**(4), 201–213 (2016). p. 13
18. Wu, Y.-C.J., Chang, W.-H., Yuan, C.-H.: Do Facebook profile pictures reflect user's personality. Comput. Hum. Behav. J. **51**, 880–889 (2014). Elsevier
19. Fonseca, H., Rocha, E., Salvador, P., Gomes, D.: A Facebook event collector framework for profile monitoring purposes. In: IEEE International Conference of Computer and Communication (ISCC) (2014)
20. Salton, G.: The SMART Retrieval System–Experiments in Automatic Document Processing. Prentice-Hall Inc, Upper Saddle River
21. Zhao, Z., Cheng, Z., Hong, L., Chi, E.H.: Improving user topic interest profiles by behavior factorization. In: International World Wide Web Conference Committee (IW3C2)
22. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Semantic enrichment of Twitter posts for user profile construction on the social web. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., Leenheer, P., Pan, J. (eds.) ESWC 2011. LNCS, vol. 6644, pp. 375–389. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21064-8_26
23. Peña, P., Hoyo, R., Vea-Murguía, J., González, C., Mayo, S.: Automatic ontology user profiling for social networks from URLs shared. In: Bielza, C., Salmerón, A., Alonso-Betanzos, A., Hidalgo, J.I., Martínez, L., Troncoso, A., Corchado, E., Corchado, J.M. (eds.) CAEPIA 2013. LNCS (LNAI), vol. 8109, pp. 168–177. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40643-0_18
24. Zemirli, N.: Modèle d'accès personnalisé à l'information basée sur les diagrammes d'infuence intégrant un profil utilisateur évolutif. in the thesis submitted for obtaining the degree of Doctor of Computing, University Paul Sabatier (2008)

Test Suite Prioritization Using Nature Inspired Meta-Heuristic Algorithms

Daya Gupta and Vishal Gupta(✉)

Department of Computer Science and Engineering,
Delhi Technological University (Formerly DCE), Delhi, India
vishalgupta013@gmail.com

Abstract. Real World is filled with various hard and complex problems. One such complex problem is an optimization problem. Optimization has been an active area of research for several decades. Optimized solutions are hard to find so there are no deterministic algorithms that can find exact solution in polynomial time. In large domain of applications of intelligence techniques we are interested in exploring the application of Biographical Based Optimization (BBO) and Grey Wolf Optimizer (GWO) meta-heuristic algorithm to the domain of software testing. The GWO mimics the leadership hierarchy and hunting mechanism of grey wolves in nature. In this paper we adapt BBO and GWO for test suite prioritization and minimization and evaluate their performance with other nature inspired meta-heuristics.

Keywords: Software testing · Meta-heuristic algorithm · Regression testing · Test prioritization · Test minimization · Grey Wolf Optimizer · Biographical based optimization

1 Introduction

Metaheuristic is one of the techniques to solve NP- Hard problems. Algorithms based on this technique are nature inspired computational techniques to address real world's complex problems for which traditional methodologies are inefficient because of the complexities of the process or because of the random and stochastic nature of problems.

Meta-heuristic algorithms are guided by the concept of exploration and exploitation. The exploration phase refers to the process of investigating the promising area of the search space as wide as possible whereas, exploitation refers to the local searching capability around the solutions obtain in exploration phase. Reason for the popularity of meta-heuristic algorithm stands on these four pillars: simplicity, flexibility, derivation-free mechanism and local optima avoidance. Majority of meta-heuristics operates on derivation free mechanism in contrast to gradient based optimization approaches since these algorithms optimizes problems stochastically. Because of this randomization meta-heuristic have the superior abilities to avoid the local optima as compared to traditional optimization techniques.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_22

Test suite is a set of test cases that are designed to be used to test some specified behavior of a software program. A single test case can be part of multiple test suites. This property of test cases results into execution of same test cases again and again. Moreover during regression testing executing all test cases are not possible, so there is need of test suite prioritization and minimization. Regression Testing is the validation to ensure that changed code has not introduced any new error in previously tested. Complete retesting is desired after maintenance but re-executing each test case is not feasible during maintenance because of time and cost constraints when the software is large. Hence in order to achieve complete coverage testing within stipulated amount of time test case prioritization and minimization is required. Selective retesting technique reduces the cost of testing a modified program by reusing the existing test cases.

The recent Meta-heuristic BBO and GWO have shown promising results in domain of remote sensing, travelling tournaments problem etc. [1, 2] In this paper we adapt them for test case prioritization and empirically analyze the performance of GWO and BBO. Thus this work compare them with other meta heuristic namely Ant Colony Optimization (ACO), Cuckoo Search (CS),) Biogeography based optimization (BBO and Simulated Annealing (SA).

Other sections of this paper have been organized in following manner. Section 2 provide literature review done in prioritization field. Section 3 discusses the implementation of proposed algorithm. In Sect. 4 experimental results has been shown. Section 5 discusses the conclusions and the future work.

2 Literature Review

Elbaum et al. [3] introduces a metric which incorporates varying test case and fault cost to predict the usability and efficiency of test case. Improved rate of fault detection is able to provide faster feedback on the system under test and help the tester in locating and correcting faults earlier than otherwise possible.

Test prioritization problem given T, a test suite; PT, the set of permutations of T; f, a function from PT to the real numbers can be stated as follows,

Problem: *Find* $T' \in PT$ *such that for all* T'' *where* $(T'' \in PT)\ (T'' \neq T')\ [f(T'') \geq f(T)]$ where, PT represents the set of all possible prioritization of T, f is function that applied to any such ordering yields an fitness value for that ordering.

Mohapatra and Prasad [4] applied evolutionary algorithm which prioritize the test case based on the code coverage of the test cases but ignoring fault detection capability of test cases. In [5] Mirarab and Talivildari uses Baysian network approach in the prioritization technique which incorporates source code changes, software fault proneness and test coverage data. But, complexity of baysian network and need of lot of training data limits their applicability in large softwares. Ryan Carlson et al. [6] implemented a prioritization technique that incorporates a clustering approach to help test case prioritization techniques on real data set. Their work proves that the test case prioritization technique which incorporates clustering improves the effectiveness of prioritization technique but clustering techniques are highly sensitive to setting of input parameters.

Tallam and Gupta [7] proposed a new greedy heuristic algorithm for choosing a subset of test cases T' from test suite T which have the ability to cover all requirements as covered by T but approach sometimes stuck at local maxima.

Metaheuristics with the capability to achieve solution with limited information and computational capacity have also been applied in past. Hla et al. [8] prioritise the test cases based on the altered part of software using PSO. Singh et al. [9] proposed Ant Colony optimization which works on the real life behavior of the ants and explains how randomness characteristic in ACO helps in exploration of optimal test case order. But results in ACO depends upon the initial pheromone value and other initial parameters. In [10] Nagar et al. tries to leverage the property of obligate broad parasitism of some cuckoo species to solve the prioritization problem of test cases but cuckoo search depends on the probability of detecting alien eggs and involves biased generation of new cuckoo eggs.

Grey Wolf Optimizer proves to be comparatively efficient than many other algorithms to solve NP-Hard problems which requires selecting the optimal solution from the combinatorial search space. Grey Wolf optimizer (GWO) has shown promising and more than satisfactory results in the various domains like feature subset selection [11], vehicle routing [12] and optimizing key values in cryptographic algorithms [13]. GWO has been successfully applied to solve binary [11] and multi-objective problems [14].

3 Proposed Algorithm

3.1 Biogeography Based Optimization (BBO)

Bob Simon [15] introduced Biogeography Based optimization (BBO) which is a population bio-inspired meta-heuristic algorithm based on evolutionary algorithm by the virtue of randomization iteratively. It has been applied for solving various NP-Hard problem which includes test suite prioritization.

Algorithm of Biogeography based Optimization for Test Case Prioritization Main requirements which are to be satisfied before applying BBO are:

- *Habitat Representation*: Each habitat represents a different test case ordering, which can be a potential solution.
- Mutation and migration operators are discretized according to test prioritization problem
- Objective function f(x): Average Percentage Fault Detection Metric
- Terminating Condition: Number of iterations
  $\{data\}_k$ represents in Boolean table, showing faults covered by each test case, $k = \{1, 2, N\}$

I. Define an Elite array for selection best individual in each generation and set its size ratio nKeep of the size of population
II. Immigration Rate = 1 − Migration Rate and Mutation probability (σ) ← values [0, 1]
III. Define an initial random population

IV. While (not the maximum generation limit achieved)
    a. Define a Random number
    b. For each {data} k, set the emigration probability δk proportional to fitness of {data} k, δ ϵ [0, 1]
    c. For each {data}k, set the immigration probability λk, λk ← 1 − δk
    d. Set {temp} ← {data}
    e. For each individual {temp}
        i. Use immigration probability to identify the individual which has be emigrated
        ii. Selection of immigrant is based on roulette wheel
        iii. Perform mutation with probability pMutation
    f. Keep nKeep number of good habitat along with Population-nKeep*Population new population for next iteration
V. End While
VI. Best Habitat obtained represents the prioritized order of test cases
VII. Identify the redundant test cases using the following function:
    a. From least ranked test case to highest ranked test case:
    If faults covered by test case is already covered by higher ranked test cases then remove the test case otherwise add test case to front of test cases to be executed
VIII. Final list contained the prioritized and minimized set of test cases

3.2 Grey Wolf Optimizer (GWO)

Mirjalili et al. [16] proposes a new metaheuristic called Grey Wolf Optimizer (GWO) inspired by grey wolves. The GWO algorithm mimics the leadership hierarchy and hunting mechanism of grey wolves in nature. Four types of grey wolves which formulate the hierarchy are alpha, delta, omega and beta. In addition three main steps of hunting, searching for prey, encircling prey and attacking prey are implemented to stimulate the wolves' behavior. The ones leading the pack are called alphas. The alpha wolves are called the dominant wolf since their instructions has been followed by the pack. The beta wolves are stated as next best candidate to become alpha in case alpha wolf retired. If a wolf is none of the above classified wolves then they belong to subordinate group named delta. Delta wolves follow the alphas and betas, but they overpower the other left wolves i.e. omega. Omega wolves take order from all the other more aggressive wolves.

Social hierarchy in the wolves while designing GWO can be outlined by considering alpha (α) as the best solution. Similarly, second and third best solutions are chosen as beta (β) and delta (Δ) respectively. All reminder solutions in the group are considered as omega (Ω). In the GWO algorithm the optimization of problem is guided by α, β, and Δ wolves. The reminder Ω wolves tend to move toward these three dominant wolves.

For mathematical representation of the hunting mechanism of grey wolves, it is considered that the alpha, beta, and delta (best solution among the pack) possess better knowledge about the location of optima. Therefore, we store the first three best solutions

obtained so far and other search wolves update their positions according to the position of the alpha agents using following formulas [16]

$$\overrightarrow{D\alpha} = |\overrightarrow{C1} \,.\, \overrightarrow{X\alpha} - \vec{X}|, \overrightarrow{D\beta} = |\overrightarrow{C1} \,.\, \overrightarrow{X\beta} - \vec{X}|, \overrightarrow{D\Delta} = |\overrightarrow{C1} \,.\, \overrightarrow{X\Delta} - \vec{X}|$$
$$\overrightarrow{Xa} = \overrightarrow{X\alpha} - \overrightarrow{A1} \,.\, \overrightarrow{D\alpha}, \overrightarrow{Xb} = \overrightarrow{X\beta} - \overrightarrow{A2} \,.\, \overrightarrow{D\beta}, \overrightarrow{Xc} = \overrightarrow{X\Delta} - \overrightarrow{A1} \,.\, \overrightarrow{D\Delta}$$
$$\vec{X}(t+1) = \frac{\left(\overrightarrow{Xa} + \overrightarrow{Xb} + \overrightarrow{Xc}\right)}{3}$$

The vectors $\vec{A}$ and $\vec{C}$ are obtained using $\vec{A} = 2\vec{a}.\overrightarrow{r1} - \vec{a}$ and $\vec{C} = 2.\overrightarrow{r2}$ where value of $\vec{a}$ is linearly decreased from 2 to 0 over each iterations and r1, r2 are random vectors such that r1, r2 € [0, 1].

Each wolf is separate independent potential single solution. A solution representation for the test suite prioritization is a permutation of test cases. In our case wolf represents single test case prioritization order of test cases and collection of wolf, pack, represents a complete set of multiple test case orderings. In this representation, elements of wolf vector represents test case and their index represents the position in test suite. Every test case can be place after every other test case in all wolves assuming that no test case repeats itself in single prioritization order. The numbers of wolves are fixed in pack. The number of wolves in the pack is equal to the size of the population. In continuous problems, the meaning of divergence and convergence towards prey is obvious. However, for combinatorial problems, the notion of divergence and convergence requires that the given solution must be generated by the perturbation. Perturbation must make the minimum changes on the candidate solution. This leads to the 2-opt move and double bridge move for a new solution as shown in Figs. 1 and 2. The minimum number of non-contiguous edges that we can delete from a path is two. So in 2-Opt Operator breaking of two edges and reconnecting them in other way to make new solution with minimum changes is done. Here a, b, c, d and dots represent the test cases and ordering of test case is path followed from S to E. Double Bridge perturbation cuts four edges and introduces four new ones.

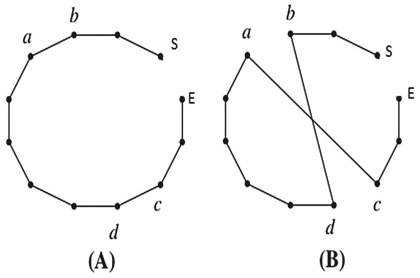


Fig. 1. 2-opt move (A) Initial ordering from S to E (B) The ordering created by 2-opt move [the edges (a, b) and (c, d) are removed, while the edges (a, c) and (b, d) are added]

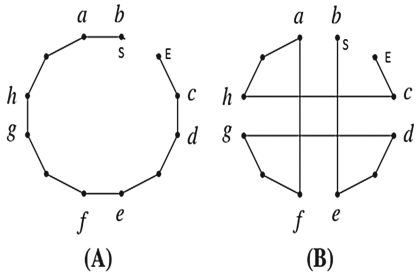


Fig. 2. Double-Bridge Move (A) Initial ordering from S to E. (B) The ordering created by double-bridge move [the edges (a, b), (c, d), (e, f) and (g, h) are replaced by the edges (a, f), (c, h), (e, b) and (g, d)]

Each wolf representing a solution in the search space is associated with a numeric objective value and the quality of a solution is proportional to the value of the objective function. In GWO, a wolf of better quality will attract other wolves to move towards itself and better wolves are also exploring the search space. This means that the quality of a wolf is directly related to its ability to converge towards global solution. For the test case prioritization problem, the quality of a solution is related to the value of Average Percentage of fault detected (APFD) metric. The best solution is the one with the highest average percentage of fault detection value.

Mathematically, Let T be a test suite containing n test cases and let F be a set of m faults revealed by T. Let TF_i be the first test case in ordering T' of T which reveals fault i. The Average Percentage of fault detected metric (APFD) for test suite T' is given by the equation in [3] as follows, $APFD = 1 - \frac{TF1 + TF2 + \quad + TFm}{n * m} + \frac{1}{2 * n}$.

Algorithm of GWO for Test Prioritization.

Objective function f(x): Average Percentage Fault Detection Metric
Input: n Number of grey wolves in the pack,
Stopping Criteria: Number of iterations for optimization
Output: x_α which represents optimal grey wolf test case ordering

I. Initialize a population of n grey wolves' positions randomly. Each wolf representing a test case ordering such that value in ordering € [1 no_testcases]. Wolf represents single test case prioritization order of test casee and pack represents a complete set of multiple test case orderings.
II. Find the α, β and δ solutions based on their fitness values. Alpha wolf is represented by the fittest solution in the iteration. Beta and Omega are represented by second and third best in the pack
III. While Stopping criteria not met do
    a. For each wolf in pack update current wolf's ordering according to getUpdatedWolf procedure
    b. Evaluate the ordering of individual wolves
    c. Update a, A, C, α, β, and δ

IV. End While
V. Identify the redundant test cases using the following function:
   a. From least ranked test case to highest ranked test case:
      If faults covered by test case is already covered by higher ranked test cases then remove the test case otherwise add test case to front of test cases to be executed
VI. Final list contained the prioritized and minimized set of test cases

Algorithm for getUpdatedWolf Procedure

The main updating equation can be formulated as $cstep = \frac{1}{1 + e^{(-10*(A*D_wolf-0.5))}}$.

Where, A is calculated used equation $\vec{A} = 2\vec{a}.\overrightarrow{r1} - \vec{a}$ and D_wolf is calculated using difference in position of two wolves which is calculated by element wise subtraction of two test ordering and then counting number of non-zero elements in the result.

I. Define two random numbers r1 and r2 € [0, 1]
II. Calculate the value of A1, A2, A3 and C1, C2, C3 using equation $\overrightarrow{A} = 2\vec{a}\,.\,\overrightarrow{r1} - \vec{a}$ and $\overrightarrow{C} = 2\,.\,\overrightarrow{r2}$ for different values of r1 and r2 while calculating

III. Calculate cstep(α), cstep(β), cstep(δ) using cstep(x) $= \frac{1}{1 + e^{(-10*(A*D_{wolf(x)}-0.5))}}$.
IV. For alpha, beta, delta
   If cstep >= rand then choose Double Bridge move otherwise choose 2-opt move
V. End for
VI. Perform Crossover(x, y, z) using simple stochastic crossover strategy where

$$X = \begin{cases} \text{x if rand} < \frac{1}{3} \\ \text{y if} \frac{1}{3} < rand < \frac{2}{3} \\ \text{z if rand} > \frac{2}{3} \end{cases}$$. Return X as updated wolf ordering

4 Experimental Results

The proposed BBO and GWO algorithm have been implemented using Matlab. The research was carried out in a system with 2 GB RAM, 500 GB HDD and 1.6 GHz core i5 Intel Processor using simulation environment shown in Table 1:

Table 1. Simulation variable

S. No	Simulation variables	Value
1.	Programming Language	MATLAB
2.	Data Base	Excel
3.	Data Set Repository	Unit test cases of Jmeter Software itself

After simulation of the algorithm GWO for 30 iterations on data set, obtained after compiling Jmeter separately for Jmeter's Junit test cases, we obtained the prioritized order displayed in Fig. 3. As shown in Fig. 4, the priorities are ordered in ascending order with test case at index 1 as the highest priority and 24 being the least. The result showed that test case 1 is much more important than rest of test cases. Test case 14 is least prioritized in testing the concerned application.

```
[1 20 23 6 21 13 19 24 12 4 5 22 9 11 17 7 3 15 8 10 16 18 2 14]
```

Fig. 3. Priority order of test cases

```
1    20    23    6    21
```

Fig. 4. Reduced test cases

Test cases after test case 21 in priority order has been reduced to zero which showed that these test case can be excluded during regression test suite selection. These test cases are found to be redundant and have been identified as worthless by Grey Wolf Optimizer (Tables 2 and 3).

Table 2. Parameter setting during research

S.no	Parameters	BBO	ACO	GWO	Cuckoo	SA
1.	Mutation Probability	0.4	-	-	-	-
2.	Elite Solution	Best habitat	Phoneme Matrix	Alpha, beta delta	Best cuckoo	
3.	Generation Limit	1000	100	1000	1000	1000 /12
4.	Cost Function	APFD	APFD	APFD	APFD	APFD
5.	Discovery rate of solutions	0.2 (keep rate)	-	-	0.25	-
6.	Population	10	10	15	15	1
7.	Initial temp	-	-	-	-	0.025
8.	Alpha	-	1	-	-	0.99
9.	Beta	-	1.5	-	-	-
10.		-	0.05	-	-	-

After careful evaluation of box graph drawn it is evident that modified GWO is much more stable in test suite minimization. APFD value of GWO is least varied in comparison to other algorithms. ACO shows highest variance in performance making it least stable. As evident from the above box graph in Fig. 5, Grey wolf show least variation in the APFD value with most of the results lie near the median resulting into vary sharp 25% and 75% boxes around the median making them even unrecognizable in graph, with

minimum value of APFD being maximum among all the compared algorithms, making GWO most efficient. However in 30 runs Cuckoo, GWO and BBO all obtain similar maximum APFD value. Minimum APFD over 30 runs is highest in GWO making GWO most promising even in its vulnerable state. Since all algorithms compared here are efficient meta-heuristic algorithms so number of worthless test case identified and usable test case are same for all algorithms but the major difference is present in convergence rate of the algorithm in term of internal processing requirement which shows that GWO requires least processing with minimum resource requirement since it converges faster. Thereby, GWO perform considerably better than all algorithms stated above in term of speed. In fact the combined execution time of reduced test cases even in the worst case is minimum in case of GWO. However in best case the proposed time is same because of meta-heuristic problem solving capability of all algorithms.

Table 3. Comparison table of various algorithms based on different factors

S.no	Characteristics		BBO	ACO	GWO	Cuckoo	SA
1	Complete coverage test cases identified (low is better)		5	5	5	5	5
2	Number of initial parameters required		4	7	2	3	4
3	Redundant test case identified (high is better)		19	19	19	19	19
4	Total Processing Iterations		24000	24000	10000	33125	12000
5	Total Execution time of reduced test cases	Worst	25	113	21	113	23
		Best	6	6	6	6	6
6	Average Convergence rate (in term of processing iterations)		19002	12989	3069	18935	11723
7	Variance in the APFD value		0.098946	0.523622	0.035818	0.251797	0.087572
8	Reduction in test suite (in %)		79	79	79	79	79

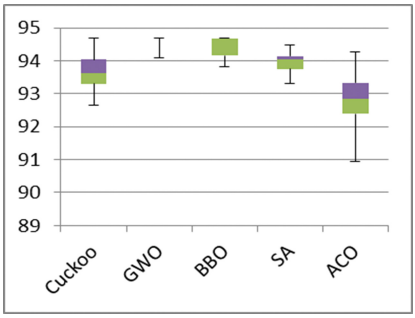


Fig. 5. Box graph of test case ordering for various algorithms (in term of APFD metric)

5 Conclusion and Future Work

Large number of algorithms have been proposed and used in the refining the process of regression test suite but all of these algorithms suffers from parameter setting problem, elitism problem and loss of coverage. BBO and Grey wolf optimization have been implemented and pioneered in various field previously have been modified in this paper and compared with other already acclaimed meta-heuristic techniques.

BBO and GWO algorithms obtained state of the art result and thereby sufficient reduction in test cases after their execution irrespective of the setting of initial parameters. The novel technique GWO can further be extended to multi-objective test suite optimization by altering the comparison mechanism using the pareto dominance and optimal front operator where a single wolf contain multiple solution and nature of wolves can be identified using pareto operators.

References

1. Gupta, D., et al.: Enhanced heuristic approach for travelling tournament problem based on extended species abundance models of biogeography. In: ICACCI 2014, pp. 1118–1124 (2014)
2. Goel, L., Gupta, D., Panchal, V.K., Abraham, A.: Taxonomy of nature inspired computational intelligence: a remote sensing perspective. In: 2012 Fourth World Congress on Nature and Biologically Inspired Computing (NaBIC), pp. 200–206. IEEE, November 2012
3. Elbaum, S., Malishevsky, A., Rothermel, G.: Incorporating varying test costs and fault severities into test case prioritization. In: Proceedings of the 23rd International Conference on Software Engineering. IEEE Computer Society (2001)
4. Mohapatra, S.K., Prasad, S.: Evolutionary search algorithms for test case prioritization. In: 2013 International Conference on Machine Intelligence and Research Advancement (ICMIRA). IEEE (2013)
5. Mirarab, S., Tahvildari, L.: A prioritization approach for software test cases based on bayesian networks. In: Dwyer, M.B., Lopes, A. (eds.) FASE 2007. LNCS, vol. 4422, pp. 276–290. Springer, Heidelberg (2007). doi:10.1007/978-3-540-71289-3_22

6. Carlson, R., Do, H., Denton, A.: A clustering approach to improving test case prioritization: An industrial case study. In: 2011 27th IEEE International Conference on Software Maintenance (ICSM). IEEE (2011)
7. Tallam, S., Gupta, N.: A concept analysis inspired greedy algorithm for test suite minimization. ACM SIGSOFT Softw. Eng. Notes **31**(1), 35–42 (2006)
8. Hla, K.H.S., Choi, Y.S., Park, J.S.: Applying particle swarm optimization to prioritizing test cases for embedded real time software retesting. In: IEEE 8th International Conference on Computer and Information Technology Workshops, CIT Workshops 2008. IEEE (2008)
9. Singh, Y., et al.: Test case prioritization using ant colony optimization. ACM SIGSOFT Softw. Eng. Notes 1–7 (2009)
10. Nagar, R., et al.: Test case selection and prioritization using cuckoos search algorithm. In: 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE). IEEE (2015)
11. Emary, E., Zawbaa, H.M., Grosan, C., Hassenian, A.E.: Feature subset selection approach by gray-wolf optimization. In: Abraham, A., Krömer, P., Snasel, V. (eds.) AECIA 2014. AISC, vol. 334, pp. 1–13. Springer, Heidelberg (2015). doi:10.1007/978-3-319-13572-4_1
12. Korayem, L., Khorsid, M., Kassem, S.S.: Using grey wolf algorithm to solve the capacitated vehicle routing problem. In: IOP Conference Series: Materials Science and Engineering, vol. 83, no. 1. IOP Publishing (2015)
13. Shankar, K., Eswaran, P.: Sharing a secret image with encapsulated shares in visual cryptography. Proc. Comput. Sci. **70**, 462–468 (2015)
14. Mirjalili, S., et al.: Multi-objective grey wolf optimizer: a novel algorithm for multi-criterion optimization. Expert Syst. Appl. **47**, 106–119 (2016)
15. Simon, D.: Biogeography based optimization. IEEE Trans. Evol. Comput. 702–713 (2008)
16. Mirjalili, S., Mirjalili, S.M., Lewis, A.: Grey wolf optimizer. Advan. Eng. Softw. **69**, 46–61 (2014)

A New Trajectory Optimization Approach for Safe Mobile Robot Navigation: A Comparative Study (Khepera II Mobile Robot)

Walid Ellili(✉), Abdelfetteh Lachtar, and Mounir Samet

Laboratory of Electronics and Information Technology,
University of Sfax, ENIS, BP 1173, Sfax 3038, Tunisia
elliliwalid@gmail.com, lachtar.abdelfetteh@gmail.com,
mounir.samet@enis.rnu.tn

Abstract. Through the present paper, a new approach useful for solving the obstacle avoidance and trajectory optimization problems during robot navigation for certain tasks to be performed at minimum costs. In its real sense, the obstacle avoidance approach is based on the application of two fuzzy controllers, the first of which is designed to join the object, while the second is conceived to serve as an obstacle avoiding device. The trajectory optimization approach is based on the gradient method. Prior to implementing the solution on the real robot, the simulation has been integrated in an immersive virtual environment, for a more effective movement analysis and safer testing purposes. The study findings prove to reveal well that the proposed approach turns out to exhibit a good average speed and a satisfactory target-reaching success rate, while the optimization oriented gradient method has turned out to be rather efficient in respect of the genetic algorithm approach.

Keywords: Fuzzy controller · Mobile robot · Obstacle avoidance · Gradient method · Genetic algorithm · Optimization

1 Introduction

A moving robot is a vehicle equipped with a sensor enabling it to move in an appropriate environment to accomplish certain tasks [1]. A remarkable progress has been made in all areas of robotics: perception, environmental modeling, automatic control of actuators, motion planning as well as scheduling tasks [2]. Technological progress has turned out to be progressively needed, sometimes critically versatile in our life for many problems to be solved. Originally, automated systems have initially been designed to substitute human beings in performing tedious tasks, dangerous applications or accomplishing works that are beyond one's physical capacities [3]. Technological developments, particularly those relating to the fields of electronics and information technology, prove to contribute a great deal in enhancing the promotion of robotics, along with intelligent machines and devices.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_23

On further equipping the system with extra perception abilities, action and decision executing capacities, roboticists are nowadays increasingly interested in providing mobile robots with extra autonomy in a bid to make them interact more easily with their environments [4]. Thus, mobile robots have more recently been conceived to move more independently around rather structured environments.

Noteworthy, however, certain problems do actually persist in the robotics field, namely, the reactive obstacle avoidance to reach a predefined target in advance [5], as well as the performance of such a task at a minimal cost. In this respect, the present work is focused on investigating the robot navigation problem in presence of environmental obstacles, while optimizing the obtained trajectories. Actually, the navigation strategy rests on two types of fuzzy controllers. First, in case of inexistent obstacles, the fuzzy controller will monitor the robot to return to the desired configuration. Second, in case of obstacle existence, the controller charged with avoiding obstacles will be activated. The resulting trajectories' optimization is maintained by means of the gradient algorithm. In a first place, detection is ensured by means of the robot's own sensors. In a second place, detection is maintained through the introduction of new infrared sensor. In a last stage, an implementation phase of the developed algorithms is implemented on the Khepera II robot, for the simulation results to be checked. For this sake, this work is structured as follows: Sect. 2 is designed to sum up the trajectory optimization methods, while Sect. 3 provides an overview of the materials applied. As for Sect. 4, it involves a presentation of the different approaches undertaken to preserve the obstacle avoidance process along with the trajectory optimization, followed by an evaluation stage maintained through simulation and experimentation, subjects of Sects. 5 and 6. Finally, Sect. 7, bears the major concluding remarks, while paving the way for prospective future work and new horizons.

2 Trajectory Optimization Framework

Among the optimization methods, one might well cite the genetic algorithms as well as the gradient one. To note, the genetic algorithm is used in a number of well' known undertaken projects. As part of the control process, genetic algorithms can be applied as powerful methods whereby optimal research could be implemented in areas the scope of which is, apriori, unknown, or for the sake of determining the relevant effective control strategies within a complex environment [7]. Actually, the genetic algorithms' interest resides essentially in the fact that they allow for searching the most optimum solutions in areas whose extent is unknown, thus, allowing for critical processes to be attained within a large number of situations. In this respect, Genetic algorithms constitute a set of procedures that may stand as natural selection mechanisms. Their basic fundamental principle lies in simulating natural evolution processes within a hostile environment. In so far as the present work is concerned, the gradient algorithm has been applied as an optimization algorithm, whose principle consists in starting from a random point prior to moving towards the steeper inclination direction [8]. On applying a number of iterations, the algorithm converges up to a point which constitutes an the criterion extremum to be minimized. In this way, the gradient algorithm consists in

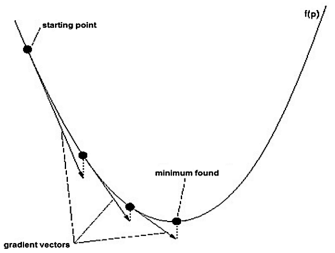


Fig. 1. The gradient descent principle, relevant to the function-to-variable case.

moving from a starting point (see Fig. 1) or one iteration u_k following the maximum slope line, as associated with the cost function f.

The descent direction, corresponding to this particular greatest slope outcome u_k line provided by the gradient $\frac{\partial f}{\partial uk}$.

A further iteration can be even rendered by:

$$U_{K+1} = U_K - \varepsilon \frac{\partial f}{\partial uk} \tag{1}$$

Where ε stands for a fixed positive parameter.

The algorithm appears to converge whenever the $\frac{\partial f}{\partial uk}$ gradient approaches zero. Hence, the algorithm turns out to involve the following steps: Initialization, the beginning of the optimization loop (iteration k), the state equations' resolution, the cost function computation, cost function gradient calculation, Computation of the new vector relevant control variables, the optimization loop end.

3 Hardware Architecture

3.1 The Khepera II Platform Description

The robot subject of study application is the khepera II mobile robot. It is a two-wheel equipped mobile vehicle, developed by the Autonomous Systems Laboratory of the Federal Polytechnic School of Lausanne (EPFL). It has been the outcome and subject of several elaborated projects.

3.2 Detection Tools

To aquire an autonomous feature, the mobile robot must fulfill a number of capabilities. It must be able to sense its environment and to be located in it. For this purpose, special sensors, such as sonars, must be installed. They consist in particular devices useful for measuring distances separating it to any nearby obstacles. The notion of perception [9] in mobile robotics is related to the robot's ability to receive such necessary information as process and format for the robot to act and react suitably well in compliance with the

surrounding environment. For this reason, we try hard to get as much information as possible on the environment for the robot behavior to be adequately adjusted. More particularly, some interesting information turns out to be essential to apply, namely, data concerning the robot distance relative to a wall or to an object on the ground. Once the distance is recognized, the robot should be able to move from one point to another through retrieving effective ways to avoid collisions with the existing obstacles. For this sake, the khepera II robot has been equipped with eight infrared sensors whose detection scope does not exceed five centimeters. To solve the problem of detectors' short scope, we considered it essential to add an extra sensor, in an attempt to further increase the detection range.

4 Applied Methods

The robot is assimilated to a particle moving along the x and y coordinates in the reference frame, in accordance with an angle orientation, as shown in the following figure (Fig. 2):

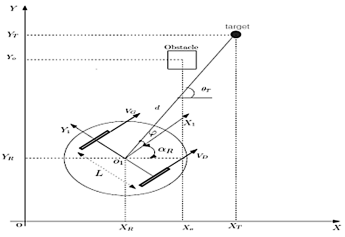


Fig. 2. Robot navigation area with an obstacle.

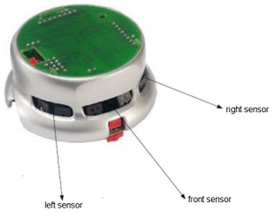


Fig. 3. Arrangement of the three obstacle-detecting sensors.

4.1 Obstacle Avoidance

A fuzzy logic method has been selected to ensure the mobile robot's navigation and acquire the obstacle avoiding behavior. The obstacle is assumed static and square in shape with a 5 cm side. According to this framework, the robot is equipped with three sensors: a frontal sensor, along with two other ones placed on both sides (right and left) (as shown in Fig. 3). These sensors are responsible for detecting the obstacle in the three indicated directions. The approach is based on application of two fuzzy controllers. While the first is designed to join the object, the second is conceived to serve as an obstacle avoiding device (Fig. 4).

The applied controller is the Takagi-Sugeno one, of order 0 [10].

Once entries are recognized, the outputs will be determined on the basis of the fuzzy rules' pertinent degrees. This controller can be activated only if the distance separating the obstacle to the robot is inferior to the distance covered by the robot sensors' range. The controller input allocating variables: For the designed controller,

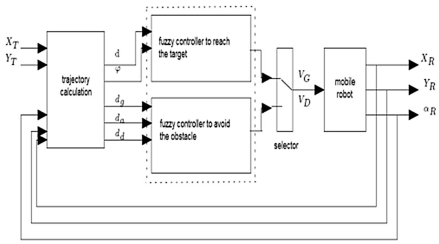


Fig. 4. Control chart with an obstacle avoidance device.

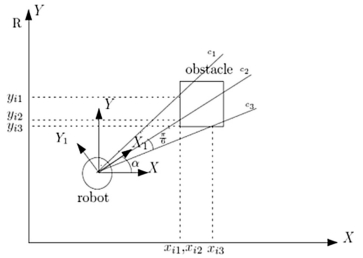


Fig. 5. Obstacle as detected by the robot's three sensors.

we have chosen as input variables the right side distance dd, the left side one dg, as well as the distance separating the robot to the obstacle front side da. Figure 5, below, displays the approach graphic presentation.

With: c1, c2 and c3 denoting the lines respectively passing through each of the left, front and right sensors' center. xi1, xi2, xi3, yi1, yi2 and yi3: are, respectively, the X and Y coordinates of the straight lines intersection points' dg, da and dd with the obstacle. If the obstacle is placed in front of the robot, it will be simultaneously detected by the three sensors so that the distances dg, da and dd separating the robot and the obstacle would stand as the fuzzy controller's inputs, allowing the robot to avoid the obstacle. The expressions corresponding to the three distances are given by:

$$d_g = \frac{1}{\tan(\alpha_R + \frac{\pi}{6})}(Y_{i1} - b_1) \tag{2}$$

$$d_a = \frac{1}{\tan \alpha_R}(Y_{i2} - b_2) \tag{3}$$

$$d_d = \frac{1}{\tan(\alpha_R - \frac{\pi}{6})}(Y_{i2} - b_2) \tag{4}$$

With

$$b_1 = Y_R - \tan(\alpha_R + \frac{\pi}{6})X_R \tag{5}$$

$$b_2 = Y_R - \tan(\alpha_R)X_R \tag{6}$$

$$b_3 = Y_R - \tan(\alpha_R - \frac{\pi}{6})X_R \tag{7}$$

(XR, YR) stands for the robot x and y coordinates. At this level, the discourse input universe is divided into three pertinent functions. The controller outputs correspond to both of the right and left wheels' respective speeds. The fuzzy controller which serves

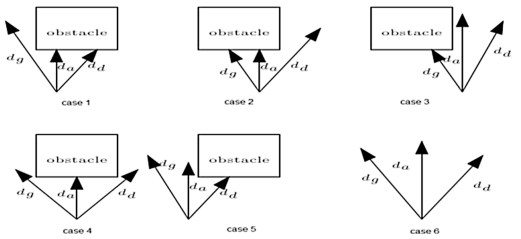


Fig. 6. Different obstacle detection cases as ensured through the three sensors.

to avoid the obstacle is activated once one of the three distances dg, da and dd proves to be inferior to the robot's sensors' range. Figure 6, below, depicts the obstacle detection capabilities via the three sensors.

Case 1: Obstacle detection case via both of the front and right sensors only. Case 2: obstacle detection case exclusively maintained by the front and left sensors. Case 3: Obstacle detection exclusively through the left sensor. Case 4: The obstacle is detected exclusively by the front sensor. Case 5: The obstacle is detected exclusively by the right sensor. Case 6: No obstacle is being detected. All possibilities have been studied for the purpose of allowing the robot to navigate and avoid obstacles without difficulty. The range reach of the robot sensors ranges between 4 and 5 cm. As such result, the robot must approach the obstacle closely to be able to detect it. Such a situation is likely to result in sudden changes to occur in the robot's speed. To overcome this problem, we have envisaged adding an extra infrared sensor whose range could reach 30 cm. Indeed, this sensor would have the capacity and advantage of detecting the obstacle from a farther distance and, thereafter, avoiding sudden speed changes on the one hand, and maintaining smoother and more optimal trajectories, on the other. Both of the dg and dd inputs' discourse universe remains unchanged (as it has been) i.e. it varies between 0 and 50 mm, whereas the previously scored distance [0; 50] has grown to reach the extent of [0; 300], thanks to the new sensor's range, which is able to reach 300 mm. These three inputs are divided into three membership functions of a Gaussian type.

4.2 Trajectory Optimization

The criterion consists in minimizing the distance separating the target position to that of the next sample step.

$$J = \frac{1}{2}\left[(X_R - X_T)^2 + (Y_R - Y_T)^2\right] \tag{8}$$

With: (XR, YR): designating the robot's current position, (XT, YT): the target's position. Each of the fuzzy controller's output, or VG VD, is calculated in terms of the expressions 9 and 10. Noteworthy, however, is that the fuzzy inference table's adjustment appears to cater exclusively for the fuzzy rules' pertinent conclusions, which can be accomplished via the following equations:

$$C_{Di}^{mouv} = C_{Di}^{anc} - \varepsilon \frac{\partial J}{\partial C_{Di}} \tag{9}$$

$$C_{Gi}^{mouv} = C_{Gi}^{anc} - \varepsilon \frac{\partial J}{\partial C_{Di}} \tag{10}$$

The derived criterion relating to the fuzzy rules' conclusions are provided by:

$$\frac{\partial J}{\partial C_G} = \frac{\partial J}{\partial X_R}\frac{\partial X_R}{\partial V_G}\frac{\partial V_G}{\partial C_G} + \frac{\partial J}{\partial Y_R}\frac{\partial Y_R}{\partial V_G}\frac{\partial V_G}{\partial C_G} \tag{11}$$

$$\frac{\partial J}{\partial C_G} = T^2 \frac{V_G + V_D}{2L} \frac{\alpha_i}{\sum_{j=0}^{r} \alpha_j} [(X_R - X_T)\sin\alpha - (Y_R - Y_T)\cos\alpha] \tag{12}$$

$$\frac{\partial J}{\partial C_D} = \frac{\partial J}{\partial X_R}\frac{\partial X_R}{\partial V_D}\frac{\partial V_D}{\partial C_D} + \frac{\partial J}{\partial Y_R}\frac{\partial Y_R}{\partial V_D}\frac{\partial V_D}{\partial C_D} \tag{13}$$

$$\frac{\partial J}{\partial C_D} = T^2 \frac{V_G + V_D}{2L} \frac{\alpha_i}{\sum_{j=0}^{r} \alpha_j} [(X_R - X_T)\sin\alpha - (Y_R - Y_T)\cos\alpha] \tag{14}$$

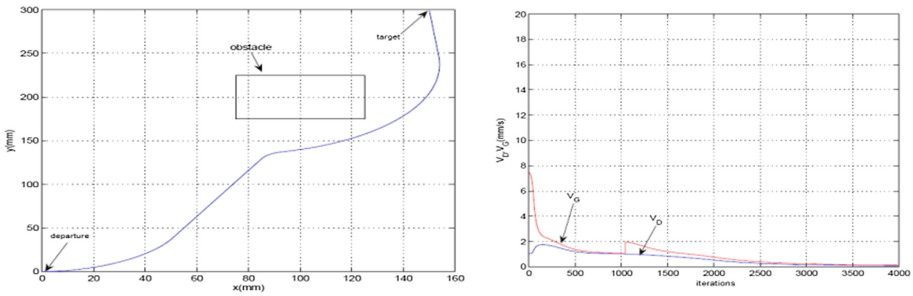
5 Simulation Results

The simulations have been undertaken within a Matlab environment for the sake of testing the controller's ability to reach the designed target, while avoiding the encountered obstacle.

5.1 Obstacle Avoidance

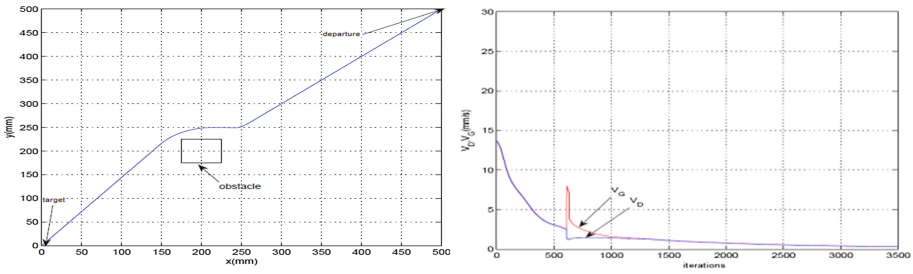
In this respect, and regarding different robot positions, several obstacle locations and different targets, have been set up in advance. The relevant simulation results are depicted in Figs. 7, 8 and 9, below.

Based on Figs. 7, 8 and 9, one can note that starting from any location point, the robot turns out to be able to detect the obstacle and avoid it during navigation, by acting on both of the right and left wheels' respective speeds. Indeed, the actual improvement has been observed on the two speeds' pertaining values on approaching the obstacle. In Fig. 9, for instance, the right wheel's speed has been greater than that of the left wheel for the robot to move away rightwards from the obstacle and achieve the set target. For the purpose of assessing the validity of these achieved findings, a robustness test concerning the controller's performance has been undertaken, to evaluate its response to any change in either the initial position or in the target one. The obtained results appear in Fig. 10, below.



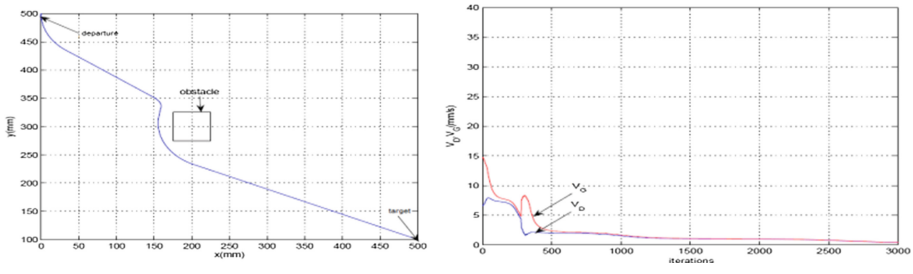
(a) (b)

Fig. 7. (a): Simulation results for an obstacle placed at a level of (100, 200) and a target at (150, 300); (b): The robot's left and right wheels' corresponding speeds.



(a) (b)

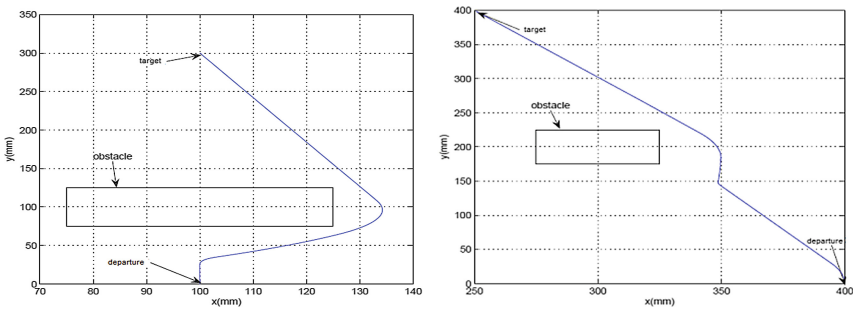
Fig. 8. (a): Simulation results for an obstacle placed at the (200, 200) level and a target at (0, 0) and a starting point of (500, 500); (b): The robot's left and right wheel's respective speeds.



(a) (b)

Fig. 9. (a): Simulation result relevant to an obstacle placed at (200, 300), (b): The robot's left and right wheels' respective speeds.

At this level, the developed fuzzy controller proves to fit perfectly well for ensuring the mobile robot's motion from any initial position to any desired position, without hurting any of the (avoided) obstacles. Figure 11, below, indicates the new sensors

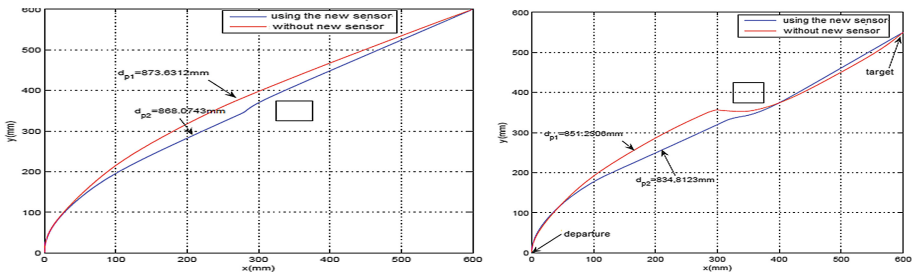


(a) (b)

Fig. 10. (a): Simulation result relevant to an obstacle placed at point (100, 100); (b): Simulation Results relevant to an obstacle placed at point (300, 200).

achieved simulation results in respect of those reached by means of the khepera II robot's proper sensors. Figure 11 reveals well that the robot's navigation performance has recorded an improvement as the distance traveled starting from the same initial conditions has been noticeably reduced with application of the new sensor. The following table depicts a comparison of the distances traveled with and without introduction of the new sensor (Table 1).

Relying on this table, one can well deduce that the new sensor is suitably fit for the application presented regarding each of the treated cases. Actually, the distance traveled proves to decrease following implementation of this newly-devised sensor, as it appears to help significantly in detecting the obstacle at a farther distance, by pursuing the most optimal path.



(a) (b)

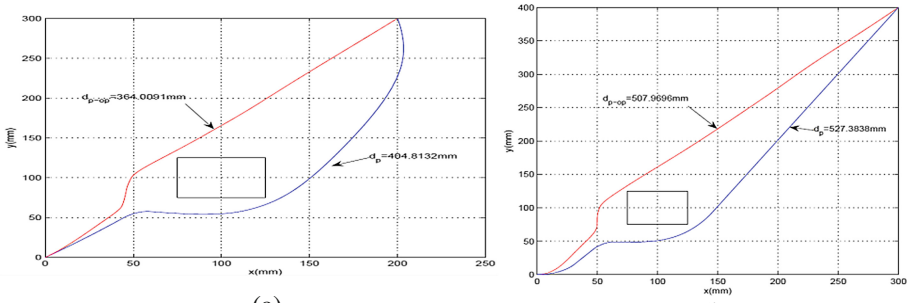
Fig. 11. (a), (b): Obstacle avoidance without and with the newly developed sensor, dp1: distance without application of the new sensor, and dp2: distance using the new sensor.

5.2 Trajectory Optimization

The Fig. 12 pertinent curves reveal well that the optimized controller reached distances turn out to be lower than the pre-optimization attained ones.

Table 1. Distances traveled with and without the new sensor with regard to different conditions

Starting point (mm)	Target (mm)	Obstacle position (mm)	Distance traveled without the new sensor (mm)	Distance traveled with new sensor (mm)
(0; 0)	(600; 500)	(350; 350)	824,2705	809,5474
(0; 0)	(600; 500)	(350; 400)	804, 8216	801,7906
(0; 0)	(600; 600)	(350; 350)	865,8441	855,9001
(0; 0)	(600; 600)	(350; 350)	873,6312	868,0743
(0; 0)	(600; 550)	(350; 400)	851,2306	834,8123
(0; 0)	(700; 600)	(350; 400)	952,1903	933,9707



(a) (b)

Fig. 12. (a): Represents the trajectory optimization regarding an obstacle set at (100, 100) and a target (200, 300), while (b) Denotes the trajectory optimization concerning an obstacle set at (100, 100) and a target (300, 400).

Table 2 depicts the gradient method as a rather efficient as compared to the trajectory optimization related genetic algorithm.

6 A Practical Implementation Case

The experiments implemented have been aimed to validate the obstacle avoidance applied approach. The mobile robot Khepera II departs from an initial configuration to reach a predefined goal, with a wooden cube being placed in front of it, which it has been able to detect by means of infrared sensors.

The following (Fig. 13) shows well that the robot khepera II has been capable of detecting the obstacle and avoiding it.

Table 2. Distances traveled with and without optimization

Starting point (mm)	Target (mm)	Obstacle position (mm)	Distance traveled without optimization (mm)	Distance traveled with optimization (Gradient method)	Distance traveled with optimization (Genetic algorithm)
(0,0)	(300,400)	(100,100)	527,3838	507,093	513,5699
(0,0)	(300,300)	(100,100)	441,3619	436,7311	440,9823
(0,0)	(200,300)	(100,100)	404,8132	364,0091	379,0217
(0,0)	(200,400)	(150,200)	458,0796	451,1089	455,1156
(0,0)	(100,200)	(100,100)	238,5747	225,5232	230,4522

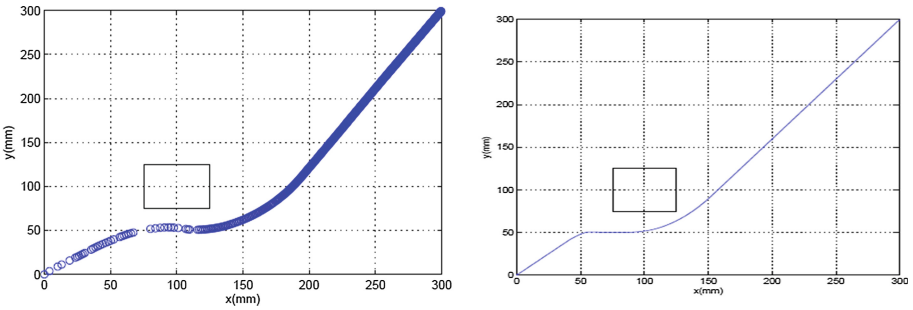


Fig. 13. (a): Experimental Results regarding the obstacle avoidance process; (b): Simulation results concerning the obstacle avoidance process.

The results provided by both figures prove to be perfectly an significantly similar.

7 Conclusion

In the present work, a particular focus has been laid on investigating the obstacle-avoidance problem by means of fuzzy systems. In a first place, the Khepera II robot specific sensors have been deployed. Noteworthy, however, sudden changes have been noticed to prevail in the robot's both wheels' speeds. For this reason, a new extra sensor has been implemented with a relatively larger distance range. As a matter of fact, while the mobile robot's navigation and obstacle avoidance features have been based on the use of two controllers, monitoring and directing the robot, concerning the case in which no obstacle exists on the way to the goal, the additional controller has been devised to avoid any obstacle likely to occur on its path. Actually, the experimental implementation of a fuzzy controller in the mobile robot has led to a remarkable validation of the achieved simulation results. It is worth highlighting, in this respect that as the curves obtained is liable to optimization Distances traveled with Gradient

method and Genetic algorithm. In addition, we envisage elaborating on comparation in two methods; the Gradient method is advantage in Genetic algorithm.

References

1. Ye, J.: Tracking control of a nonholonomic mobile robot using compound cosine function neural networks. Intell. Serv. Robot. (2013). doi:10.1007/s11370-013-0136-4
2. Cheng, P.-Y., Chen, P.-J.: Navigation of mobile robot by using D++ algorithm. Intell. Serv. Robot. **5**, 229–243 (2012). doi:10.1007/s11370-012-0120-4
3. Schiffer, S., Ferrein, A., Lakemeyer, G.: CAESAR: an intelligent domestic service robot. Intell. Serv. Robot. **5**, 259–273 (2012). doi:10.1007/s11370-012-0118-y
4. Ryu, H., Chung, W.K.: Local map-based exploration for mobile robots. Intell. Serv. Robot. **6**, 199–209 (2013). doi:10.1007/s11370-013-0137-3
5. Hossain, A.: Undefined Obstacle Avoidance and Path Planning, American Society for Engineering Education (2013)
6. Hermawanto, D.: Genetic Algorithm for Solving Simple Mathematical Equality Problem. Indonesian Institute of Sciences (LIPI), Indonesia (2013)
7. Vandenberghe, L.: «Gradient method», EE236C (Spring 2016)
8. Bayle, B.: "mobile robotic", Project at the National School of Physics of Strasbourg (2008)
9. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. IEEE Trans. Syst. Man Cybern. **15**, 116–132 (1985)

The Improvement of an Image Compression Approach Using Weber-Fechner Law

Mourad Rahali[1,2(✉)], Habiba Loukil[1], and Mohamed Salim Bouhlel[1]

[1] Sciences and Technologies of Image and Telecommunication, High Institute of Biotechnology, University of Sfax, Sfax, Tunisia
rahali.mourad@yahoo.fr, loukil_habiba@net-c.com, medsalim.bouhlel@enis.rnu.tn
[2] National Engineering School of Gabes, University of Gabes, Gabès, Tunisia

Abstract. The idea behind image compression is the reduction of the average number of bits per pixel needed for its representation. Indeed, our study of still image compression is based on the non-conservative compression method. This means that the reconstructed image, after a compression/decompression cycle, will be different from the original one. In fact, such a difference brings about a degradation of the original image. In the first part of still image compression, we suggest a chain composed of a discrete wavelet transform followed by neural-networks quantization and a binary encoder. In this paper, we improve the quality of compression by adding a pretreatment phase through the use of the principle of Weber-Fechner law which considers the human-eye sensitivity to luminance as a logarithmic function.

Keywords: Weber-Fechner law · Logarithmic quantization · Discrete wavelet transform · Neural-networks

1 Introduction

The approaches using the artificial neural networks [1, 9] for data intelligent processing seem to be very promising. Essentially, this is due to their structures which provide the opportunities for parallel computations and for the use of the learning process which allows the network to adapt to the data to be processed [2]. Again, the discrete wavelet transform [10, 11] can produce different size sub-images and can identify the relevant information from the details of an image. To improve the compression quality, we use a pretreatment step which employs the principle of Weber-Fechner law [4, 12] for the logarithmic quantization of the original image. In this article, we disclose the extent to which it is important to use the logarithmic quantization in the improvement of the compression quality by taking the wavelet transform and Kohonen's networks [3] as a basis. To show how important the compression law is, we will, first, study the principle of Weber-Fechner law in quantifying the original image. Second, we will detail our approach to the compression of still images. Finally, we will compare the two approaches so as to assess our new method.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_24

2 The Proposed Image Compression Approach

Our new image compression approach consists of three steps: the quantization of the original image by Weber-Fechner law, the discrete wavelet transform and the quantization by Kohonen's network. Figure 1 elucidates the steps of compression by means of our approach.



Fig. 1. Image compression steps

2.1 Weber-Fechner Law

Sensitivity to contrast [12, 13] is the ability of the human visual system to detect the changes in luminance (achromatic) and the chromatic changes. Any measure of sensitivity to contrast depends on the level of luminance of the stimuli, their spatial frequencies, and their chrominance as well as the human observer adaptation level [5]. Indeed, contrast allows measuring the relative change in luminance with respect to the neighborhood. This property is known as Weber-Fechner law:

$$C^w = \frac{\Delta L}{L} \tag{1}$$

Where ΔL is the difference in luminance between the stimulus and its neighborhood and L is the luminance of the neighborhood. The Contrast Sensitivity Functions (CFS) [13] is typically used to quantify these dependencies (Fig. 2).

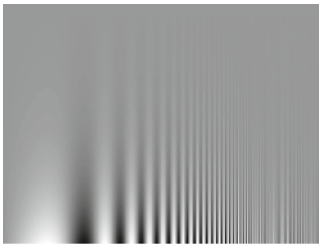


Fig. 2. Contrast sensitivity functions

Weber law regards the sensitivity of the human eye to luminance as a logarithmic function [6, 18]. Weber developed a quantitative description of the relationship between the stimulus intensity and the discrimination which is now known as Weber law.

$$\frac{\Delta S}{S} = K \tag{2}$$

Where ΔS is the perceived intensity difference with respect to a stimulus S and K which is a constant. Fechner-Weber applied the law to the sensory experience. He found that the intensity of sensation is proportional to the logarithm of the stimulus intensity.

$$\Delta S = k * S \Rightarrow \Delta P = k' \tag{3}$$

$$\Delta P = k' = k * \frac{\Delta S}{S} = k * \Delta(\log S) \tag{4}$$

$$\Delta P = P_1 - P_0 = k * \log(\frac{S_1}{S_0}) \tag{5}$$

If the absolute threshold is S0 = 1 and the associated sensation is P0 = 0, Fechner assumes that the amplitude of sensation is proportional to the logarithm of the stimulus. This relationship is called Weber-Fechner law.

$$P = k * \log(S) \tag{6}$$

Thus, according to Weber-Fechner law, the quality measurement must take into account the logarithmic sensitivity of the eye to light and the decreases in the image gray levels are considered imperceptible to the human eye [4]. On the other side of the coin, the logarithmic quantization reduces the entropy of the image which will lead to a higher compression ratio without noticeable degradation of the original image. We use the principle of Weber-Fechner law to quantify the image signal S (Fig. 3) as a pre-treatment phase before the wavelet transform.

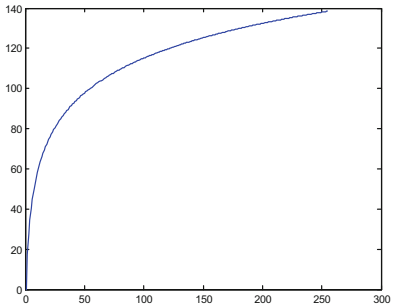


Fig. 3. Logarithmic quantization of the original image

2.2 Discrete Wavelet Transformation

The dimensional discrete wavelet transformation (DWT) [10] is based on the image multi-resolution analysis concept [14]. It is divided into a set of sub bands representing the information carried by the source image at different levels of resolution: the image approximation (LL) and the images of the horizontal details (HL), the vertical details (LH), the diagonal details (HH) [11]. The procurement of those pictures is the result of the following steps (Fig. 4):

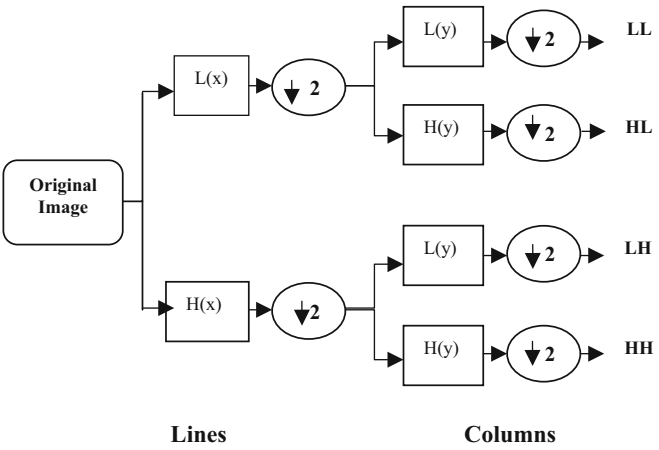


Fig. 4. Decomposition steps of an image by the discrete wavelet

Figure 5 is an example of the wavelet transformation of the Lena image at two levels of decomposition.



Fig. 5. Representation of a wavelet transform of the Lena image at 2 levels

2.3 Kohonen's Network

Kohonen's network came into existence in 1990 (Self Organizing Map) [3, 15]. It is a competitive network. It is made up of a layer of neurons reflecting the input-data presented to the network and a topological map. Self organizing map is a neural network in a single layer in which the neurons are usually arranged on a matrix. Each of these neurons is connected to the input via variable synaptic weights. The general structure of Kohonen's network is illustrated in Fig. 6. When an input is presented to

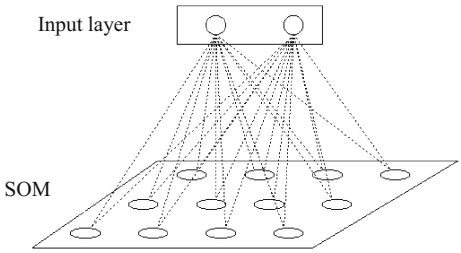


Fig. 6. Self organizing map

the network via the input layer, the neurons of the self organizing map are activated differently and a competition starts between these neurons. The neuron with the closest weight value to that of the input value is deemed the winner [17].

The adaptation of the network weight to learning is done according to the following rules:

1. Find the winner neuron in the competition as:

$$d(X, w_c) \leq d(X, w_i), \forall i \neq c \tag{7}$$

Where
X: input vector
w_c: the weight vector of the winner neuron c
w_i: the weight vector of the neuron i

2. Update weight w_i of the network:

$$w_i(t+1) = w_i(t) + h(c, i, t) * [X - w_i(t)] \tag{8}$$

where $w_i(t)$ is the weight vector of the neuron i at time t, h is a function defined by:

$$h(c, i, t) = \begin{cases} \alpha(t), i \in N(c, t) \\ 0, \text{si non} \end{cases} avec \alpha(t) \in [0, 1] \tag{9}$$

The function h defines the magnitude of the correction made to the winner neuron c and its neighborhood. The neighborhood, at time t, of the winner neuron c is determined by the function $N(c, t)$ which is a decreasing function over time. The final neighborhood of a neuron consists of the neuron itself. The function $h(c, i, t)$ allows assigning the same correction $\alpha(t)$ to all the neurons belonging to the neighborhood of the winner neuron at time t.

3 Image Compression by the Proposed Approach

To compress the gray level image, we first apply a logarithmic quantization to the original image using Weber-Fechner law. Second, we apply a discrete wavelet transformation to the quantified image according to the decomposition level (j) and the type of wavelet. Then, we cut the three detail sub-images (LH, HL, and HH) in a block (BS) in a definite block width (e.g. 4×4, 8×8 or 16×16) by keeping the approximation sub-image (LL). Third, we look for the codebook for each block representing the code-word with the minimum distance from the block. The index of the selected word is added to the index vector, and then we encode this vector by Huffman coding [7, 8]. The last step is to save the index vectors coded in Huffman and the approximation sub-image representing the compressed image.

4 Experimental Study and Results

In our work, we compare two compression methods: the classic method and the new method using Weber-Fechner law. We change the compression parameters: the wavelet decomposition level (j), the size of the input block (BS) and the size of Self Organization map (SOM). To evaluate our new approach, we use three compression evaluation criteria: the compression ratio (CT), the means square error (MSE) and the Peak Signal to Noise Ratio (PSNR).

- The compression ratio gives a measurement of the performance of the compression methods of still images. Also, it is an evaluation criterion for the compression algorithms. The compression ratio is defined by:

$$TC = \left(1 - \frac{k'}{k}\right) * 100 \tag{10}$$

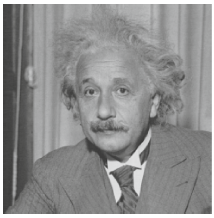
  where TC: the compression ratio, k': the number of bits per pixel in the compressed image and k: the number of bits per pixel in the original image (k = 8 for the images in gray level).
- MSE (Means Square Error), in the context of the compression method with losses, can measure the quality of the reconstructed image with respect to the original image.

$$MSE = \frac{1}{m * n} \sum_{i=1}^{m} \sum_{j=1}^{n} (I(i,j) - J(i,j))^2 \tag{11}$$

  where m and n represent the dimensions of the image, I and J stand for the original image as well as the reconstructed image.
- PSNR (Peak Signal to Noise Ratio) measures the signal to noise ratio. The signal represents the original image, whereas the noise represents the difference between the original image and the image reconstructed after compression.

$$PSNR = 10 * \log_{10}\left(\frac{(2^n - 1)^2}{MSE}\right) \tag{12}$$

The following tables correspond to the tests carried out on Einstein, House, Boat and Lena gray level images of the size 512*512 pixels (Fig. 7). We use the Haar wavelet transform because we show in [16] it is better for the reconstructed image quality by the comparison to other types of wavelets.


Einstein.bmp


House.bmp


Boat.bmp


Lena.bmp

Fig. 7. Original Images.

By using the same compression parameters, the analysis of the Tables 1, 2, 3 and 4 show that the compression ratio (the number of bits per pixel) of the image compressed by the new approach is improved with respect to the former approach. In this work, we present a subjective comparison between the synthetic images compressed by the classic approach and our new approach which utilizes Weber-Fechner law. Both compressions are made at the same compression ratio.

Table 1. Experimental compression results of einstein image

Classic approach				Weber-Fechner approach				Parameters
Nbpp	CR(%)	MSE	PSNR	Nbpp	CR(%)	MSE	PSNR	
4.490	43.87	34.627	32.736	2.957	63.02	34.240	32.785	J = 1; BS = 2; SOM = 16
3.879	51.51	34.390	32.766	2.612	67.34	33.637	32.862	J = 1; BS = 2; SOM = 8
2.606	67.42	41.796	31.919	1.770	77.86	43.441	31.751	J = 1; BS = 4; SOM = 16
2.255	71.81	42.927	31.803	1.554	80.56	45.755	31.526	J = 1; BS = 4; SOM = 8
1.818	77.27	44.37	31.659	1.132	85.84	55.845	30.660	J = 1; BS = 4; SOM = 2
1.624	79.70	44.803	31.617	0.988	87.64	56.049	30.645	J = 1; BS = 16; SOM = 2
1.237	84.53	96.691	28.276	0.977	87.78	102.899	28.006	J = 2; BS = 4; SOM = 8
0.638	92.25	103.63	27.975	0.438	94.51	115.736	27.496	J = 2; BS = 8; SOM = 8
0.487	93.91	104.61	27.934	0.299	96.25	116.937	27.451	J = 2; BS = 16; SOM = 8
0.438	94.52	103.32	27.988	0.266	96.66	122.282	27.257	J = 2; BS = 16; SOM = 2
0.138	98.27	188.588	25.375	0.092	98.84	219.585	24.714	J = 3; BS = 16; SOM = 2

Table 2. Experimental compression results of house image

Classic approach				Weber-Fechner approach				Parameters
Nbpp	CR(%)	MSE	PSNR	Nbpp	CR(%)	MSE	PSNR	
4.396	45.05	95.531	28.329	2.592	67.59	96.560	28.282	J = 1; BS = 2; SOM = 16
3.716	53.55	100.827	28.095	2.251	71.861	98.822	28.182	J = 1; BS = 2; SOM = 8
2.607	67.41	111.318	27.665	1.707	78.66	119.358	27.362	J = 1; BS = 4; SOM = 16
2.284	71.45	112.99	27.600	1.447	81.90	123.361	27.219	J = 1; BS = 4; SOM = 8
1.826	77.17	115.71	27.497	1.069	86.63	139.326	26.690	J = 1; BS = 4; SOM = 2
1.667	79.16	116.372	27.472	0.954	88.07	144.335	26.537	J = 1; BS = 16; SOM = 2
1.188	85.15	295.88	23.419	0.852	89.34	297.751	23.392	J = 2; BS = 4; SOM = 8
0.642	91.97	327,871	22.974	0.427	94.65	336.749	22.857	J = 2; BS = 8; SOM = 8
0.495	93.81	345,151	22.751	0.307	96.15	344.926	22.753	J = 2; BS = 16; SOM = 8
0.448	94.40	367,421	22.479	0.276	96.55	363.384	22.527	J = 2; BS = 16; SOM = 2
0.141	98.23	528.691	20.898	0.094	98.81	654.78	19.969	J = 3; BS = 16; SOM = 2

Table 3. Experimental compression results of boat image

Classic approach				Weber-Fechner approach				Parameters
Nbpp	CR(%)	MSE	PSNR	Nbpp	CR(%)	MSE	PSNR	
4.864	32.20	71.972	29.559	3.161	60.47	67.74	29.822	J = 1; BS = 2; SOM = 16
4.025	49.68	76.681	29.283	2.777	65.28	72.495	29.527	J = 1; BS = 2; SOM = 8
2.703	66.21	87.312	28.720	2.043	74.46	92.124	28.487	J = 1; BS = 4; SOM = 16
2.374	70.32	88.727	28.650	1.687	78.90	97.617	28.235	J = 1; BS = 4; SOM = 8
1.964	75.45	92.864	28.452	1.182	85.21	125.068	27.159	J = 1; BS = 4; SOM = 2
1.742	78.22	93.392	28.427	1.099	86.25	113.826	27.568	J = 1; BS = 16; SOM = 2
1.252	84.35	222.975	24.648	1.057	86.78	241.295	24.305	J = 2; BS = 4; SOM = 8
0.666	91.67	233.746	24.443	0.496	93.79	268.219	23.845	J = 2; BS = 8; SOM = 8
0.517	93.53	235.984	24.402	0.352	95.59	269.455	23.825	J = 2; BS = 16; SOM = 8
0.468	94.15	232.396	24.468	0.312	96.08	278.112	23.688	J = 2; BS = 16; SOM = 2
0.144	98.20	418.641	21.912	0.105	98.67	492.372	21.207	J = 3; BS = 16; SOM = 2

Table 4. Experimental compression results of lena image

Classic approach				Weber-Fechner approach				Parameters
Nbpp	CR(%)	MSE	PSNR	Nbpp	CR(%)	MSE	PSNR	
4.755	40.56	34.60	32.739	2.763	65.45	34.885	32.704	J = 1; BS = 2; SOM = 16
3.958	50.52	36.902	32.460	2.399	70.008	37.253	32.419	J = 1; BS = 2; SOM = 8
2.673	66.58	43.062	31.789	1.790	77.616	46.277	31.477	J = 1; BS = 4; SOM = 16
2.312	71.11	44.289	31.667	1.494	81.32	48.185	31.301	J = 1; BS = 4; SOM = 8
1.892	76.35	46.754	31.432	1.062	86.71	54.738	30.747	J = 1; BS = 4; SOM = 2
1.677	79.03	47.458	31.367	0.937	88.28	56.842	30.584	J = 1; BS = 16; SOM = 2
1.258	84.27	129.087	27.022	0.915	88.55	135.757	26.803	J = 2; BS = 4; SOM = 8
0.657	91.77	135.75	26.803	0.444	94.44	155.445	26.215	J = 2; BS = 8; SOM = 8
0.507	93.66	138.216	26.725	0.308	96.149	156.748	26.178	J = 2; BS = 16; SOM = 8
0.464	94.21	137.228	26.756	0.273	96.58	162.486	26.022	J = 2; BS = 16; SOM = 2
0.151	98.11	289.429	23.515	0.097	98.77	340.524	22.809	J = 3; BS = 16; SOM = 2

The reconstructed images are compared with the original images in terms of PSNR according to the number of bits per pixel (Nbpp). Actually, Figs. 8, 9, 10 and 11 indicate that the visual quality metric (PSNR) of the compressed images of the approach using Weber-Fechner law (green curve) in terms of the compression ratio (Nbpp) improves with respect to the classic approach (blue curve). We notice that there is a significant improvement in the quality of the reconstructed image if the compression ratio is higher than 1 bits per pixel since the shape of the green curve (new approach) is going up for the four images. Also, there is a little improvement if the compression ratio is less than 0.5 bits per pixel. According to the four figures, we notice a significant drop in the image quality metric (30.6 to 28 for the image Instein, 26.5 to 23 for the image House and 27.5 to 23.3 for the image Boat) if the number of bits per pixel (NBPP) is between, approximately, 1 and 1.7 and the second drop if the NBPP is less than 0.5 for the reconstructed images. This significant degradation of the image quality is due to the change of the wavelet transform level to another level.

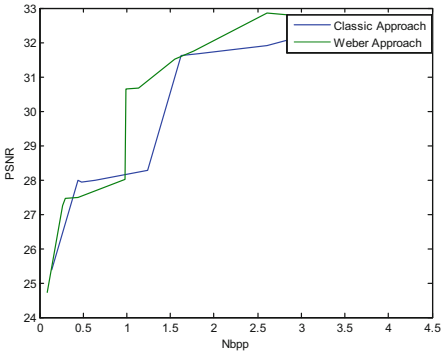


Fig. 8. PSNR = F(Nbpp) of Einstein image

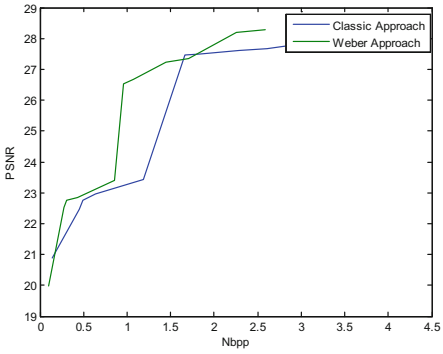


Fig. 9. PSNR = F(Nbpp) of house image

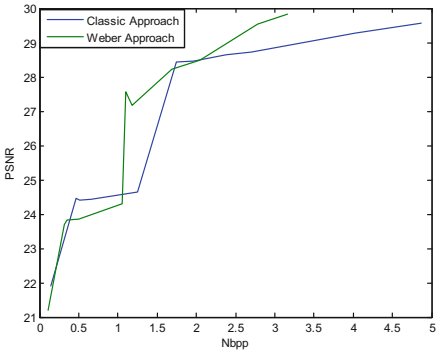


Fig. 10. PSNR = F(Nbpp) of boat image

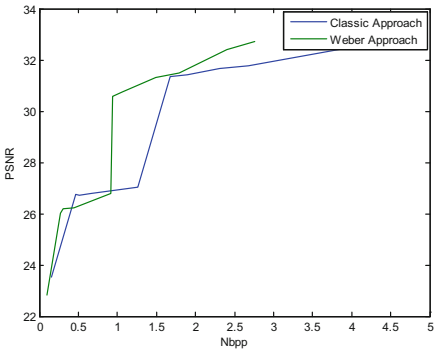


Fig. 11. PSNR = F(Nbpp) of lena image

5 Conclusion and Future Scope

In this paper, we have enhanced the compression quality of a still-image-compression approach based on the discrete wavelet transformation and neural networks by adding a new pre-treatment phase to quantify the signals of the original image through the use of the principle of Weber-Fechner law. We notice that the new approach is better than the classic approach in terms of quality according to the compression ratio feature if the number of bits per pixel is higher than 0.5. To boost our approach, we will quantify the original image in a semi-logarithmic manner by using the law of compression 'A' which is used in mobile telephone networks.

References

1. Gupta, P., Alam, M.S.: Advancement of neuro-wavelet network over neural network in image compression. Int. J. Adv. Res. Comput. Sci. Manag. Stud. **2**(11), 405–411 (2014)
2. Khashman, A., Dimililer, K.: Image compression using neural networks and Haar wavelet. WSEAS Trans. Sig. Process. **4**(5), 330–339 (2008)
3. Kohonen, T.: The self-organizing map. Proc. IEEE **78**(9), 1464–1480 (1990)
4. Dehaene, S.: The neural basis of the Weber – Fechner law: a logarithmic mental number line. TRENDS in Cogn. Sci. **7**(4), 145–147 (2003)
5. Xia, Q., Li, X., Zhuo, L., et al.: Visual sensitivity-based low-bit-rate image compression algorithm. IET Image Process. **6**(7), 910–918 (2012)
6. Loukil, H., Kacem, M.H., Bouhlel, M.S.: A new image quality metric using system visual human characteristics. Int. J. Comput. Appl. **60**(6) (2012)
7. Pujar, J., Kadlaskar, L.: A new lossless method of image compression and decompression using huffman coding techniques. J. Theoret. Appl. Inf. Technol. **15**(1/2), 18–23 (2010)
8. Mathur, M.K., Loonker, S., Saxena, D.: Lossless Huffman coding technique for image compression and reconstruction using binary trees. Int. J. Comput. Technol. Appl. **3**(1), 76–79 (2012)
9. Hore, S., Bhattacharya, T., Dey, N., Hassanien, A.E., Banerjee, A., Bhadra Chaudhuri, S.R.: A real time dactylology based feature extractrion for selective image encryption and artificial neural network. In: Awad, A.I., Hassaballah, M. (eds.). SCI, vol. 630, pp. 203–226. Springer, Heidelberg (2016). doi:10.1007/978-3-319-28854-3_8
10. Boopathi, G.: An image compression approach using wavelet transform and modified self organizing map. Int. J. Comput. Sci. Issues **8**(5), 323–330 (2011)
11. Lewis, A.S., Knowles, G.: Image compression using the 2-D wavelet transform. IEEE Trans. Image Process. **1**(2), 244–250 (1992)
12. Aydin, T., Mantiuk, R., Seidel, H.: Extending quality metrics to full luminance range images. In: Human Vision and Electronic Imaging, February 2008
13. Pelli, D.G., Bex, P.: Measuring contrast sensitivity. Vis. Res. **90**, 10–14 (2013)
14. Vasuki, A., Vanathi, P.T.: Image compression using lifting and vector quantization. Graph. Vis. Image Process. J. GVIP, ICGST **7**(1), 7–15 (2007)
15. Zribi, M., Boujelbene, Y., Abdelkafi, I., Feki, R.: The self-organizing maps of Kohonen in the medical classification. In: 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, SETIT 2012, 21–24 March 2012, Sousse-Tunisia, pp 852–856 (2012)

16. Rahali, M., Loukil, H., Bouhlel, M.S.: Theoretical improvement of the image compression method based on wavelet transform. In: 13th International Conference Computer Graphics, Imaging and Visualization (CGIV), Morocco, April 2016, pp 273–277 (2016)
17. Ribes, J.M.M., Simon, B., Macq, B.: Combined Kohonen neural networks and discrete cosine transform method for iterated transformation theory. Sig. Process. Image Commun. **16**(7), 643–656 (2001)
18. Ma, J., Duan, F., Guo, P.: Improvement of texture image segmentation based on visual model. In: 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), Tunisia, March 2012, pp 151–154 (2012)

A Minimal Rare Substructures-Based Model for Graph Database Indexing

Mehdi Azaouzi(✉) and Lotfi Ben Romdhane

Modeling of Automated Reasoning Systems (MARS) Research Lab,
Higher Institute of Computer Science and Telecom (ISITCom),
University of Sousse, Sousse, Tunisia
mehdi.azaouzi@gmail.com, lotfi.ben.romdhane@gmail.com

Abstract. Systems such as proteins, chemical compounds, and the Internet are stored as graph structures in graph databases. A basic, common problem in graph related applications is to find graph data that contains a query. It is not possible to scan the whole data in graph databases since subgraph isomorphism testing is an NP-complete problem. In recent years, some effective graphs indexes have been proposed to first obtain a candidate answer set and then performing verification on each candidate by checking subgraph isomorphism. However, candidate verification is still inevitable and expensive when the size of the candidate answer set is large. In this paper, we propose a new Structural Graph Indexing, called GIRAS, based on RAre subGraphs (RGs) as the basic indexing feature. The idea is to have a single characteristic that can uniquely identify a graph in a database. Few substructures are ideal candidates since they are rare graphs, which means they occurs in only a small number of graphs in the database. Thus, in confronting a query using these indexes, the size of the candidate answer set is close to that of the exact answer set, and the number of subgraph isomorphism tests is small. Therefore, the time of the candidate verification step is reduced to a minimum.

Keywords: Graph databases · Graph indexing · Graph querying · Rare subgraphs · gSpan-Modified

1 Introduction

Graphs have become popular for modeling structured data. In particular, *graph databases* facilitate the applications needed in the data management organized as graphs. A basic operation is *the search* of a query graph in a target graph or, more generally, in a database of graphs. In many cases, the success of a graph database application is directly dependent on the efficiency of the graph query processing. Let $D = \{G_1, G_2, .., G_n\}$ be a graph database that contains n graphs. A search of

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_25

a query graph can be described as follows: given a graph Q, retrieve all $G_i \in D$ such that G_i is a supergraph of Q. The problem of searching for a query graph in a database of graphs appeals to the procedure *subgraph isomorphism*, known to be NP-complete [1]. Since the subgraph isomorphism test is expensive, checking all graphs of a large database can be unfeasible. Therefore, the solution is to index the graph database. The query processing using these indexes can be divided in two phases. In the first step (the preprocessing phase), we analyze the database of graphs and an index is built. The second step is the filtering-and-verification framework: the index is used to discard the graphs of the database which cannot contain the query, producing a small set of candidate set C_Q of graphs that may potentially contain the query (the filtering step). The set of candidates is then verified (the verification step) by a subgraph isomorphism algorithm and all the resulting matches are reported, in order to finally obtain the answer set D_Q. C_Q is typically larger than the answer set D_Q, $|C_Q| \geq |D_Q|$. A subgraph isomorphism test is performed to check that Q is contained in each candidate graph in C_Q. Thus, $|C_Q|$ subgraph isomorphism tests must be performed between C_Q and Q to eliminate graphs not containing Q to obtain the answer set D_Q. Naturally, the verification step is computationally more expensive since it requires a subgraph isomorphism.

In this paper, we explore this direction and we introduce GIRAS, a novel inverted index based on subgraphs with low frequency. A key feature of GIRAS is that it indexes a graph with the most discriminative rare structure. Intuitively, a rare structure is a subgraph with the minimal frequency that discriminates the indexed graph from the rest of the graph database.

2 Related Work

The differences among the various graph indexing systems lie mainly in the choice of the feature set whose main purpose is to minimize the set of candidates C_Q. Existing research works on graph indexing can be grouped into two categories. The *first* category is feature-based graph indexing which is divided into two classes; Path-based indexing methods [2–5] and Tree-based indexing methods [6–9]. The main advantages facing this category is that, if the choice of the feature is carried out correctly then it is more likely, that several queries are included in feature set. Thus, the answers are found directly. To keep this advantage, we must have a higher number of features to be indexed. The more possibility that they comprise a sufficiently discriminative feature set. Furthermore, the common challenge of feature-based indexing is the index size. The higher the size index, the longer will be filtering time. However, we must seek a balance between the sufficiently discriminative feature set and the size index. The *second* category is hierarchical indexing [10–16]. The main characteristic of these approach is the direct access to each graph in the database. This advantage corrects the problem of size index in feature-based indexing. They guarantee firstly the discriminative feature summarized in the direct access. Second, they prevent a large set of index by the hierarchical structure. Our algorithm GIRAS fits into the category of the

feature-based graph indexing. Our indexing feature is proved to circumvent the size index problem in the feature-based indexing and has the advantage of the hierarchical indexing systems for the direct access to each graph. This is detailed subsequently.

3 Our Model GIRAS

The indexing features in GIRAS[1] based on low-frequency. Indeed, in GIRAS, the basic idea of our model is to associate, if possible, *a unique index* to each graph in the base. Otherwise, we choose indexes with low frequency, called *rare substructures*. The kernel of GIRAS is the fixing of f_{min} as a threshold to determine rare subgraphs. The threshold f_{min} is calculated by the recursive calls of the algorithm *gSpan-Modified*. gSpan-Modified seeks substructures having a support equal to the values given $\{1, 2, ..\}$. In GIRAS an inverted index is built on graph database D. The construction of the index is composed of two main steps. In the first step, we search all substructures having a frequency equal to 1 and we index them as filtering features using inverted lists. Each entry in the inverted list contains the graph identity and the frequency. These substructures represent a unique index. In this stage, if the database is fully indexed using these substructures, then it leads to unique indexes for all graphs in the database which is our ultimate goal. Assume that all substructures of 1-$frequency$ have been indexed. Given a query graph Q, if Q has a 1-$frequency$, then the set of candidate answers C_Q is of small size and can be recovered directly since Q is indexed. In this way we reduce the time of the verification phase and thus the total response time. Otherwise, we look for the index containing Q as a subgraph. These indexes are 1-$frequency$ and the set of candidate answers $|C_Q|$ is of small size and the exact set of query answers is returned without performing candidate verification. One of the problems that can arise when using these substructures of frequency 1 is that we may have *incomplete* indexing, because not all graphs of the database contain subgraphs of frequency 1. This requires other substructures to index all the graphs. Therefore, we focus primarily on determining a threshold frequency. The frequency of index will be less than or equal to this threshold. This threshold represents the minimum frequency that can index our whole database. The use of the concept of "minimum frequency" lead to a new concept, we named "*rare minimum substructures*". Intuitively, the latter are subgraph with the smallest size and the least frequency. This is detailed subsequently.

3.1 Rare Minimum Substructures

In this subsection, we will discuss how GIRAS seeks to propose an approach that guarantees both a uniqueness constraint and a complete indexing of the base. The idea is to define a threshold and perform an indexing using all substructures with a frequency less than or equal to this threshold. This set of

[1] GIRAS is an acronym for Graph database Indexing based on the RAre Structures.

indices represents substructures of low frequencies. It contains substructures of frequency equal to 1. It also ensures a complete indexing of the base. Unfortunately, the determination of the value of this threshold is challenging and raises several fundamental questions. On one hand, how can we determine the lowest frequency that allows complete indexing of a database? On the other hand, how can we differentiate between substructures of low frequency and frequent substructures? This lead us to define a threshold that guarantees the resolution of these issues. This threshold is defined as the minimum frequency which can index the database. A new threshold value which guarantees the indexing of the database graphs using substructures of frequency 1 if it is possible. Otherwise, we seek to find the minimum frequency that provides a complete indexing to the base. To do this, we will first recall what is an index of a base; and second, we will define a threshold which indicates this minimum frequency.

Definition 1. *We say that* $I = \{I_1, I_2, .., I_i\}$ *is indexing a graph database* $D = \{g_1, g_2, .., g_n\}$ *if and only if* $\forall g_k \in D$; $\exists I_j \in I$ *such as* $I_j \subseteq g_k$.

Definition 2. *Let* D *be a graph database. The minimum frequency that allows indexing* D *is the smallest integer* f_{min} *such that there is a set of index* I *of* D, $\forall I_k \in I$; *frequency*$(I_k) \leq f_{min}$.

The principle of the determination of f_{min} is given as follows: given a graph database D, we first initialize f_{min} to 1. Then, we index D using the substructures of frequency equal to 1. Afterward, we test if the base is fully indexed, if it is not the case then we increase the value of f_{min} by 1. This process is repeated until the base is completely indexed. It eventually leads to a reasonable and moderate threshold value f_{min}, providing an index of the database with substructures having the lowest possible frequency. In the best case, f_{min} is equal to 1, and we guarantee the uniqueness constraint previously presented and a complete indexing. We will adapt these tools to define the rare substructures used as index. For this, we define the concept of rare subgraphs as follows.

Definition 3. *Let* $D = \{g_1, g_2, .., g_n\}$ *be a graph database that contains a set of* N *graphs and* f_{min} *the minimum frequency that indexes* D. *A substructure* pr *is said rare relative to* D *if frequency*$(pr) \leq f_{min}$.

Our choice of the rare substructures will be based on the frequency value. In the best case, all indexes are frequency equal to 1, and in other cases the index may have frequency values very close to 1. At this stage, all supergraphs of a substructure with frequency f, are a substructures with a frequency less than or equal to f. That is to say, if g contains pr_x as an index and pr_x is rare subgraph, these supergraphs $pr_{1x}, pr_{2x}, .., pr_g$, ($pr_g$ is the graph g itself) are also rare substructures and will be included in the index set. Therefore, as soon as we find a rare substructure then all of these supergraphs are also rare substructures. Hence the choice of these rare substructures can lead to an exponential number of rare substructures. This problem is circumvented as follows. First, we see in the selection of rare substructures that among the similar substructures with the same support, it is often sufficient to index the smallest common substructure;

i.e., if pr' (a supergraph of pr) has the same support as pr, it will not be able to provide more information than pr if both are selected as indexing features. Thus pr' should be removed from the feature set. It is better to include only one of them in the feature set. Generally, among similar substructures with the same support, it is often sufficient to index only the smallest common substructure since more query graphs may contain the smallest substructure. In this case, we say that pr' is excessive and there is no need to include it in the rare subgraphs set. In contrast, there are substructures which are not excessive, called *the rare minimum substructure*.

Definition 4. *We consider a graph database* $D = \{ g_1, g_2, .., g_n\}$ *and* f_{min} *the minimum frequency that indexing* D*. A substructure* pr *is said rare minimum relative to* D *if frequency(*pr*)* $\leq f_{min}$ *and if no other substructure* pr' *exists such that:*

- $pr' \subset pr$.
- *frequency(*pr'*)* = *frequency(*pr*).*

So, for each selected frequency, we compute the smallest substructures having this frequency value. This is achieved by a modified gSpan [17] as we will see in the next section.

3.2 gSpan-Modified

Computing graph substructures whose frequency is above a specified threshold is a common task in graph mining and deeply studied in the research literature. One of the most known algorithm is gSpan [17]. However, computing graph substructures whose frequency is exactly equal to a given value is not a trivial problem. The main purpose of this section is to a propose a modified gSpan algorithm for computation of graph substructures with a specified frequency.

A preliminary version of *gSpan-Modified* was already published in [18]. The details of *gSpan-Modified* are represented in Algorithm 1. gSpan-Modified tests if s is a minimum DFS code pattern to eliminate duplicate subgraphs and their descendants (line 2–3). In fact, extensions are only allowed on minimal DFS codes to avoid computation of duplicate graphs. We compute the frequency value of s, if it is equal to f, then gSpan-Modified stops the extension (line 7) and returns s as a substructure with the frequency f (line 6) and gSpan-Modified investigates other candidates for expansion (line 12). In this case, s does not contain a subgraph with a frequency f. However, as we show in Sect. 3.1, each new selected subgraph with a frequency f is a minimum rare substructure and it does not contain a subgraph with frequency f. Unfortunately, gSpan-Modified does not guarantee that a graph with the requested frequency that is generated by extending a graph with a higher frequency is not already contains a subgraph generated by the extension of another graph. However, in order to guarantees the criterion of the minimum rare substructure, soon as *gSpan-Modified* found the fragments with the requested frequency, it verifies that there are these subgraphs already found in advance (line 5, second test).

Algorithm 1. gSpan-Modified(s, D, f, S)

Data : A DFS code **s**; a graph data set **D**; the frequency **f**; and A graph set **S**
Result : A graph set **S** with frequency **f**

```
1  begin
2      if s ≠ min(s) then
3          return ;
4      end
5      if support(s) = f and (∄ s' ∈ S ∧ s' ⊂ s) then
6          insert s into S ;
7          return ;
8      end
9      C ⟵ ∅;
10     Scan D once, find all the edges e such that s can be right-most extended to
         s.e; insert s.e into C and count its frequency ;
11     Sort C in DFS lexicographic order ;
12     foreach s.e in C do
13         if support(s) ≥ f then
14             gSpan-Modified(s.e, D, f, S) ;
15             return S;
16         end
17     end
18 end
```

gSpan-Modified is recursively called to extend the graphs with a frequency higher than f and it stops either when the support of a new graph is less than f (line 13). In this case, *gSpan-Modified* stops searching without return s and without extending since any supergraph of a graph with frequency lower than f is also with frequency lower than f. If the frequency of s is greater than f, we must extend s, and we make a recursive call to each supergraph of s. This extension stops when we discover a model with a frequency equal to f (line 5–8), this model will be a minimum rare subgraph with a frequency f and it does not contains subgraph with this frequency value. We must ensure that our model does not generate an beforehand a subgraph s' of s, This is done using a simple verification (line 5). In this way, all the substructures will be returned with a frequency f and they are minimal. With the algorithm gSpan-Modified at hand, now we are able to compute graph substructures for a database D with a specified frequency f. In addition, the computed subgraphs are of minimum size; i.e. rare minimum substructures. Now, we are ready to outline our proposal for computing the index using minimum rare substructures for a given graph database.

3.3 Our Algorithm GIRAS

In our algorithm GIRAS is described in Algorithm 2. Initially, we start with an unindexed database D and a frequency equal to 1 (line 2). Initially we call

the algorithm *gSpan-Modified* (line 5). The algorithm gSpan-Modified returns all minimal rare substructures with a frequency equal to 1. These substructures may satisfy the complete indexing of D. To check the full indexing of D, we delete all graphs of the database containing the selected substructures (line 7–8). If these is still unindexed graphs (line 10), then we increment the frequency value by one (line 9). Subsequently, a second call is done on gSpan-Modified to determine the rare minimum substructure with frequency 2. We increase the frequency for the chance to index the graphs that are not indexed yet. We repeat the same steps as there are graphs D which are not indexed yet (line 10). Finally, we return the substructures of the lowest frequency that can index D; and we obtain the value of the minimum frequency f_{min} (defined in Sect. 3.1).

Algorithm 2. GIRAS

Data : A graph data set **D**
Result : A rare minimum substructures set **R**

1 **begin**
2 $R = \{pr_{\emptyset}\}$, $D_c = D$ and frequency $= 1$;
3 **repeat**
4 $S \longleftarrow \emptyset$;
5 $S \longleftarrow$ **gSpan-Modified**(s,D,frequency,S) ;
6 $R \longleftarrow R \cup S$;
7 $D_s \longleftarrow \{g \supseteq S \mid g \in D_c\}$;
8 $D_c \longleftarrow D_c \setminus D_s$;
9 frequency $\longleftarrow$ frequency $+ 1$;
10 **until** $(D_c = \emptyset)$;
11 **return** R;
12 **end**

Summarizing, we have presented in this section our model GIRAS for indexing graph databases using rare substructures which have high discriminative power and minimal size. Thereafter, these rare substructures are organized into an index tree using DFS coding. Using this index tree, user queries can be processed efficiently. In order to assess the efficiency of our proposal, we will conduct an experimental analysis in the next section.

4 Simulation Results

We performed a series of experiments the performance of our proposed model GIRAS. As a testbed, we considered both real and synthetic graph databases. As a real database, we considered AIDS Antiviral dataset[2] (see Table 1). We considered also, synthetic datasets generated by GraphGen[3]. For querying the

[2] Aids antiviral dataset used for igraph. http://urlm.co/www.igraph.or.kr.
[3] Graphgen: a synthetic graph data generator. http://www.cse.ust.hk/graphgen/.

Table 1. Characteristics of the used database

Graph database	#graph	#vertices	#edges	#vertex labels	#edge labels
AIDS	10000	25.42	27.40	51	4
Synthetic.10K.E30.D5.L50	10000	12.45	30.00	50	50
Synthetic.10K.E30.D5.L20	10000	12.45	30.00	20	20

database using the constructed index, we considered a set of standard queries $Q = \{Q_4, Q_8, Q_{12}, Q_{16}, Q_{20}, Q_{24}\}$ used, approximately, in all indexing methods. Each query Q_n contains 1000 graphs and each graph in Q_n consists of n edges.

4.1 Evaluation on AIDS

Simulation results on the dataset graphs AIDS, are summarized in Tables 2 and 3. We can see that our model GIRAS significantly reduces the total response time. This is explained by the reduction of verification time made in GIRAS which is a little higher in the model *gIndex*. And also this reduction is clear in the case of the query with larger sizes. This is expected since large query have low frequency. On the other hand, experiments performed on real database *AIDS* shows that we can have rare substructures with small size. This expresses the effectiveness of GIRAS also for some small queries. We can also guarantee that GIRAS can recover most of the answers presented in the database D.

Table 2. The average number of candidates responses for each query Q_i for AIDS with size 10000

Query	*GIRAS*	*gIndex*	*FG-Index*	*Tree+Δ*	*C-Tree*
Q_4	2315.790	2821.890	2242.700	**2040.290**	2300.750
Q_8	**181.560**	291.117	812.085	326.671	447.990
Q_{12}	**19.770**	62.347	307.863	79.538	100.213
Q_{16}	**10.891**	22.233	142.129	26.002	28.530
Q_{20}	**4.663**	12.526	87.856	13.666	10.994
Q_{24}	**3.089**	7.407	57.779	7.919	4.941

4.2 Evaluation on Synthetic Datasets

As a summary to these simulations (in Tables 4 and 5), we can state that the variation in the number of distinct labels affects the indexing using rare substructures. GIRAS is better for a larger number of distinct labels because its pruning in this case is more effective. In fact, GIRAS finds the candidates in both cases

Table 3. Query Processing Time (in milliseconds) for AIDS of size 10000

Query	*GIRAS*	*gIndex*	*FG-Index*	*Tree+Δ*	*C-Tree*
Q_4	221.980	610.797	**102.203**	255.125	1187
Q_8	**80.710**	107.593	151.906	173.172	929
Q_{12}	**77.890**	100.953	100.453	120.781	720
Q_{16}	**62.342**	95.672	212.844	112.422	533
Q_{20}	**24.210**	37.375	94.297	98.687	364
Q_{24}	**22.580**	34.172	71.703	97.281	249

Table 4. The average number of candidates for Synthetic.10K.E30.D5.L20 and Synthetic.10K.E30.D5.L50

Query	*GIRAS*	*gIndex*	*FG-Index*	*Tree+Δ*	*C-Tree*	*GIRAS*	*gIndex*	*FG-Index*	*Tree+Δ*	*C-Tree*
Q_4	108.793	119.876	214.836	194.947	**106.986**	251.164	276.058	336.953	325.624	**207.992**
Q_8	9.023	11.095	47.325	21.012	**7.076**	96.212	111.162	175.405	124.721	**59.570**
Q_{12}	**1.513**	2.478	18.556	2.858	1.568	25.142	64.766	122.670	43.782	**23.261**
Q_{16}	**1.039**	1.243	7.578	1.092	1.053	**6.459**	44.393	91.243	10.904	8.959
Q_{20}	**1.002**	1.062	4.239	1.006	**1.002**	**2.240**	32.289	76.177	2.659	3.496
Q_{24}	**1.001**	1.012	2.623	1.002	**1.001**	**1.260**	26.250	64.355	1.287	1.716

Table 5. Query Processing Time in milliseconds ($T_{reponse}$) on Synthetic.10K.E30.D5.L20 and Synthetic.10K.E30.D5.L50

Query	*GIRAS*	*gIndex*	*FG-Index*	*Tree+Δ*	*C-Tree*	*GIRAS*	*gIndex*	*FG-Index*	*Tree+Δ*	*C-Tree*
Q_4	104.584	244.672	**99.672**	153.094	1045	201.287	**182.078**	432.422	240.000	456
Q_8	68.712	208.156	**51.922**	242.610	349	163.195	**159.765**	401.520	325.000	407
Q_{12}	**19.641**	201.109	25.560	729.656	255	**130.433**	141.343	334.454	862.000	382
Q_{16}	**12.594**	230.672	15.410	943.022	215	**44.190**	68.500	288.500	2145.000	182
Q_{20}	**10.266**	320.297	14.612	2130.435	65	**38.740**	57.977	245.340	3516.170	155
Q_{24}	**10.040**	439.531	12.012	4226.440	54	**20.290**	48.240	210.640	6913.920	139

the number of distinct labels, in a small time compared to other methods. It is close to *Tree+Δ* and it is still far from the execution time of *gIndex* and *FG-Index*. This efficiency of GIRAS is due to the existence of several rare structures in the synthetic data set.

5 Conclusion

In this paper, we have presented GIRAS, an effective index for an efficient query processing on graph databases. A distinguishing feature in GIRAS is the use of a new type substructure called "minimal rare substructures". The latter has a high discriminative power and can be intuitively defined as subgraphs with minimal size and minimal frequency (ideally one) which can distinguish graph. Simulation results on real-world and synthetic large datasets reveal a good performance of GIRAS in terms of recall, precision and query processing time especially for large

queries compared to other well-known proposals. These results should stimulate future research. A challenging direction is the computation of the whole set of subgraphs having a given frequency. To the best of our knowledge; actually, there is no simple way for computation of all subgraphs with a fined frequency. This last point constitutes our immediate focus.

References

1. Cook, S.A.: The complexity of theorem-proving procedures. In: STOC, pp. 151–158 (1971)
2. Giugno, R., Shasha, D.: GraphGrep: a fast and universal method for querying graphs. In: ICPR, pp. 112–115 (2002)
3. Yan, X., Yu, P.S., Han, J.: Graph indexing: a frequent structure-based approach. In: SIGMOD, pp. 335–346 (2004)
4. Cheng, J., Ke, Y., Ng, W., Lu, A.: FG-index: towards verification-free query processing on graph databases. In: SIGMOD, pp. 857–872 (2007)
5. Lee, C.H., Chung, C.W.: Efficient search in graph databases using cross filtering. Inf. Sci. **286**, 1–18 (2014)
6. Zhao, P., Yu, J.X., Yu, P.S.: Graph indexing: tree + delta <= graph. In: The VLDB Journal, pp. 938–949 (2007)
7. Zhang, S., Hu, M., Yang, J.: Treepi: a novel graph indexing method. In: ICDE, pp. 966–975 (2007)
8. Lin, Z., Bei, Y.: Graph indexing for large networks: a neighborhood tree-based approach. Knowl.-Based Syst. **72**, 48–59 (2014)
9. Fan, Z., Peng, Y., Choi, B., Xu, J., Bhowmick, S.S.: Towards efficient authenticated subgraph query service in outsourced graph databases. IEEE Trans. Serv. Comput. **7**(4), 696–713 (2014)
10. He, H., Singh, A.: Closure-tree: an index structure for graph queries. In: ICDE, pp. 38–49 (2006)
11. Williams, D.W., Huan, J., Wang, W.: Graph database indexing using structured graph decomposition. In: ICDE, pp. 976–985 (2007)
12. Yuan, D., Mitra, P.: Lindex: a lattice-based index for graph databases. VLDB J. **22**, 229–252 (2013)
13. Park, C.S., Lim, S.: Efficient processing of keyword queries over graph databases for finding effective answers. Inf. Process. Manage. **51**(1), 42–57 (2015)
14. Katsarou, F., Ntarmos, N., Triantafillou, P.: Performance and scalability of indexed subgraph query processing methods. VLDB Endowment **8**(12), 1566–1577 (2015)
15. Gani, A., Siddiqa, A., Shamshirband, S., Hanum, F.: A survey on indexing techniques for big data: taxonomy and performance evaluation. Knowl. Inf. Syst. **46**(2), 1–44 (2015)
16. Zhao, X., Xiao, C., Lin, X., Wang, W., Ishikawa, Y.: Efficient processing of graph similarity queries with edit distance constraints. VLDB J. **22**(6), 727–752 (2013)
17. Yan, X., Han, J.: gSpan: graph-based substructure pattern mining. In: ICDM, pp. 721–724 (2002)
18. Azaouzi, M., Ben Romdhane, L.: A modified gSpan for computing rare substructures in graph databases. In: ICAI, Sousse, Tunisia, pp. 51–57, June 2013

Multibiometrics Enhancement Using Quality Measurement in Score Level Fusion

Saliha Artabaz[1(✉)], Layth Sliman[2], Hachemi Nabil Dellys[1], Karima Benatchba[1], and Mouloud Koudil[1]

[1] Ecole nationale Supérieure d'Informatique ESI, Oued-Smar, Algiers, Algeria
{s_artabaz,h_dellys,k_benatchba,m_koudil}@esi.dz
[2] Ecole d'ingénieur généraliste en informatique et technologies du numérique Efrei, 30-32 Avenue de la République, Villejuif, 94800 Paris, France
layth.sliman@efrei.fr

Abstract. Multi-biometrics is a solution which is sought to overcome functional and security deficiency in a baseline biometric configuration. In this paper, we propose a multi-biometrics scheme and we apply the cross validation between two databases to study the Equal Error Rate improvement of score level fusion. Our fusion function is constructed using an evolutionary GA on the XM2VTS score database. The best one is tested on a sub-sequence of the BioSecure Score database. As this database offers quality measurement, we transform our function into a weighted function with user-specific approach to study performance enhancement with quality integration. The results are significantly improved with a high confidence and quality measurement becomes inherent to reduce recognition errors.

Keywords: Weighted fusion · User-specific · Tree structure · Cross-validation

1 Introduction

Multi-biometrics is well-considered to deal with traditional biometric systems drawbacks. Major motivations are performance improvement and universality fulfilment. In fact, multiple biometrics allows handling enrolment errors by offering alternative possibilities when a biometric is missing. In addition, experiments show that combining different evidences can enhance performances [1–6]. However, the fused evidences cannot be equivalent and must not be considered as the same in the fusion process [8]. That is why recent researches focus on evaluating quality of each sub-system as entry of the fusion process. As the quality is one of the main factors affecting the overall performance of biometric systems, measuring quality information aims to overcome fusion process errors. Multi-biometrics fusion considers different levels: image, feature, score, decision and rank level. The most used in the literature is the score level [16, 17]. At this level, combined scores are easier to fuse and provide rich information at the same time [4]. Furthermore, the ease of accessing to scores and good performances that outperform other levels make it as the best level [17]. The quality assessment considers

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_26

three points of view depending on the considered information: features, relation between sample and source and impact on the performances. In this paper, we are interested in the score fusion with quality weighting. We conduct a previous work in fusion level presented in [7] by optimizing a fusion function using a tree structure. This work gives sufficient results on the XM2VTS[1] score database. An extended evaluation is needed to prove the effectiveness of the proposed approach. In addition, we aim at optimizing the score distribution by integrating the template-query[2] quality as weights. We choose the Biosecure score database [13] as it is, to the best of our knowledge, the only database that offers quality and cost evaluation. We propose a weighted function based on quality measurement between sample and source. This is simple to apply in real case by including the quality value in the template and integrating the measurement function in the biometric system. The proposed approach outperforms the basic biometric subsystems.

This paper is organized as follows: First, in Sect. 2, we introduce the studied field. Section 3 illustrates the proposed fusion approach and explains database used to conduct experiments. After that, we describe experimental results in Sect. 4. Finally, we conclude and list some perspectives of our work.

2 Multi-biometrics and Quality Fusion

Multi-biometrics is an emerging field that deals with unimodal biometric system problems. Many researchers conduct their experiments on different systems to demonstrate the effectiveness of the fusion in performances enhancement. However, their approaches combine heterogeneous systems and data. The fusion requires normalized data to avoid biased results. Furthermore, weighting systems outputs is required to preserve and improve resulting system performances. This can be done using statistical estimation from system performances or quality measurement [3, 5]. The quality measurement [14] is a promising key that can be used as predictor of the system performances. The authors in [14] define sample quality as "*scalar quantity that is related monotonically to the performance of biometric matchers*". They discuss the effectiveness of a quality evaluation and the different ways to provide this scalar. The sample quality is a result of different points of view [8]:

- Character: refers to the quality (properties) of the physical features of the individual.
- Fidelity: refers to the degree of similarity between a biometric sample and its source (faithfulness).
- Utility: refers to the impact of the individual biometric sample on the overall performance of a biometric system (predicted contribution in performance).

The usefulness of quality measurement and integration is unavoidable in the fusion process since fused systems are heterogeneous. The computed scalars for each sample

[1] XM2VTS is score database built with Lausanne Protocol and based on Face and speech [18].
[2] Template is the data sample used to represent the claimed identity and Query is the sample obtained from the <true ID>.

can be used as weights to fix unbalanced fusion function. This is the subject of many recent researches in order to improve fusion performances [9]. The authors in [9] explain the main raisons that led to use quality measures. They focus on the need to design systems that work in an unsupervised environment, to match with new biometric systems requirements such as portable or low-cost devices, remote access, distant acquisition or forensics. [10] Uses quality information to switch between different system modules depending on the data source. In this work, the authors use linear logistic regression to obtain an efficient combination of matching scores. In [11], authors propose weighting modalities with ancillary information based on relative degradation to enhance performances under noise conditions. [12] processes the data uncertainty concept and fusion confidence with uncertainty function of factors affecting system performance. Their proposition is built upon Dempster-Shafer approach. [15] Uses Score Reliability Based Weighting (SRBW) to estimate matcher weights. The reliability is computed using raw matching scores of each system.

3 Proposed Scheme

Fusion function is the major step to define multi-biometric system. The rule-based function is considered as the best and the most intuitive method to use, exceptionally the sum and weighted sum. In this paper, we aim at testing fusion function based on tree structure with advanced experiments [7]. The function is computed using evolutionary algorithm 'GA'. We use tree structure to generate and apply modifications on fusion function with mutation and crossover operations. Our preliminary tests give high performances with 50 generations. In practice, the fitness function stabilizes within this number. We conduct additional tests with extended number of generations up to 100. We adjust the convergence criteria to 0,001.

Figure 1 illustrates initial population and its evolution with GA simulation according to the crossover operation used to evolve the population, and roulette selection upplied to get the best population list with error under 0,05. A ratio of this population number to the total number of individual is computed. The experimented process progresses to

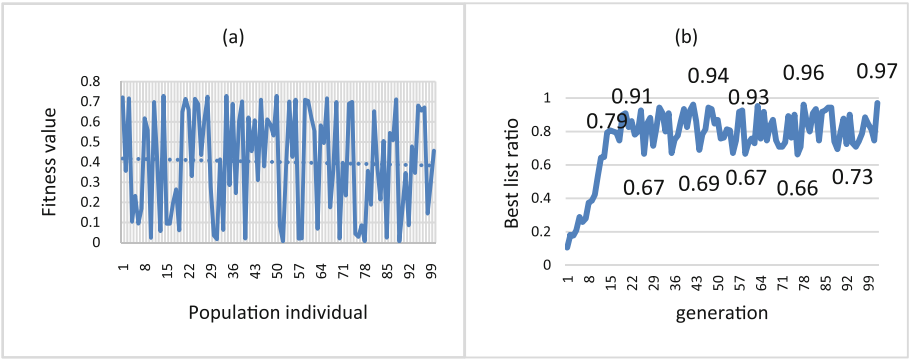


Fig. 1. (a) Fitness of initial population (b) Evolution of best population ratio during generations.

reach a steady range between 60% and 100%. The graph oscillations show the progress of the GA algorithm, which does not converge systematically.

This simulation allows a significant improvement of the fused system. Figure 2 shows the min, max, mean and standard deviations. The max and standard deviation curves illustrate better results and stabilize within 100 generations. We reach an HTER = 0,0018 with 60% improvement (Last HTER = 0,0045).

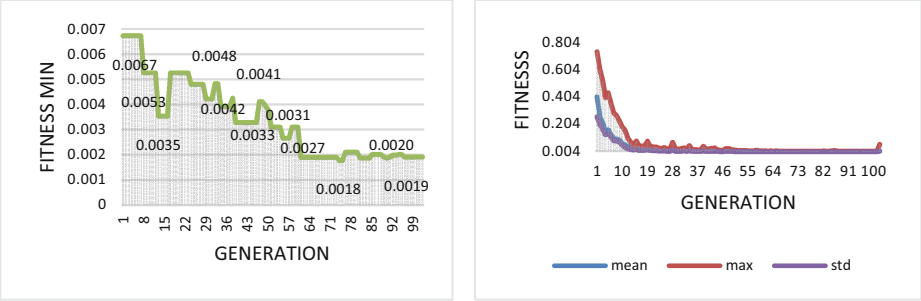


Fig. 2. Statistical values of obtained population during the 100 generations.

We select the best tree obtained from the simulation to establish a cross-validation on another database in order to prove the efficiency of the proposed method. In our work, we aim at studying the impact of quality integration on multi-biometric system. We test the use of quality measurement for weighting our fusion function. We use a user-specific weighting in order to improve baseline systems performances. Which means that weights are adapted to quality measurement of each individual. Each computed weight reflects the sample fidelity to the claimed ID. It represents the gap between sample and its source. We compute the weight as follows:

$$\left(\frac{1}{2} * \left(\left(x_t - \bar{x}\right)^2 + \left(x_q - \bar{x}\right)^2\right)^{1/2}\right)^{-1} \tag{1}$$

where x_t is the template quality (claimed id) and x_q is the query quality (true id).

We use this equation to compute inverse of the standard deviation between the template quality and query quality. A small value means that the two measurements are very different and vice-versa.

To normalize the quality, we use the standard deviation between max and min values computed from the development set. The resulting score fusion is applied on the normalized scores, with normalization to 1 value, using the following function:

$$avg(S_1, avg(\min(\min(S_2, S_3) + S_4 * S_5, S_6), \min(S_7, S_8))) \tag{2}$$

where S_i are the selected scores.

4 Experiments and Discussion

We choose the Biosecure score database [13] as it is the only score database that offers a quality evaluation of its sets. We perform the experiments, using the Biosecure protocol [13] on two sessions of development and evaluation set, according to these steps:

1. Generate 1000 different configurations to select 8 scores from 24 scores of the database;
2. Test the baseline systems of these configurations;
3. Test each configuration with the proposed fusion function on three multi-biometric systems using development set (session 1) to fix experimental parameters:
   (a) fusion of the baseline scores;
   (b) fusion of the normalized scores;
   (c) fusion using the weighted function with quality scalars with normalized and non-normalized scores.
4. Check results on session 2 to avoid biased performances;
5. Repeat tests (a–c) on evaluation set to render system performances (session 1 and 2);
6. Replace dummy values in the normalized baseline system with max and min value according to their sign.

The preliminary results seem interesting since it optimizes scores under the mean of scores 0,54 and 0,58 for session 1 and session 2 respectively. The normalization involves 60% improvement. Results for session 2 are quite better with (AVG = 8%, STD = 22%) vs. (AVG = 16%, STD = 26%) for session 1 (see Fig. 3).

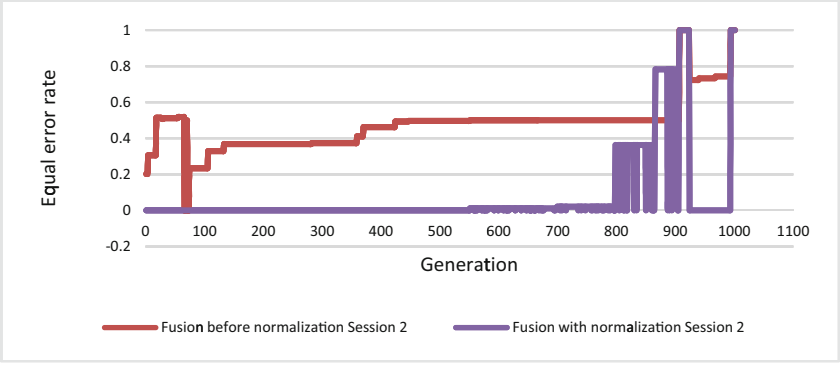


Fig. 3. Normalization impact on EER for 1000 configurations with development set (session 2).

The normalization of baseline systems before fusion reduce fused system errors. However, the curve is not regular. As shown in Fig. 3, the curve for session 2 is chunked into two parts: best EER and worst EER. This is due to the standard deviation that exceeds STD in quality based system. By applying quality measurement, we achieve significant improvement. Despite the worst values above the baseline values, the curve is regular around the mean, which is equal to 9% and 12% for test session 1 and session

2 respectively. This can be seen by the small standard deviation (STD = 11%) comparing to normalization results. These results outperform the fusion of the best-selected systems, which achieve 6% for EER. The quality weighting reaches an EER equal to 0,6%. In addition, our method succeed to enhance worst systems (82% of generations) to achieve an EER between 0 and 20%.

As it is recommended to use session 2 of the evaluation set, we apply the proposed fusion on it to get performances (EER and ROC curve). Figure 4 shows a comparison between development set and evaluation set (session 2). The two curves seem to be similar with a comparable standard deviation (16 for the first and 15 for the last). However, the mean decreases for the evaluation set to half value (12 to 23). We can notice that system performances are preserved even tested on different sets.

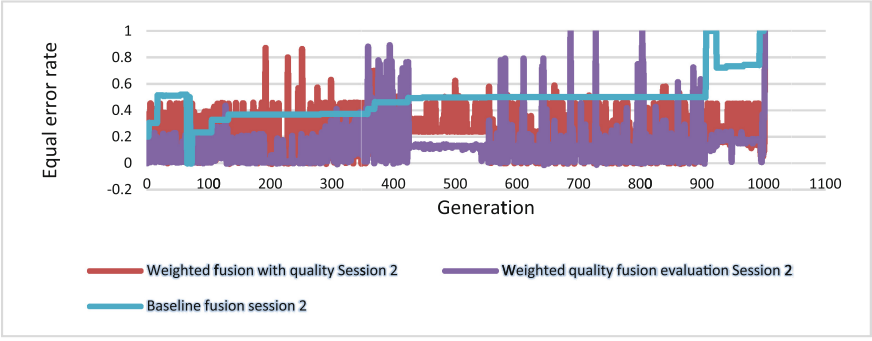


Fig. 4. Weighted fusion function applied on development and test set for session 2.

The ROC curve, of one achieved fusion system under the mean EER = 1,3%, is shown in Fig. 5. It compares between the proposed approach and the eight selected

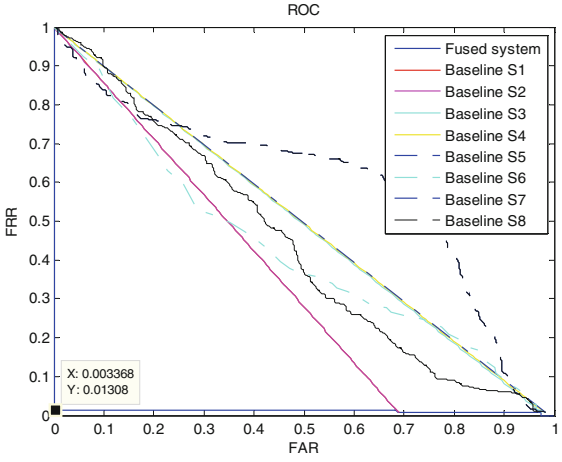


Fig. 5. ROC curves (EER = 0,013) for weighted fusion function and selected baseline systems.

baseline systems. The weighing achieve an important enhancement compared to initial systems. The AUC is much higher for these systems than the fused weighted system. This illustrates the importance of the quality information in the fusion process.

5 Conclusion and Future Work

In this paper, we propose a weighted fusion function based on primitive fusion rule. We use Genetic algorithm to explore search space and find out the best rules combination using the XM2VTS. Next, we apply user-specific weighting on the acquired function. Experiments on the Biosecure score database provide promising improvement. In addition of multi-biometric advantages, these results reach a significant optimization of EER. This work can be extended to compare with other normalization methods and study the impact on density distribution. We can select a sub-list of the obtained trees in order to establish comparative study between them.

References

1. AlMahafzah, H., AlRawashdeh, M.Z.: Performance of multimodal biometric systems at score level fusion. In: Zeng, Q.-A. (ed.) Wireless Communications, Networking and Applications, vol. 348, pp. 903–913. Springer, Heidelberg (2016)
2. Ghulam Mohi-ud-Din, S., Mansoor, A.B., Masood, H.: Personal identification using feature and score level fusion of palm- and fingerprints. Sig. Image Video Process. **5**, 477–483 (2011). Springer
3. Anzar, S.M., Sathidevi, P.S.: Optimization of integration weights for a multibiometric system with score level fusion advances in computing and information technology. In: Meghanathan, N., Nagamalai, D., Chaki, N. (eds.) Advances in Computing and Information Technology. AISC, vol. 177, pp. 833–842. Springer, Heidelberg (2013). doi:10.1007/978-3-642-31552-7_85
4. Conti, V., Militello, C., Sorbello, F., Vitabile, S.: A frequency-based approach for features fusion in fingerprint and iris multimodal biometric identification systems. IEEE Trans. Syst. Man Cyber. Part C App. Rev. **40**, 384–395 (2010)
5. Damer, N., Opel, A.: Multi-biometric score-level fusion and the integration of the neighbors distance ratio. In: Campilho, A., Kamel, M. (eds.) ICIAR 2014. LNCS, vol. 8815, pp. 85–93. Springer, Heidelberg (2014). doi:10.1007/978-3-319-11755-3_10
6. Eskandari, M., Toygar, Ö.: Score level fusion for face-iris multimodal biometric system. In: Gelenbe, E., Lent, R. (eds.) Information Sciences and Systems 2013, pp. 199–208. Springer, Cham (2013)
7. Artabaz, S., Sliman, L., Benatchba, K., Dellys, H.N., Koudil, M.: Score level fusion scheme in hybrid multibiometric system. In: Badioze Zaman, H., Robinson, P., Smeaton, A.F., Shih, T.K., Velastin, S., Jaafar, A., Mohamad Ali, N. (eds.) IVIC 2015. LNCS, vol. 9429, pp. 166–177. Springer, Heidelberg (2015). doi:10.1007/978-3-319-25939-0_15
8. El-Abed, M., Charrier, C., Rosenberger, C.: Quality assessment of image-based biometric information. EURASIP J. Image Video Process. **2015**, 1–15 (2015)
9. Alonso-Fernandez, F., Fierrez, J., Bigun, J.: Quality measures in biometric systems. In: Li, S.Z., Jain, A.K. (eds.) Encyclopedia of Biometrics, pp. 1287–1297. Springer, New York (2015)

10. Alonso-Fernandez, F., Fierrez, J., Ramos, D., Gonzalez-Rodriguez, J.: Quality-based conditional processing in multi-biometrics: application to sensor interoperability. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **40**, 1168–1179 (2010)
11. Mohammed Anzar, S.T., Sathidevi, P.S.: On combining multi-normalization and ancillary measures for the optimal score level fusion of fingerprint and voice biometrics. EURASIP J. Adv. Sig. Process. **2014**, 1–17 (2014)
12. Nguyen, K., Denman, S., Sridharan, S., Fookes, C.: Score-level multibiometric fusion based on dempster shafer theory incorporating uncertainty factors. IEEE Trans. Hum. Mach. Syst. **45**, 132–140 (2015)
13. Poh, N., Bourlai, T., Kittler, J.: A multimodal biometric test bed for quality-dependent, cost-sensitive and client-specific score-level fusion algorithms. Pattern Recogn. J. **43**, 1094–1105 (2009)
14. Grother, P., Tabassi, E.: Performance of biometric quality measures. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 531–543 (2007)
15. Kabir, W., Ahmad, M.O., Swamy, M.N.S.: Score reliability based weighting technique for score-level fusion in multi-biometric systems. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–7 (2016)
16. Lip, C.C., Ramli, D.A.: Comparative study on feature, score and decision level fusion schemes for robust multibiometric systems. In: Sambath, S., Zhu, E. (eds.) Frontiers in Computer Education. AISC, vol. 133, pp. 941–948. Springer, Heidelberg (2012). doi: 10.1007/978-3-642-27552-4_123
17. Jain, A.K., Ross, A.: Introduction to biometrics. In: Jain, A.K., Flynn, P., Ross, A.A. (eds.) Handbook of Biometrics, pp. 1–222. Springer, New York (2008)
18. Poh, N., Bengio, S.: Database protocol and tools for evaluating score-level fusion algorithms in biometric authentication. Pattern Recogn. **39**(2), 223–233 (2006)

Effects of Random Sampling on SVM Hyper-parameter Tuning

Tomáš Horváth[1(✉)], Rafael G. Mantovani[2], and André C.P.L.F. de Carvalho[2]

[1] Faculty of Informatics, Eötvös Loránd University, Budapest, Hungary
tomas.horvath@inf.elte.hu
[2] Institute of Mathematical and Computer Sciences,
University of São Paulo, São Carlos, Brazil
{rgmantov,andre}@icmc.usp.br

Abstract. Hyper-parameter tuning is one of the crucial steps in the successful application of machine learning algorithms to real data. In general, the tuning process is modeled as an optimization problem for which several methods have been proposed. For complex algorithms, the evaluation of a hyper-parameter configuration is expensive and their runtime is speed up through data sampling. In this paper, the effect of sample sizes to the results of hyper-parameter tuning process is investigated. Hyper-parameters of Support Vector Machines are tuned on samples of different sizes generated from a dataset. Hausdorff distance is proposed for computing the differences between the results of hyper-parameter tuning on two samples of different size. 100 real-world datasets and two tuning methods (Random Search and Particle Swarm Optimization) are used in the experiments revealing some interesting relations between sample sizes and results of hyper-parameter tuning which open some promising directions for future investigation in this direction.

Keywords: Sampling · Hyper-parameter tuning · Suport vector machines

1 Introduction

Hyper-parameter (HP) tuning of Machine Learning (ML) algorithms is a challenging task with some practical difficulties. First, good HP settings depend on the dataset used, so the HP tuning should take into account the algorithm-dataset combination. Second, the individual HP values in good HP settings are often related to each other, thus, independent tuning of individual HP values should be avoided. Finally, evaluating the fitness of a specific HP setting can be computationally expensive, an issue on which this paper is focused on.

The choice of the most adequate technique for HP tuning approach depends on the complexity of the ML algorithm used with relation to the given data, i.e.,

Tomáš Horváth is also a member of the Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Slovakia.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_27

if the evaluation of the HP setting for a ML algorithm has a low (many different HP settings can be easily evaluated) or high (HP have to be tuned from only a small number of evaluations) computational cost.

An alternative to speed up the runtime of ML algorithms is sampling [5], when a reduced but representative sample of the data is used to induce a model, instead of using the whole dataset. In addition to random sampling [13], there are more sophisticated (stratified) sampling methods [9,19] which can be used.

The motivation behind the research presented in this paper lies in the following question: to what extent does sampling of a dataset affect the results of HP tuning? A preliminary research, an experiment, was conducted in order to investigate the previous question. 100 real-world datasets and two tuning techniques for HP tuning of Support Vector Machines (SVM) were used in the experiment. For comparing the results of HP tuning on different sample sizes, the use of the Hausdorff distance was proposed. The results of the experiment reveal interesting relations which serve as the basis for the further research in this direction.

2 Hyper-parameter Tuning

In general, a ML algorithm $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of ML algorithms, induces a model $\hat{m}$ with a set of parameter values $\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_l$. However, the algorithm itself has its own parameters $h_1, h_2, \ldots, h_k$, called hyper-parameters (HP), the choice of which affects the resulting parameter values of the induced model. Thus, the HPs of a need to be tuned in order to achieve a good model $\hat{m}$. An example of a HP is the complexity constraint, called cost, C for SVM algorithms.

HP tuning is, in general, treated as an optimization problem whose objective function $f : \mathcal{A} \times \mathcal{D} \times \mathcal{H} \rightarrow \mathbb{R}$ captures the predictive performance of an algorithm $a \in \mathcal{A}$ with the HP setting $\mathbf{h} = (h_1, h_2, \ldots, h_k) \in \mathcal{H}$ on the dataset $\mathbf{D} \in \mathcal{D}$ where $\mathcal{D}$ is the set of datasets and $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2 \times \cdots \times \mathcal{H}_k$ is the space of admissible values of HPs for the algorithm a. The task of HP tuning is, given a, $\mathcal{H}$ and $\mathbf{D}$, to find $\mathbf{h}^\star \in \mathcal{H}$ such that

$$\mathbf{h}^\star = \underset{\mathbf{h} \in \mathcal{H}}{arg\,max}\ f(a, \mathbf{D}, \mathbf{h}) \tag{1}$$

Several techniques for algorithm HP tuning have been proposed. Some of them, although very simple, are widely used, like Grid Search (GS) [3] for low-dimensional $\mathcal{H}$ and Random Search (RS) [10] for high-dimensional $\mathcal{H}$. Local and pattern search techniques [14] extend the palette. However, these techniques tend to get stuck in local minima. Another large family of HP tuning techniques includes Sequential Model-Based Optimization (SMBO) techniques [8]. Nature inspired techniques, like Genetic algorithms (GAs) [6] and Particle Swarm Optimization (PSO) [12], have also been largely used for HP tuning. Recent works have focused on Bayesian techniques [16] and meta-learning [17].

All these techniques have been recently applied to Support Vector Machines (SVM) HP tuning. SVM have been shown to be very efficient when used for

classification tasks. Initialization of optimization methods with promising HP configuration values [12] and recommendation of optimization techniques to be used [11] have also been investigated.

3 Experiment Settings

Classification experiments were carried out using SVM with the Radial Basis Function (RBF) kernel [7]. Two SVM HPs were tuned: the cost (C) and the width of the kernel (γ). Given a dataset (and its sample size) these HPs are tuned using the average per-class accuracy measured over a 10-fold cross-validation resampling strategy.

3.1 Experimental Methodology

The experimental methodology is illustrated in Algorithm 1 and may be described as follows: a random sample of instances is extracted from each dataset. For this sample, the best HP setting is found using a given HP tuning technique. Iteratively, the sample is increased with some of the remaining dataset's instances. The process continues while the extended sample does not contain all of the instances. Since the HP tuning techniques used are stochastic, the whole process is repeated 30 times for each dataset. The resulting set of the best found HP settings is further analyzed.

Algorithm 1. Experiment methodology

Require: $K, \mathcal{D} = \{\mathbf{D}_1, \ldots, \mathbf{D}_N\}, \{p_1, \ldots, p_n\}, f, a, \mathcal{H}, hpt$

for $k = 1 \rightarrow K$ **do**
  for $i = 1 \rightarrow N$ **do**
    $\mathbf{D}_i^k \leftarrow \emptyset$
    $\mathcal{H}_i^{k\star} \leftarrow \emptyset$
    for $j = 1 \rightarrow n$ **do**
      $\mathbf{S} \leftarrow$ sample $p_j\%$ from $\mathbf{D}_i$ $(\mathbf{D}_i^k \cap \mathbf{S} = \emptyset)$
      $\mathbf{D}_i^k \leftarrow \mathbf{D}_i^k \cup \mathbf{S}$
      $\mathbf{h}_{i_j}^{k\star} \leftarrow \underset{\mathbf{h} \in \mathcal{H}}{arg\,max}\, f(a, \mathbf{D}_i^k, \mathbf{h})$ ▷ using hpt
      $\mathcal{H}_i^{k\star} \leftarrow \mathcal{H}_i^{k\star} \cup \{\mathbf{h}_{i_j}^{k\star}\}$
  return $\mathcal{H}_i^{k\star}$

3.2 Parameters Used in the Experiments

The parameters of the experiment methodology, described in Algorithm 1, are the following:

1. *Datasets* ($\mathcal{D}$): 100 (multi-class) classification datasets from the UCI repository[1] were used. Their main characteristics, such as number of instances (Inst.), number of attributes (Attr.) and number of classes (Class.), are summarized in Table 1. Thus, $\mathcal{D} = \{\mathbf{D}_1, \ldots, \mathbf{D}_{100}\}$ and $N = 100$.

[1] http://archive.ics.uci.edu/ml/.

2. *Sample sizes* (p_j): the parameter n of the Algorithm 1 is set to 11, such that $\{p_1, \ldots, p_{11}\} = \{50, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5\}$. The initial sample size is 50% of the dataset's size and at each iteration 5% of the remaining instances are added to the sample. Thus, sample sizes are 50%, 55%, ..., 95%, 100%, respectively, of the original dataset. A simple random sampling mechanism is utilized in experiments.
3. *SVM hyper-space* ($\mathcal{H}$): the chosen algorithm a is SVM with RBF kernel. The fitness function f is a simple classification predictive accuracy defined as the ratio of correctly classified instances to all instances averaged over 10-folds of cross-validation. The SVM hyper-space is set to $\mathcal{H} = \mathcal{H}_C \times \mathcal{H}_\gamma = [2^{-2}, 2^{15}] \times [2^{-15}, 2^3] \subset \mathbb{R}^2$.
4. *HP-Tuning methods* (*hpt*): two HP tuning techniques are used in the experiments:
   - Exhaustive RS, suggested in [2] as a good alternative for HP tuning, with 2500 random trials; and
   - PSO[2] [20]. The number of maximum evaluations was set to 2500. It corresponds to a maximum of 100 iterations with a population composed by 25 particles. The default HP values recommended by LibSVM [4] were added to the initial population. The other PSO parameters are those recommended by the package used for the experiments (see next).

   Since both of techniques are stochastic, each HP tuning process was run 30 times, i.e. the parameter $K = 30$.

All techniques were implemented using the `R` framework[3]. The `e1071` package, which has an interface to the LibSVM[4] library [4], was used for the SVM implementation. For PSO, the `pso` package [1] was used. RS was implemented by the authors.

4 Results and Discussion

Figure 1 shows the results for the *'analcatdata_germangss'* and *'artificial-characters'* datasets, in the first and second rows, respectively. In the columns from left to right, the 30 best[5] HP settings found for samples with 50%, 60%, 70%, 80%, 90% and 100%, respectively, of the original data, are presented.

The first finding is that, not only for the two illustrated dataset, but in general, the results from HP tuning using RS are very similar to those using PSO. This can be due to the relatively large number of evaluations in the experiment, which can allow RS to cover a large portion of the HP space $\mathcal{H}$, enough to find good HP values.

[2] Particle swarm optimization has been successfully used in partially irregular or noisy optimization problems, and, often performs well, finding good solutions because it does not make any assumption about the search landscape.

[3] https://www.r-project.org/.

[4] https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[5] According to the parameter K of the Algorithm 1 set to 30 in the experiment.

Table 1. Classification datasets used in the experiment. Datasets with the maximum (analcatdata_germangss) and minimum (artificial-characters) average distances (d^H_{mean}) are highlighted.

Name	d^H_{RS}	d^H_{PSO}	d^H_{mean}	Name	d^H_{RS}	d^H_{PSO}	d^H_{mean}
abalone-11class	**3.53**	4.09	3.81	ozone-eighthr	6.79	**6.40**	6.60
abalone-28class	**3.30**	3.34	3.32	ozone-onehr	**3.48**	4.84	4.16
abalone-3class	3.20	**2.70**	2.95	page-blocks	**2.40**	3.31	2.86
abalone-7class	**3.43**	4.03	3.73	parkinsons	4.87	**4.31**	4.59
acute-infl.-nephr.	3.90	**2.87**	3.38	pima-ind.-diab.	3.62	**3.09**	3.36
analcat._author.	**3.72**	6.96	5.34	planning-relax	**6.20**	6.86	6.53
analcat._boxing2	4.00	**3.60**	3.80	plant-sp.-leav.-marg.	7.68	**6.95**	7.31
analcat._credit.	**3.38**	3.52	3.45	plant-sp.-leav.-shape	4.40	**2.89**	3.65
analcat._dmft	**4.05**	4.08	4.06	plant-sp.-leav.-text.	**2.44**	3.47	2.96
analcat._germ,	**12.49**	13.24	**12.87***	prnn_crabs	**2.43**	3.36	2.90
analcat._lawsuit	**4.02**	7.11	5.56	qsar-biodegr.	**5.45**	7.38	6.42
appendicitis	**5.55**	5.68	5.62	qualit.-bankruptcy	**3.50**	4.60	4.05
artif.-charact.	**1.17**	1.23	**1.2****	ringnorm	**5.81**	8.15	6.98
autoU.-au1-1000	9.78	**5.49**	7.63	robot-failure-lp4	**3.84**	6.80	5.32
autoU.-au4-2500	3.25	**2.93**	3.09	robot-failure-lp5	**2.67**	2.80	2.73
autoU.-au6-1000	5.64	**5.56**	5.60	saheart	2.93	**2.57**	2.75
autoU.-au6-250-dr.	**3.57**	5.92	4.75	seeds	6.52	**5.92**	6.22
autoU.-au6-cd1-400	**1.95**	3.08	2.52	seismic-bumps	10.34	**10.24**	10.29
banknote-auth.	**2.84**	3.80	3.32	spambase	**2.08**	2.31	2.20
breast-canc.-wisc.	**4.13**	4.22	4.17	spectf-heart	**4.90**	4.99	4.95
breast-tiss.-4class	**4.22**	4.99	4.61	spect-heart	4.53	**4.00**	4.27
breast-tiss.-6class	7.24	**6.31**	6.78	statlog-austr.-cr.	**5.06**	5.87	5.46
bupa	2.73	**2.69**	2.71	statlog-ger.-cr.	3.72	**2.74**	3.23
car-evaluation	**1.79**	2.02	1.9	statlog-ger.-cr.-num.	2.53	**2.33**	2.43
climate-sim.-crach.	6.84	**4.13**	5.48	statlog-heart	3.85	**3.58**	3.71
cloud	6.25	**5.86**	6.05	statlog-im.-segm.	**2.10**	2.34	2.22
cmc	2.61	**2.33**	2.47	statlog-land.-sat.	1.60	**1.23**	1.42
conn.-mines-vs-rocks	**2.87**	3.50	3.19	statlog-veh.-silh.	**1.94**	2.00	1.97
conn.-vowel	**2.17**	2.20	2.19	teaching-assist.-eval.	**3.05**	3.09	3.07
conn.-vowel-reduced	**2.28**	2.39	2.34	thoracic-surgery	**6.11**	12.21	9.16
contrac.-meth.-choice	2.17	**2.00**	2.08	thyroid-allhyper	4.62	**2.88**	3.75
dermatology	4.51	**4.13**	4.32	thyroid-allrep	**2.24**	2.34	2.29
ecoli	3.52	**3.27**	3.4	thyroid-ann	**1.79**	1.87	1.83
fertility-diagnosis	**5.74**	6.73	6.24	thyroid-dis	**2.55**	4.01	3.28
glass	**2.69**	3.32	3.00	thyroid-hypothyroid	**2.76**	3.71	3.23
habermans-survival	**5.14**	8.02	6.58	thyroid-newthyroid	**4.64**	5.18	4.91
hayes-roth	2.79	**2.27**	2.53	thyroid-sick	3.02	**2.70**	2.86
hepatitis	**4.15**	5.05	4.6	thyroid-sick-euthyr.	**6.83**	10.45	8.64
horse-colic-surgical	10.57	**9.31**	9.94	tic-tac-toe	6.93	**5.3**	6.12
indian-liver-patient	**5.85**	7.84	6.84	user-knowledge	2.47	**2.43**	2.45
ionosphere	8.49	**6.78**	7.63	voting	**3.74**	4.38	4.06
iris	**3.30**	4.08	3.69	wdbc	**3.08**	3.62	3.35
kr-vs-kp	**2.11**	4.69	3.4	wholesale-channel	**3.44**	3.78	3.61
leaf	3.22	**2.66**	2.94	wholesale-region	**3.55**	4.72	4.13
led7digit	4.60	**4.00**	4.30	wilt	**2.74**	2.90	2.82
leukemia-haslinger	4.40	**3.62**	4.01	wine	**5.58**	7.72	6.65
mammographic-mass	**4.67**	5.44	5.05	wine-quality-red	**11.49**	12.74	12.12
mfeat-fourier	**4.30**	4.63	4.46	wpbc	**4.34**	7.61	5.97
movement-libras	**2.96**	3.86	3.41	yeast	**2.79**	4.14	3.46
optdigits	**4.54**	6.04	5.29	yeast-4class	3.12	**2.78**	2.95

However, Fig. 1 shows two different situations: for the *artificial-characters* dataset, the HP tuning results for different sample sizes are very similar, regardless whether RS or PSO was used (the six plots in the second row of charts). On the other hand, for the *analcatdata_germangss* dataset, the best HP configurations found differ more for distinct sample sizes. Besides, the 30 HP settings

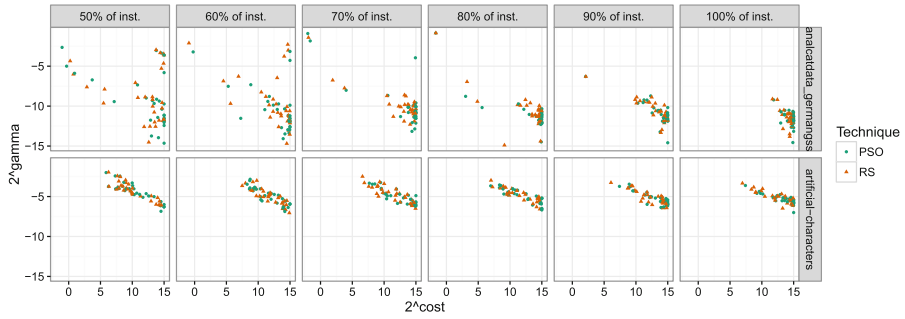


Fig. 1. The 30 best HP configurations found by RS (orange triangles) and PSO (green squares) for samples with 50%, 60%, 70%, 80%, 90% and 100% (1st, 2nd, 3rd, 4th, 5th and 6th columns, respectively) of the analcatdata_germangss (first row) and artificial-characters (bottom row) datasets.

found are less dispersed across $\mathcal{H}$ for larger sample sizes. It means that tuning the C and γ HPs is less sensitive to sampling in the *artificial-characters* dataset than in the *analcatdata_germangss* dataset.

4.1 Hausdorff Distance

One way to measure the difference in the HP tuning results for various sample sizes is to use the Hausdorff distance [18], a commonly used dissimilarity measure between two sets $\mathcal{A}$ and $\mathcal{B}$ of points, defined as

$$d^H(\mathcal{A}, \mathcal{B}) = max\{d(\mathcal{A}, \mathcal{B}), d(\mathcal{B}, \mathcal{A})\} \tag{2}$$

where

$$d(\mathcal{A}, \mathcal{B}) = \underset{\mathbf{a}\in\mathcal{A}}{max}\{\underset{\mathbf{b}\in\mathcal{B}}{min}\{||\mathbf{a}, \mathbf{b}||\}\} \tag{3}$$

and $||.,.||$ is any norm, e.g. the Euclidean distance. Two sets are close according to the Eq. 2 if every point of either set is close to some point of the other set.

According to the Algorithm 1, for a $\mathbf{D}_i$ dataset and a fixed sample of size $p_1 + \cdots + p_j$, such that $1 \leq j \leq n$, there are K HP settings $\mathbf{h}_{i_j}^{1\star}, \mathbf{h}_{i_j}^{2\star}, \ldots, \mathbf{h}_{i_j}^{K\star}$. Let $\mathcal{A} = \{\mathbf{h}_{i_j}^{1\star}, \mathbf{h}_{i_j}^{2\star}, \ldots, \mathbf{h}_{i_j}^{K\star}\}$ and $\mathcal{B} = \{\mathbf{h}_{i_l}^{1\star}, \mathbf{h}_{i_l}^{2\star}, \ldots, \mathbf{h}_{i_l}^{K\star}\}$, where $1 \leq j \neq l \leq n$, be two sets of HP settings computed by a HP tuning technique for two samples of different sizes sampled from the same dataset. Pairwise Hausdorff distances between various sets $\mathcal{A}$ and $\mathcal{B}$ can be calculated.

4.2 Pairwise Differences

The pairwise distances, averaged over all the datasets used in the experiment, are shown in Table 2. Although the results are very similar for the two HP tuning techniques evaluated, the RS results are less "distant" with relation to different sample sizes than the PSO results. One reason is that PSO is a more

Table 2. Pairwise Hausdorff distances between HP tuning results for samples of different sizes averaged over all the datasets. Results for the RS and PSO HP tuning techniques are shown above and below the diagonal, respectively.

Sample size	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
50%	–	3.53	3.58	3.72	3.78	3.84	4.11	4.02	4.32	4.47	**5.03**
55%	4.01	–	3.68	3.64	3.62	3.83	4.01	4.01	4.27	4.58	**4.95**
60%	4.01	3.78	–	3.57	3.57	3.59	3.83	3.72	4.10	4.25	**4.58**
65%	4.24	3.82	3.78	–	3.52	3.57	3.88	3.75	3.75	4.27	**4.46**
70%	4.53	3.92	3.93	4.18	–	3.46	3.53	3.63	3.82	4.09	**4.53**
75%	4.24	4.14	3.88	4.01	4.00	–	3.66	3.57	3.62	4.01	**4.07**
80%	4.55	4.16	4.03	3.95	4.05	4.00	–	3.48	3.46	3.72	**4.16**
85%	4.54	4.36	4.00	4.01	3.88	3.86	3.72	–	3.49	3.69	**3.80**
90%	4.78	4.53	4.20	4.42	4.27	4.02	3.68	3.76	–	3.31	**3.57**
95%	5.27	4.94	4.60	4.72	4.52	4.24	3.94	3.84	3.58	–	**3.27**
100%	**5.61**	**5.30**	**4.85**	**4.85**	**4.76**	**4.73**	**4.22**	**4.10**	**4.14**	**3.78**	–

sophisticated optimization algorithm than RS and tends to narrow down the search space of candidate solutions. Thus, if $\mathcal{A}$ and $\mathcal{B}$ are less dispersed, then there is a higher chance that the overlap of the areas covered by these two sets is smaller.

4.3 Tuning Based on Samples vs. the Whole Dataset

The distances between HP tuning results for samples of different size and the whole dataset, averaged over the 30 runs and all the datasets, are illustrated by bold in the last row and last column of Table 2. For both RS and PSO, the differences between the results (HP settings) found for the whole dataset and the results found for samples (abbreviated as "average differences" below) of size 60%, 65% and 70% are very similar. For RS in particular, average differences in results are similar for the samples of size of 50% and 55% and for samples of size 75% and 80%. Regarding PSO, similar average differences were recorded for sample of sizes 80%, 85% and 90%. Besides that, average differences for the samples of size 75% are similar to those of sizes 60%, 65% and 70%.

These results can support decisions about the size of samples to be used for HP tuning. For example, using a sample size of 60% would probably lead to similar results but in a shorter time as if the a sample size of 75% would be used for tuning the HP with PSO.

Table 1 contains, for each dataset, the averaged Hausdorff distances between the HP tuning results (averaged over the 30 runs) using the whole data and samples of different size, for RS (denoted as d^H_{RS}) and PSO (denoted as d^H_{PSO}) as well as their mean (denoted as d^H_{mean}). Datasets with the minimum and the maximum values for d^H_{mean} are highlighted.

4.4 Significant Differences Between Sample Sizes

Table 3 contains significant differences between HP tuning results for samples which size differ in 5%, 10%, ..., 45% and 50% of instances. The results were computed as follows: For each dataset, Hausdorff distances between its samples differing in 5%, 10%, ..., 50%, respectively, were averaged (e.g. in case of 10% difference, these are the distances between sample sizes of 50–60%, 60–70%, ..., 90–100%) resulting in ten 100-dimensional vectors. Wilcoxon test was applied to each pair of these vectors.

The results indicate that there are no significant differences when sample sizes differ in less than 25% or more than 35% of the size of the whole dataset.

Table 3. Statistically significant differences according to Wilcoxon test between the results of HP tuning for samples differing in 5%, ..., 50% of instances with relation to the size of the whole dataset. ∘ denote p-value ≤ 0.05 and • denote p-value ≤ 0.01. Results for RS and PSO are above and below the diagonal, respectively.

Difference in sample sizes	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
5%	–					•	•	•	•	•
10%		–				∘	∘	•	•	•
15%			–			∘		∘	•	•
20%				–					∘	•
25%					–				∘	•
30%	•	∘				–				
35%	∘	∘					–			
40%	•	•	•	∘				–		
45%	•	•	•	•	•	∘			–	
50%	•	•	•	•	•	∘				–

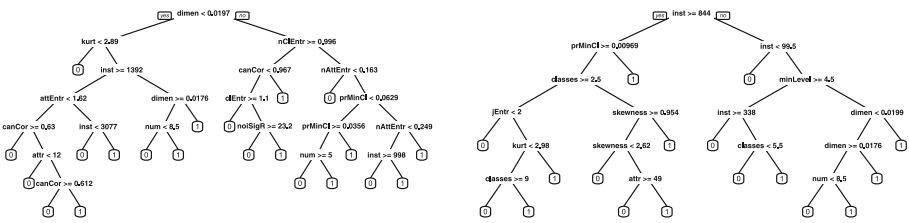


Fig. 2. Decision trees induced from dataset characteristics for a target variable d^H_{RS} (left) and d^H_{PSO} (right) from the Table 1. Leaves labeled with 1 mean that $d^H_{RS} > 3$ or $d^H_{PSO} > 3$, respectively, otherwise the leaves are labeled with 0.

4.5 Data Characteristics vs. Sensitivity of HP-Tuning on Sampling

In order to reveal relations between the sensitivity of HP tuning on sampling from a dataset and some characteristics of the dataset, a decision tree was induced using the `rpart` package of `R`. As explanatory measures, some simple [12], statistical [17] and information-theoretic [15] measures were extracted from each dataset. These measures, known as meta-features, are well explored in meta-learning research [11]. A dataset was created where each dataset is represented by predictive attributes, a.k.a. the meta-features, and a target attribute. The target attribute can assume the values from d^H_{RS} or d^H_{PSO}, from Table 1, depending if the HP tuning for the dataset used RS or PSO, respectively. The continuous target values (d^H_{RS} or d^H_{PSO}) were transformed to binary values, as follows: values larger than 3 were encoded as 1 (True), otherwise as 0 (False), indicating that HP tuning on the given dataset is sensitive or not, respectively, to sampling. The resulting trees, illustrated in Fig. 2 were not pruned, in order to reveal all meta-features considered relevant to the classification.

For RS, the tree contains the following characteristics: dimensionality of the dataset (dimen), average kurtosis of continuous attributes (kurt), normalized class entropy (nClEnt), number of instances (inst), canonical correlation between attributes and labels (canCor), normalized attribute entropy (nAttEntr), attribute entropy (attEntr), class entropy (clEntr), probability of the minority class (prMinCl), noise-to-signal ratio computed from the average attribute entropy and the mutual information in the data (noiSigR), number of attributes (attr) and number of numeric attributes (num). Regrading PSO, the tree presents the following measures: number of classes (classes), minimum number of levels of nominal attributes (minLevel), joint entropy (jEntr), skewness and the inst, prMinCl, dimen, kurt, attr and num measures introduced above.

In both cases, the number of attributes and the number of instances in the dataset were among the most important measures determining the sensitivity of HP tuning regarding the sampling.

5 Conclusions and Future Work

This paper investigated the effect of random sampling on SVM HP tuning. For such, several experiments were performed, where various samples were generated for 100 datasets using PSO and RS HP tuning techniques. The main goal of this study was to investigate how the size of the sample affects the HP tuning results, when compared with the results obtained for the whole dataset.

The presented research[6] is in its early stages and need to be further pursued but the experimental results obtained indicate some future research directions, which include: to investigate the use of the Hausdorff distance for HP tuning in data stream mining, where new instances arrive continuously, investigate how

[6] Supported by the Brazilian Funding Agencies CAPES, CNPq and São Paulo Research Foundation FAPESP (CeMEAI-FAPESP process 13/07375-0 and grant #2012/23114-9), and, the Slovakian project VEGA 1/0475/14.

the results would be affected by the use of other HP tuning techniques [14,16] or some stratified sampling techniques [9]. To the best of the authors' knowledge, this research is the first to investigate the effects of sampling on the HP tuning process, is relevant for ML and worth further research.

References

1. Bendtsen, C.: pso: Particle Swarm Optimization, R package version 1.0.3 (2012)
2. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**, 281–305 (2012)
3. Braga, I., Carmo, L.P., Benatti, C.C., Monard, M.C.: A note on parameter selection for support vector machines. In: Castro, F., Gelbukh, A., González, M. (eds.) MICAI 2013. LNCS (LNAI), vol. 8266, pp. 233–244. Springer, Heidelberg (2013). doi:10.1007/978-3-642-45111-9_21
4. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27:1–27:27 (2011)
5. Cochran, W.G.: Sampling Techniques, 3rd edn. Wiley, New York (1977)
6. Friedrichs, F., Igel, C.: Evolutionary tuning of multiple SVM parameters. Neurocomputing **64**, 107–117 (2005)
7. Hofmann, T., Schölkopf, B., Smola, A.J.: Kernel methods in machine learning. Ann. Stat. **36**(3), 1171–1220 (2008)
8. Hutter, F., Hoos, H.H., Leyton-Brown, K.: Sequential model-based optimization for general algorithm configuration. In: Coello, C.A.C. (ed.) LION 2011. LNCS, vol. 6683, pp. 507–523. Springer, Heidelberg (2011). doi:10.1007/978-3-642-25566-3_40
9. Lang, K., Liberty, E., Shmakov, K.: Stratified Sampling Meets Machine Learning (2015)
10. Mantovani, R.G., Rossi, A.L.D., Vanschoren, J., Bischl, B., de Carvalho, A.: Effectiveness of random search in SVM hyper-parameter tuning. In: International Joint Conference on Neural Networks, pp. 1–8 (2015)
11. Mantovani, R.G., Rossi, A.L.D., Vanschoren, J., Bischl, B., de Carvalho, A.: To tune or not to tune: recommending when to adjust SVM hyper-parameters via meta-learning. In: 2015 International Joint Conference on Neural Networks, pp. 1–8 (2015)
12. Mantovani, R.G., Rossi, A.L.D., Vanschoren, J., Carvalho, A.C.P.D.L.: Meta-learning recommendation of default hyper-parameter values for SVMs in classification tasks. In: 2015 International Workshop on Meta-Learning and Algorithm Selection at ECML/PKDD, pp. 80–92 (2015)
13. Meng, X.: Scalable simple random sampling and stratified sampling. In: JMLR Workshop and Conference Proceedings, vol. 28, pp. 531–539 (2013)
14. Momma, M., Bennett, K.P.: A pattern search method for model selection of support vector regression. In: SIAM International Conference on Data Mining. SIAM (2002)
15. Reif, M., Shafait, F., Goldstein, M., Breuel, T., Dengel, A.: Automatic classifier selection for non-experts. Pattern Anal. Appl. **17**(1), 83–96 (2014)
16. Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. In: NIPS, pp. 2960–2968 (2012)
17. Soares, C., Brazdil, P.B.: Selecting parameters of SVM using meta-learning and kernel matrix-based meta-features. In: ACM Symposium on Applied computing, pp. 564–568. ACM (2006)

18. Taha, A.A., Hanbury, A.: An efficient algorithm for calculating the exact hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. **37**(11), 2153–2163 (2015)
19. Tillé, Y.: Sampling Algorithms. Springer, Heidelberg (2006)
20. Yang, X.-S., Cui, Z., Xiao, R., Gandomi, A.H., Karamanoglu, M.: Swarm Intelligence and Bio-Inspired Computation: Theory and Applications. Elsevier, Amsterdam (2013)

Bayesian Networks for Identifying Semantic Relations in a Never-Ending Learning System

Edimilson Batista dos Santos[1], Massilon Lourenço Fernandes[1], Estevam R. Hruschka Jr.[2], and Maísa Cristina Duarte[3(✉)]

[1] DComp/UFSJ – Department of Computer Science, Federal University of São João del-Rei, São João del-Rei, MG, Brazil
edimilson.santos@edu.ufsj.br, massilon@live.com

[2] DC/UFSCar – Computer Department, Federal University of São Carlos, São Carlos, SP, Brazil
estevam@dc.ufscar.br

[3] Univ. Lyon, UJM-Saint-Etienne, CNRS, Laboratoire Hubert Curien UMR 5516, 42023 Saint Etienne, France
maisa.cristina.duarte@univ-st-etienne.fr

Abstract. A new paradigm of Machine Learning named Never-Ending Learning has been proposed through a system known as NELL (Never-Ending Language Learning). The major idea of this system is to learn to read the web better each day and to store the gathered knowledge in a knowledge base (KB), continually and incrementally. This paper proposes a new method that can help NELL populating its own KB using Bayesian Networks (BN). More specifically, we use facts (knowledge) already stored in NELL's KB as input for a BN learning algorithm named VOMOS (Variable Ordering Multiple Offspring Sampling) by aiming at representing the acquired knowledge by NELL system. In addition, we propose to use the BN induced by VOMOS for identifying new semantic relations to be added to NELL's KB, expanding thus its initial ontology.

Keywords: Bayesian networks · Never-Ending Learning · Machine Learning

1 Introduction

Machine Learning [1] is a research area that has received a lot of attention within the Artificial Intelligence and Computer Science in general. This fact has contributed for a great advance and progress in results obtained by methods and algorithms from this research area in recent decades. Currently, however, there are not yet many computer systems able to learn cumulatively, forever, in a never-ending learning fashion. More important, there are not many systems that use the knowledge acquired yesterday to improve their ability of learning today, in a continuous and never-ending process.

The first never-ending learning system reported in literature is named NELL (Never-Ending Language Learning) [2, 3]. NELL is a computer system that runs 24 h per day, 7 days per week, extracting information from web text to populate and extend its own knowledge base. The main goal of the system is to learn to read the web better each day, and to store the gathered knowledge in a never-ending growing knowledge

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_28

base (KB). The system takes advantage of many different components like CPL [4], CSEAL [5], Prophet [6], OntExt [7], and Conversing Learning [8], in order to be self-supervised and avoid semantic drifting [9, 10]. Semantic drifting is expected to happen in semi-supervised systems after too many iterations without supervision. NELL's key principle to minimize semantic drifting is to couple many different views and many different tasks, in a multi-view [11] and multi-task [12] learning approach. The idea is combining different components with different approaches aiming to obtain a confident learning and minimize the semantic drifting. NELL's KB is represented by an ontology-based structure characterized by categories, relations and their instances. In [3], an extended version of the prototype reported in [2] is presented with the currently subsystems (modules or components).

Since NELL's KB continuously grows each day, it does not contain all instances of every category, neither all instances of every relation described in the ontology [13]. The main idea behind the work described in this paper is to show how we can help NELL populating its own knowledge base using Bayesian Networks (BN) [15]. In this paper, we use facts (knowledge) already stored in NELL's KB as input for a BN learning algorithm named VOMOS (Variable Ordering Multiple Offspring Sampling) [16] aiming to discover new relations (Relation Discovery - RD), as well as to extract new instances for relations (Relation Extraction – RE) that NELL was not able to obtain from web text before.

Related works focus on the same task of RD and RE using NELL's KB, as well as other KB's. In [17], the authors propose the use of association rules in order to populate NELL's KB with instances and generalized association rule to investigate how useful they can be to extend the relations between the KB categories. In [14] a very similar idea is proposed, but based on Bayesian Sets (BS). However, developing methodologies to help both, extending and populating such KB and improving their coverage is still a challenge. In this paper we are interested in starting to explore Bayesian Networks (BN) to help in such tasks.

BN are probabilistic models that allow inference and may be used to discover relationships between variables, thus the methods and algorithms applied to induce the BN structure may also be used to guide NELL's ontology automatic extension and population. To do so, we propose to use VOMOS, a hybrid adaptive algorithm based on MOS approach [18] and designed for improving the process of inducing a BN structure from data. The main idea behind VOMOS is to explore the power of different evolutionary operators to search for suitable variable orderings.

Therefore, this paper presents, as specific goals, (i) the application of VOMOS to induce a BN capable of representing the acquired knowledge by NELL and allow inference (based on such representation); (ii) explore VOMOS induced BN to identify new semantic relations to be added to NELL's KB (expanding the initial ontology).

The remainder of this paper is organized as follows. Section 2 provides a brief overview of BN and it also presents VOMOS algorithm. Section 3 shows how VOMOS has been applied to NELL's knowledge base to induce a BN and to identify new semantic relations. Section 3.2 also presents some experiments with the proposed approach. Finally, Sect. 4 brings the concluding remarks and points out some future work.

2 Bayesian Networks

Bayesian Networks (BN) [15] are graphical representations of multivariate joint probability distributions. They are described by directed acyclic graphs in which the nodes represent the variables and the arcs represent probabilistic dependencies between connected nodes (variables). The strength of each dependence is given by the conditional probability $P(x_i \mid \pi_{x_i})$, where x_i and π_{x_i} are the i-th variable and the set of parents of x_i in the graph, respectively. The use of conditional independence is the key to the ability of BN to provide a general-purpose compact representation for complex probability distributions.

Computational methods for learning BN may be seen as a manner to identify a probabilistic model which describes the dependence (and independence) among variables from a given domain [19]. Thus, BN learning algorithms may be used as a tool for the discovery of relationships among variables and therefore they are suitable for the purposes of this work.

Among several BN learning algorithms available in the literature, this work applies a specific method named VOMOS [16]. This method, presented in Subsect. 2.1, has important characteristics for the solution of the problem of identifying semantic relations in NELL's knowledge base.

2.1 VOMOS (Variable Ordering Multiple Offspring Sampling)

VOMOS [16] is a Multi-Off-Spring (MOS) algorithm [18] aiming at optimizing the learning of Bayesian network structure by searching for a suitable Variable Ordering (VO). The basic idea of VOMOS is to use a hybrid adaptive algorithm (MOS) that simultaneously handles several evolutionary approaches and dynamically adjusts the participation of each one of them in the overall search process, as described in [18].

The main characteristics of the general evolutionary strategy employed by VOMOS are the following. Each individual of the population represents a possible VO. The variable identification (ID) is coded as an integer number. Therefore, for a problem described by M variables, V_1, V_2, …, V_M having as ID the integers 1, 2, …, M respectively, there are M! possible VO and therefore M! possible individuals.

Figure 1 presents VOMOS algorithm pseudo-code; the input for the algorithm is the training dataset. MOS starts the search process by generating the initial population randomly (P_0) and evaluating each individual. The individuals are evaluated by the fitness function and the best ones are then selected to generate the next generation. The new individuals will be created using a set of recombination operators (crossover and/or mutation operators). Each set of these operators create their own individuals $O_i^{(j)}$ (i is the generation and j is the set of recombination operators).

The idea of using this set of operators embedded in MOS is to explore the potential of each one collectively. In this sense, we are interested in approaching the variable ordering problem integrating previously defined evolutionary operators. The idea of integrating different methods and algorithms in order to enhance the results in a specific problem has shown good results in works related to Ensembles of Classifiers [20], as well

```
Algorithm: VOMOS Algorithm
{Input: training dataset.}
{Output: Best_VO, Best_BN.}
1  begin
2      Create initial global population of candidate solutions P0.
3      Evaluate initial population P0.
4      while termination criterion not reached do
5        for every crossover operator do
6                for every mutation operator to (if there are any)
7                    Create new individuals from current population
                 Pi.
8                    Evaluate new individuals.
9                   Add new individuals to an auxiliary population
                 Oi(j).
10               end
11               end
12       Combine populations O i (j) and Pi according to a pre-
         established criterion to generate Pi+1.
13     end
14   end
```

Fig. 1. VOMOS algorithm pseudo-code (extracted from [16]).

as in the "Never-Ending Learning" paradigm [2]. Therefore it motivated us to explore these integration principles by using the evolutionary approach given in MOS [18].

According to empiric results presented in literature [16, 21], VOMOS is able to induce good BNs which can be used for identifying relationships among variables from given domain. Thus, VOMOS seems suitable for the search of semantic relations between variables from NELL's knowledge base.

3 VOMOS Algorithm Applied to NELL's Knowledge Base

This paper aims at to apply VOMOS algorithm to NELL's KB. The BN structure induced by VOMOS allows identifying the semantic relations existing in the KB. Thus, it will be possible to discover new relations that may be identified through BN structure and inserted in the knowledge base in the future, expanding the NELL's initial ontology.

Since VOMOS seeks a suitable variable ordering for the BN structure learning, when it is applied to NELL's KB, it may identify the correct relationships (of dependence and independence) among the variables from the specific domain. Thus the BN induced by VOMOS allows, as direct consequence, a representation of NELL's functional relations in a representation structure that enables inference.

3.1 Preparation of the Datasets

NELL's KB was initially pre-processed. This data pre-processing is very important because the data format stored in the KB is not suitable for BN learning algorithms.

The pre-processing procedure allowed building a dataset containing only the NELL's functional relations, named DataSet1. Thus, VOMOS was ran using DataSet1 as input to induce a BN able to allow inference in NELL's KB and identify new semantic relations.

NELL's complete KB is available at http://rtw.ml.cmu.edu/rtw/resources. To apply the pre-processing step, we have used the tab-separated-value file with every belief in the KB, one per line. The accessed file was a subset built from iteration 945. The tab-separated-value file has thirteen columns (or variables), of which only the first three were used: Entity, Relation and Value (forming a kind of SPO, Subject-Predicate-Object, triple). The variable Relation presents relations existing between two categories stored in Entity and Value. Table 1 shows two examples of instances for these three variables. The first instance brings the relation *teamplaysinleague* that allowed learning two new values for the categories *sporteam* (in Entity) and *sportsleague* (in Value). This line in Table 1 means that *Montana_tech_orediggers* is a sport team which plays in league *ncaa*. Similarly, the second instance stores the value *hunterdon*, learnt for the category *county* (in Entity) and the value *raritan*, learnt for the category *river* (in Value). Both values were learnt from relation *cityliesonriver*.

Table 1. Examples of instances for the variables *Entity, Relation, Value*, provided in NELL's KB.

Entity	Relation	Value
concept: sportsteam:montana_tech_orediggers	concept: teamplaysinleague	concept:sportsleague: ncaa
concept:county:hunterdon	concept:cityliesonriver	concept:river:raritan

The dataset for VOMOS was built from values of those three variables provided in NELL's KB. It can also be noticed, in examples shown in Table 1, that data still need to be cleaned before being used (e.g., in *concept:river:raritan*, it might be that we are only interested in *raritan*). Thus, the dataset for VOMOS is composed of variables (that represent categories) and their values which form semantic relations in NELL's KB. For example, the relation *teamplaysinleague* (see Table 1) provides both variables *team* (*sportsteam*) and *league* (*sportsleague*) to compose the dataset for VOMOS; their values (*montana_tech_orediggers* and *ncaa*) are also stored. From this dataset, VOMOS may induce a BN able to represent the relation *teamplaysinleague* through a directed arc between variables *team* and *league*, existing in induced BN structure, reinforcing NELL's belief that this relation is correct.

3.2 Experiments and Analysis of Results

This section describes the experiments performed using VOMOS algorithm having as input datasets obtained from a relation set and values which were stored in NELL's knowledge base (KB). The main goal described here is to induce Bayesian networks (BN) able to represent the semantic relations existing in NELL's KB and identify new relations not present yet. VOMOS algorithm was chosen to carry out these experiments

mainly because of its properties of being able to find a suitable variable ordering (VO) to optimize BN's learning. Thus, the induced BN structure may be considered the most probable to represent the semantic relations existing in NELL's KB, given a set of explored VO.

Two datasets were built based on DataSet1, described in Subsect. 3.1. These two new datasets have their characteristics summarized in Table 2. The first dataset called *DataSet_NELLValues* has 6 variables and 31000 instances which were obtained from relations representing knowledge about sports. In each line of this dataset, the conjunction of the values of all variables represents a true knowledge, as shown in Table 3. Notice that the first line in Table 3 brings the following true information: *ahman_green* (athlete) is a *football* (sport) player that plays for *texans* (team), as a *running_back* (sportposition), having fans in *germany* (country of fans) and uses *cleats* (equipment). In this case, this dataset tends to present relationships among all variables aiming to check if VOMOS can represent these relationships in a BN.

Table 2. Datasets used in experiments.

Dataset names	Number of attributes	Number of instances
DataSet_NELLValues	6	31000
DataSet_Relations	10	100000

The second dataset called *DataSet_Relations* has 10 variables and 100000 instances and, like the first dataset, it has data regarding sports. However, unlikely the first dataset, the instances do not clearly suggest the relationships among every variable. In this scenario, we are interested in assessing the ability of the VOMOS algorithm to find new relationships among distinct variables from NELL's data.

As previously mentioned in Subsect. 2.1, VOMOS is an evolutionary algorithm and it was applied here following the same setting and methodology described in [16]. Each individual (representing a possible variable ordering) is evaluated by an objective function. The objective function, which expresses the quality of the BN structures, used here is the natural logarithm of a posteriori probability of the database of cases, given the structure to be evaluated, following the definition given by Cooper and Herskovits (g function) [19]. The initial population has 100 individuals. The new individuals are created using different combinations of crossover and mutation operators. Each combination is applied to same population. The crossover rate was defined as 0.9 and the mutation rate as 0.1. For the stopping criterion, it was defined a maximal number of 5 generations without improvements. These parameter values were empirically defined.

Next, the VOMOS algorithm was run for the two datasets. Since the algorithm has stochastic nature, more than one run is necessary to verify the final solution. For this reason, VOMOS was run 35 times for each dataset. Table 4 presents the average Bayesian score (g function) obtained by VOMOS. Since the two datasets have few variables, the VOMOS algorithm converged fast. As soon in first generation, in all runs, for two datasets, VOMOS has already obtained the best result. This result suggests that VOMOS is more adequate for larger datasets, with more variables, as already stated in [16].

Table 3. Examples of two instances existing in *DataSet_NELLValues*.

Variables	Athlete	Sport	Team	Sport position	Country of fans	Equipment
Instances	ahman_green	football	texans	running_back	Germany	cleats
	yogi_berra	baseball	yankees	Catcher	Japan	uniform

Table 4. Bayesian score (g function) for VOMOS and K2 (random VO).

Dataset names	VOMOS	K2 (random VO)
DataSet_NELLValues	−329686.72	−336145.49
DataSet_Relations	−1156046.51	−1158252.47

In addition, and aiming at conducting a more robust comparative analysis, the datasets were also used as input to run the K2 algorithm [19], considered an efficient algorithm when a suitable VO is supplied. Since it is not possible to know such ordering in this case, K2 was performed using a random VO. The results obtained by K2 are also shown in Table 4.

Considering the results presented in Table 4, some observations are possible. Taking into consideration the assessed datasets, the Bayesian scores obtained with VOMOS are better than the ones achieved when using the K2 algorithm. As the Bayesian score reflects the quality of the induced BN structure, it is possible to state that VOMOS tends to induce the best BN.

Figure 2(a) and (b) present the BN structures induced by VOMOS, when applying it to data sets *DataSet_NELLValues* and *DataSet_Relations*, respectively. Observing the induced structures, it is possible to take some conclusions. These BN structures demonstrate the capacity of the VOMOS algorithm for representing the functional relations from NELL's KB.

Observing Fig. 2(a), it is possible to notice that VOMOS identified the dependence relationships among the variable from *DataSet_NELLValues*, which has been built based on some functional relations existing in NELL's KB. Thus, VOMOS demonstrates the ability to represent these relations in a BN structure. VOMOS has identified, for example, the relationship between variables *sport* and *team*, which is represented through an arc in the structure, as can be seen in Fig. 2(a). This arc in the structure reinforces the belief of NELL's system on the relation *teamplayssport* to be correct and so it can learn true values.

The BN structure shown in Fig. 2(b) supports further our thesis that VOMOS may represent functional relations existing in NELL's KB. Notice that *DataSet_Relations* has more variables than *DataSet_NELLValues* and the conjunction of the values of all variables do not suggest relationships among the variables, as done in *DataSet_NELLValues*. Therefore, the structure induced by VOMOS from *DataSet_Relations* reflects its ability in representing dependence relationships among variables from data of the functional relations in NELL's KB. Another interesting issue is that *country* has been the only variable that is not connected in the BN structure (see Fig. 2(b)). This suggests that *country* has no relationship with other variables, according in *DataSet_Relations*. Actually, there is no functional relation defined between *country* and other variables in

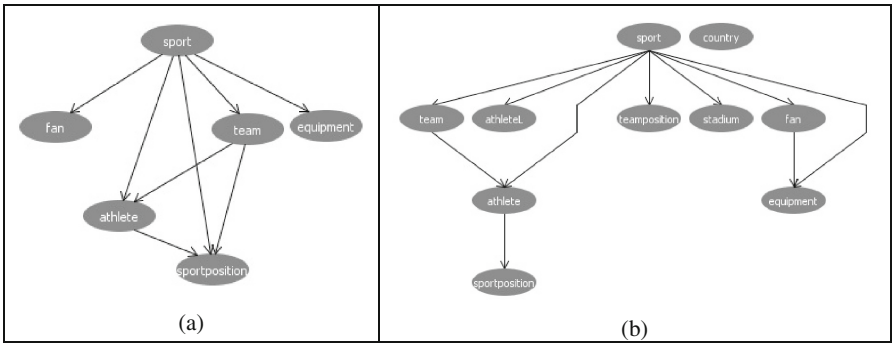


Fig. 2. (a) BN structure induced by VOMOS from the *DataSet_NELLValues* dataset. (b) BN structure induced by VOMOS from the *DataSet_Relations* dataset.

the subset of NELL's KB used in these experiments. However, our concern is also to verify whether VOMOS could induce a BN structure to represent some functional relations, which did not exist in NELL's KB. As an example, we can analyze the arc in Fig. 2(b) between *fan* and *equipment* variables. This arc represents a functional relation not existing in NELL's KB that makes sense. Thus, VOMOS has found a new functional relation that may be inserted in NELL's ontology and that will allow the system to expand its knowledge on both categories. Other example is shown by the arcs between *team* and *athlete* and *athlete* and *startposition*. The arcs indirectly suggest a new functional relation. This same relation is directly represented in structure of the Fig. 2(a) through arc between team and *sportposition*. This will be investigated later.

These results show initial promising empirical evidence on the feasibility of using Bayesian Networks as a tool for representation of knowledge stored through semantic relations in NELL's KB. From this representation, it is still possible to discover new relations to expand its initial ontology. We intend, in sequence, to augment the number of variables in datasets, aiming at to represent and identify different knowledge areas learnt by NELL system.

4 Conclusions and Future Works

This paper proposes a new method for representing and identifying semantic relations from knowledge base on a never-ending system named NELL. This new method employs a Bayesian network learning algorithm called VOMOS aiming to induce Bayesian Networks that identify dependence relationships among variables and thus representing the semantics relations.

The presented results show the feasibility of application of VOMOS for NELL`s knowledge base (KB). In the performed experiments, a subset of NELL's KB was pre-processed for providing datasets to VOMOS. Some relations existing in NELL's KB are drawn and their variables and values form two datasets that serve as input to run VOMOS. From these datasets, VOMOS induced Bayesian Network structures that represent the dependence relationships among the variables that form the semantic

relations. The initial experiments have shown the ability of VOMOS to learn a Bayesian Network capable of identifying the semantic relations existing in NELL's ontology. Besides, VOMOS has been able to find new functional relations that may allow expanding the NELL's initial ontology.

Based on these initial results, we intend to investigate new Bayesian Network learning algorithms (based on VO [22], or not) for being applied to NELL's KB. We also pretend to study inference algorithms aiming at inferring new facts which can populate NELL's KB. In addition, new experiments will be run in future from NELL's KB in Portuguese [13].

Another very interesting line of investigation for future work is to explore relational (and first order) approaches to Bayesian Networks [23], such as the more recent approaches proposed in [24, 25].

Acknowledgments. The authors acknowledge the Brazilian Institutional Program – PIBIC/FAPEMIG/UFSJ – by the scholarship granted through document no. 002/2014/PROPE to develop this research and the Univ. Lyon - UJM-Saint-Etienne (CNRS,Laboratoire Hubert Curien) in France for the support.

References

1. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
2. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr., E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence. AAAI Press (2010)
3. Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Dalvi, B., et al.: Never-ending learning. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pp. 2302–2310. AAAI Press (2015)
4. Carlson, A., Betteridge, J., Hruschka Jr., E.R., Mitchell, T.M.: Coupling semi-supervised learning of categories and relations. In: Proceedings of the NAACL HLT 2009, Work on Semisupervised Learning for Natural Language Processing, New Jersey, pp. 1–9 (2009)
5. Carlson, A., Betteridge, J., Wang, R., Hruschka Jr., E.R., Mitchell, T.M.: Coupled semi-supervised learning for information extraction. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (2010)
6. Appel, A.P., Hruschka Jr., E.R.: Prophet - a link-predictor to learn new rules on NELL. In: Proceedings of the IEEE ICDM2011 Workshop on Data Mining in Networks. IEEE Press, Los Alamitos (2011)
7. Mohamed, T.P., Hruschka Jr., E.R., Mitchell, T.M.: Discovering relations between noun categories. In: Proceedings of the 8th Conference on Emp. Methods in Natural Language Processing, Stroudsburg, pp. 1447–1455 (2011)
8. Pedro, S.D.S., Hruschka, E.R.: Conversing learning: active learning and active social interaction for human supervision in never-ending learning systems. In: Pavón, J., Duque-Méndez, N.D., Fuentes-Fernández, R. (eds.) IBERAMIA 2012. LNCS (LNAI), vol. 7637, pp. 231–240. Springer, Heidelberg (2012). doi:10.1007/978-3-642-34654-5_24
9. Curran, J.R., Murphy, T., Scholz, B.: Minimising semantic drift with mutual exclusion bootstrapping. In: Proceedings of the 10th Conference on the Pacific Association for Computational Linguistics, pp. 172–180 (2007)

10. Hruschka Jr., E.R., Duarte, M.C., Nicoletti, M.C.: Coupling as strategy for reducing concept-drift in never-ending learning environments. Fundamenta Informaticae **124**(1–2), 47–61 (2013)
11. Pennacchiotti, M., Pantel, P.: Entity extraction via ensemble semantics. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 1, pp. 238–247. Association for Computational Linguistics (2009)
12. Caruana, R.: Multitask learning. Mach. Learn. **28**, 41–75 (1997)
13. Duarte, M.C., Hruschka, E.R.: How to read the web in Portuguese using the never-ending language learner's principles. In: 2014 14th International Conference on Intelligent Systems Design and Applications, pp. 162–167. IEEE (2014)
14. Verma, S., Hruschka, E.R.: Coupled Bayesian sets algorithm for semi-supervised learning and information extraction. In: Flach, P.A., Bie, T., Cristianini, N. (eds.) ECML PKDD 2012. LNCS (LNAI), vol. 7524, pp. 307–322. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33486-3_20
15. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo (1988)
16. Santos, E.B., Hruschka Jr., E.R., Ebecken, N.F.F.: Learning Bayesian network structures using multiple offspring sampling. In: Proceedings of the 11th International Conference on Intelligent Systems Design and Applications (ISDA 2011). IEEE Press, Los Alamitos (2011)
17. Miani, R.G.L., Hruschka Jr., E.R.: Exploring association rules in a large growing knowledge base. Int. J. Comput. Inf. Syst. Ind. Manage. Appl. **7**, 106–114 (2015)
18. La Torre, A., Pena, J.M., Muelas, S., Freitas, A.A.: Learning hybridization strategies in evolutionary algorithms. Intell. Data Anal. **14**, 333–354 (2010)
19. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. Mach. Learn. **9**, 309–347 (1992)
20. Gama, J.: Combining classification algorithms. Ph.D. thesis, University of Porto, Portugal (2000)
21. Santos, E.B., Ebecken, N.F.F., Hruschka Jr., E.R., Elkamel, A., Madhuranthakam, C.M.R.: Bayesian classifiers applied to the tenessee eastman process. Risk Anal. **34**, 485–497 (2013)
22. dos Santos, E.B., Hruschka Jr., E.R., do Carmo Nicoletti, M., Ebecken, N.F.F.: The influences of canonical evolutionary algorithm operators and variable orderings in learning Bayesian classifiers from data. Int. J. Hybrid Intel. Syst. **11**(3), 183–195 (2014)
23. Poole, D.: First-order probabilistic inference. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, pp. 985–991. Morgan Kaufmann Publishers Inc. (2003)
24. Kimmig, A., Mihalkova, L., Getoor, L.: Lifted graphical models: a survey. Mach. Learn. **99** (1), 1–45 (2015)
25. Schulte, O., Qian, Z., Kirkpatrick, A.E., Yin, X., Sun, Y.: Fast learning of relational dependency networks. Mach. Learn. **103**(3), 377–406 (2016)

Training a Spiking Neural Network to Generate Online Handwriting Movements

Mahmoud Ltaief(✉), Hala Bezine, and Adel M. Alimi

REGIM-Lab.: Research Groups in Intelligent Machines,
University of Sfax, ENIS, BP 1173, 3038 Sfax, Tunisia
mahmoud.ltaief@ieee.org

Abstract. In this paper we developed a spiking neural network model that learns to generate online handwriting movements. The architecture is a feed forward network with one hidden layer. The input layer uses a set of Beta elliptic parameters. The hidden layer contains both excitatory and inhibitory neurons. Whereas the output layer provides the script coordinates x(t) and y(t). The proposed spiking neural network has been trained according to Sander Bohet model. The trained spiking neural network has been successfully tested on MAYASTROUN data base. Also a comparative stady between the proposed spiking neural network and an artificial neural network proposed in a previous work is established.

Keywords: Spiking neural network · Artificial neural network · Beta-elliptic model · Handwriting generation · Similarity degree

1 Introduction

Spiking neural networks (SNN) is a particular category of artificial neural networks (ANN), in which neurons intercommunicate by sequences of spikes. Spiking neural networks are able to handle large quantity of data using relatively a few number of spikes [20]. Because of their functional likeness to biological neurons, spiking models provide efficient tools for analyzing the basic processes in the brain, comprising neural plasticity, learning and information processing.

Spiking neural networks are designed to model more closely the human observed behavior and neuron. There are many parts that are computationally important; the neuron itself, the synapses connecting the neurons and the signal or spike which moves between neurons through the synapses.

The neuron: A neuron will send out an action potential, or spike along her axon when the inputs (dendrites) become sufficiently excited. There is a threshold value which must be surpassed by the sum of the inputs before the neuron discharge a spike.

The synapse: Synapses are the connections between neurons through which the spike circulates from one neuron to another. A synapse has the capability to

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_29

transfer a spike from the output of one neuron to the input of another neuron. This capability is known as the synaptic efficacy of the synapse. The synaptic efficacy changes from spike to spike even when no learning is involved [13]. Some models for example, Natchlager 2001, use the timing between spikes to change the synaptic efficacy [14]. It is clear that the synapse performs computations, in addition to those of the neuron itself.

The spike: The spike or action potential is of short duration (1 ms) and unvarying shape and amplitude, even during its transmission along the axon. Just after this transmission, the potential drops below the resting level (hyper-polarization) and after a refractory phase where the neuron is no longer excitable, it joined its resting level and can restart the cycle. Is the arrival of an action potential in the axon terminals that is causing the release of neurotransmitters in the synaptic cleft, which will be collected by the postsynaptic neuron and produce a post-synaptic potential (PSP) at the postsynaptic dendritic membrane. The inhibitory or excitatory nature of the synapse will distinguish between postsynaptic excitatory or inhibitory potential (EPSP or IPSP). Such potential is low amplitude, relatively extended in time and delayed relative to the spike reception time. Computational models of spiking neuron have the advantage of providing a temporal integration of information. The response of the neuron to stimulation is no longer a value representing the discharge rate (threshold or sigmoidal neuron principle), but the emission time of the next spike. The information transmitted is not clocked by the presentation of a stimulus. A spike is emitted at the moment neuron excitation exceeds the threshold. It is therefore necessary to adopt a virtual clock that, within a ANN (Artificial Neural Network) to delay emission spikes of neurons, which can then be considered dated events. The spiking neuron models have the property to detect the temporal correlation of the associated action potentials, surpassing in this classical neuron models. The spiking neuron models proposed in the literature can be grouped into four main categories:

The Hodgkin & Huxley model (HH) [8]: This mathematical model describes the temporal dynamics of the neuron by the variations of ion concentration (K +, Ca2 +) in the different compartments and reproduces primary "modes" of operation of the biological neuron.

The Integrate & Fire model (IF) and its derivatives: Lapicque proposed in 1907 (article translated into English by Brunel & van Rossum) [5] moves to a modeling detail level less accurate than Hodgkin & Huxley model. The neuron is modeled as an electrical circuit comprising a capacitor and a resistor (Rc dipole). Thus, the dynamics of the neuron is described by its membrane potential and the current flowing therethrough. The model is described mathematically by a single differential equation. This model does not account for changes in membrane potential but only spikes emissions.

The Spike Response Model (SRM) [6]: it is a phenomenological model of the neuron in the sense that it does not consider the underlying mechanisms, but only the behavior of the neuron. The formalization of this model consists of

a core function giving the dynamics of the membrane potential of the neuron, and a function calculating the summation of post-synaptic potentials (PPS). Its mathematical formulation is derived from an integration of the underlying differential equations, which makes it closer to that of classical neurons, even if it is based on time. We will detail this model in the next section.

The Izhikevich model [9]: involves two coupled differential equations, in order to model the flow of ions and the dynamics of neuron potential. It reproduces twenty different dynamic behaviors, which makes it valuable for modelers who want their models remain close to biological observations.

In the present paper we use the spike response model for our neural network trained with the supervised learning rule, "SpikeProp", which is a generalization of back-propagation for spiking neurons [4].

2 The Spiking Neural Network Architecture

The architecture of our network consists of a feed-forward spiking neural network see Fig. 1. The input layer is composed of 11 neurones, the first 10 neurones are the Beta-elliptic parameters (the static parameters given by elliptic equation and the dynamic parameters given by Beta equation). Also we use one neurone for the timing network.

The timing network is used as a synchronizer between the input and output of our network: in the parameters extraction stage every stroke was described with a set of beta profile and set of elliptic arc. Each couple (beta profile, elliptic arc) belongs to a time interval $[t_0, t_1]$. The timing network duplicates the parameters given by each pair (Beta profile, elliptic arc) in the appropriate time interval. Duplication is done so that each point $(x(t), y(t))$ of the network output corresponds to an input (the beta elliptic parameters) at time t. In our neural network we use one hidden layer with 10 neurones (7 excitatory neurones and 3 inhibitory neurones). Finally the output layer contain 2 neurones that represent the coordinates of script pen tip $(x(t), y(t))$ estimated by the network. In our spiking neural network every two interconnected neurons have a number of m=8 synaptic terminals with variant delays d^k and an individual weight Fig. 5(B). Between the firing time of the pre-synaptic neuron, the post-synaptic potential begins to rise or fall if the neuron is excitatory, or inhibitory. Inhibitory neurons generate only negative post-synaptic potentials [4]. The neuron activity is described by the Spike Response Model [7], describing the neuron with one variable, the membrane potential $x(t)$. A neuron j receives input from a set of pre-synaptic neurons $i \in \Gamma_j$ is characterized by the variable $x_j(t)$. In that case, the effects of the arrived spikes are summed. If this sum (noted $x_j(t)$) achieves the threshold v a potential action is triggered (see Fig. 2). The membrane potential of one neuron j is described with the equation below:

$$x_j(t) = \sum_{i \in \tau_j} \sum_k w_{ij}^k y_i^k(t) = \sum_{i \in \tau_j} \sum_k w_{ij}^k \varepsilon\left(t - t_i - d^k\right) \tag{1}$$

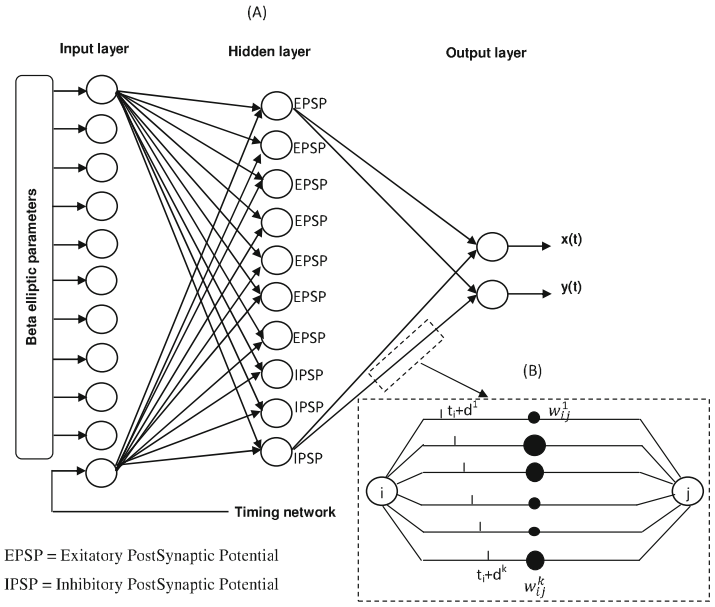


Fig. 1. (A) The spiking neural network architecture. (B) connection consisting of multiple delayed synaptic terminals.

Where w_{ij}^k is the synaptic weight. The sum covers every presynaptic spikes where Γ_j is the totality of the presynaptic neurons of neuron j, and t_i is the firing time of presynaptic neuron $i \in \Gamma_j$. The term w_{ij}^k represents the weight of the synaptic terminal k, between the neurons j and i, having the delay d_k.

3 The Learning Algorithm

The learning algorithm only considers the spikes of each neuron arriving before the action potential exceeds the threshold. So, the shape of the post-potential and all other following spikes are ignored. The spike response model is defined by the Eq. 2 and the threshold criteria [4].

The spike response function $\varepsilon(t)$, at time t can be modelled by the equation bellow.

$$\varepsilon(t) = \begin{cases} \frac{t}{\tau} \exp\left(1 - \frac{t}{\tau}\right) \text{ if } t > 0 \\ 0 \text{ if } t \leq 0 \end{cases} \tag{2}$$

With τ the modeling of time-decay constant of the membrane potential.

The learning process is to explicitly assess the gradient error relative to the weight of every synaptic terminal. The error Eq. 3 is defined as the sum of the squared differences between the actual spike times t^a and the desired spike times t^d [4].

$$E = \frac{1}{2} \sum_{j \in J} \left(t_j^a - t_j^d\right) \tag{3}$$

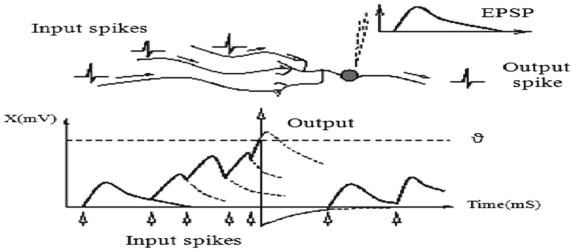


Fig. 2. Response of the membrane potential to incoming spikes. If the threshold ϑ is crossed, the membrane potential is reset to a low value, and a spike is fired.

In this paper Error-backpropagation learning algorithm is used to update each synaptic weight w_{ij}^k to reduce the error E. The E variation rate with respect to w_{ij}^k must be known. The back-propagation algorithm derived from the same way as Bohte [4]. We note that for the hole algorithm we denoted by:

H:neurons in the input layer; $h \in H$.
I:neurons in the hidden layer; $i \in I$.
J:neurons in the output layer; $j \in J$.

For each synaptic terminal we use the updating rule (Eq. 4) to calculate the new weights:

$$\Delta w_{ij}^k = -\eta \frac{\partial E}{\partial w_{ij}^k} \tag{4}$$

Where w_{ij}^k is the synaptic weight of sub-connection from neuron i to neuron j, and η is the learning rate. The derivative of E with respect to w_{ij}^k can be developed as:

$$\frac{\partial E}{\partial w_{ij}^k} = \frac{\partial E}{\partial t_j^a}\left(t_j^a\right) \frac{\partial t_j}{\partial x_j\left(t\right)}\left(t_j^a\right) \frac{\partial x_j\left(t\right)}{\partial w_{ij}^k}\left(t_j^a\right) \tag{5}$$

Where t_j is expressed as a function of the membrane potential $x_j\left(t\right)$ around $t = t_j^a$. To simplify the evaluation of the gradient, it is assumed for a small area around t_j^a the function $x_j\left(t\right)$ is treated like a linear time function. Therefore, the local derivation of t_j is constant.

The equation of the error backpropagation derived for a feed-forward network fully connected is:

$$\frac{\partial E}{\partial w_{ij}^k} = y_{ij}^k\left(t\right)\delta_j \tag{6}$$

Where y_{ij}^k is the signal spread from neuron i to j through connection k.

For output layer neurons, δ_j is:

$$\delta_j = -\frac{t_j^a - t_j^d}{\sum_{i\in\Gamma_j}\sum_{k=1}^{m} w_{ij}^k \frac{\partial y_i^k\left(t_j^a\right)}{\partial t_j^a}} \tag{7}$$

The weights updating for the output layer is as follows:

$$\Delta w_{ij}^{k} = -\eta y_{i}^{k}\left(t_{j}^{a}\right)\delta_{j} \tag{8}$$

For the hidden layer neurons, δ_i is:

$$\delta_{i} = -\frac{\sum_{j\in\Gamma^{i}}\delta_{j}\sum_{k=1}^{m} w_{ij}^{k}\frac{\partial y_{i}^{k}\left(t_{j}^{a}\right)}{\partial t_{i}^{a}}}{\sum_{h\in\Gamma_{i}}\sum_{k=1}^{m} w_{hi}^{k}\frac{\partial y_{h}^{k}\left(t_{i}^{a}\right)}{\partial t_{i}^{a}}} \tag{9}$$

Where the set Γ^i represents all the postsynaptic neurons of the neuron i.

The weights updating for the hidden layer is as follows:

$$\Delta w_{hi}^{k} = -\eta y_{h}^{k}\left(t_{i}^{a}\right)\delta_{i} \tag{10}$$

With η the learning rate.

The SpikeProp algorithm	
1	The firing time of each neuron is found when $x_j(t)$ using the Eq. 1
2	Calculate δ_j for every output neurons in J, with the Eq. 7. Using the desired and actual firing time of each output neurons
3	Calculate δ_i for all neurons in I with the Eq. 9. Using the firing time of each neurons in hidden layer and the δ_j values
4	For the output layer J, update w_{ij}^k by the Eq. 8
5	New weight= Old weight + Δw_{ij}^k
6	For the hidden layer I, update w_{hi}^k by the Eq. 10
7	New weight= Old weight + Δw_{hi}^k

4 Application to Handwriting Generation

The suggested model is concerned to online handwriting generation and it is inspired from the Beta-elliptic model for complex handwriting movements modeling. In our model, a handwritten stroke is approximated in the dynamic domain by a Beta profile which in turn corresponds to an elliptical arc in the trajectory domain. Also, a stroke runs from random starting position is typified by ten parameters. The first five parameters concern neuromuscular networks synchronization properties involved in the movement generation, although the final five parameters depict the geometrical properties of the generated trace. As illustrated in Fig. 1, these parameters are used as input for the spiking neural network to produce the initial script.

4.1 The Beta-Elliptic Model for Handwriting Movements Modelling

The beta-elliptic model is founded on certain hypothesis: first of all, it supposes handwriting movement like all other extremely qualified motor process is partly programmed beforehand. Secondly, it assumes that movements are planned and represented in the velocity field since the invariant most accepted in handwriting movement generation is the beta function for velocity profile modeling. In its most simple form, the model is funded on a beta equation β(t, t_0, t_1, t_c, p, q), where t_0 is the beginning time, t_1 is the ending time, p and q are intermediate parameters. This equation represents in the kinematics domain the velocity prole which is in turn described by an elliptic arc which typify the trajectory in the static domain [1–3]. For an elementary movement named stroke the Beta-elliptic model produces a set of ten parameters that characterizes the movement in both fields static and kinematic. Then for each stroke we have ten parameters (t_c, p, q, t_0, t_1, a, b, x_0, y_0, θ). These parameters will be used as input for our spiking neural network for complex handwriting movements generation [10,11].

5 Simulation Results

Mayastroun database [15] is used to test the performance of our model. This database was written by many writers, and composed by a large variability of online handwriting scripts. It contains 5000 isolated digits, 13036 isolated lower and upper case letters, and 35017 isolated words from a 500 word lexicon (English, French and Arabic) are collected using digitizing tablet. To train our spiking neural network 1000 scripts (digits, Latin and Arabic words and letters) are used for learning and 500 scripts for testing.

Simulation results are given in the following figures (Fig. 3), where represented respectively the initial scripts (A) and generated script (B) of digits, Latin and arabic lettre, and Latin and Arabic word. We note a good agreement between initial scripts and the scripts generated with the spiking neural network. Also we note the efficacy of our model to generate both type of scripts Latin and Arabic. In the applied field, the potential of the current model to generate synthetic writing can probably be operated as a generator of "handwritten CAPTCHAs" [16].

5.1 Similarity Degree Measure

We measure the similarity degree between the original script and the produced script to properly assess the accuracy of our spiking neural network for online handwriting generation. We use the equation below to evaluate our model [21].

$$S\left(S_o, S_g\right) = 1 - \frac{\sqrt{\sum_{i=1}^{n}\left(x_{io} - x_{ig}\right)^2 + \left(y_{io} - y_{ig}\right)^2}}{2n} \tag{11}$$

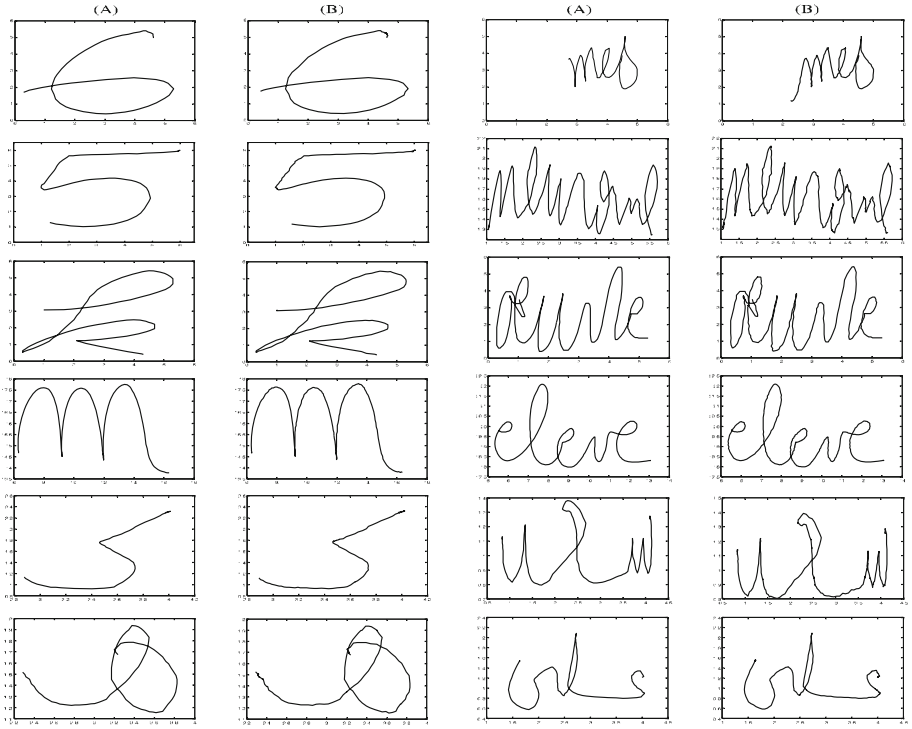


Fig. 3. (A) The original simple scripts and (B) the generated simple scripts

Where:

$S(S_o, S_g)$ is the similarity degree between original script S_o and generated script S_g.
(x_{oi}, y_{oi}) are the coordinates of the i^{th} point in the original script.
(x_{gi}, y_{gi}) are the coordinates of the i^{th} point in the generated script.
n is the length of the original script.

We note that:

- $S(S_o, S_g) \in [0, 1]$;
- $S(S_o, S_g) = 1$ then $S_o = S_g$;
- $S(S_o, S_g) = \alpha$ then S_g (generated script) is similar with α degree to the S_o (original script).

Table 1 represents the similarity degree measure between original scripts (the first colon) and the scripts generated by the spiking neural network (second colon) and the scripts generated by the artificial neural network (ANN) (third colon). This ANN which is proposed in previous work [11] contains an input layer, four hidden layers each one is composed with 10 neurons, and an output layer. Referring to this measure we can note the reliability of the proposed spiking

Table 1. The measure of similarity degree.

Scripts	The similarity degree (SNN)	The similarity degree (ANN)
Digit: 6	0.9192	0.8792
Digit: 5	0.9143	0.8531
Lettre: k	0.9097	0.8602
Lettre: m	0.9134	0.8687
Letter: ک	0.9076	0.8542
Letter: ھ	0.8902	0.8234
Word: normes	0.8822	0.8227
Word: neurone	0.8607	0.8020
Word: oeuvre	0.8873	0.8193
Word: eleve	0.8903	0.8320
Word: سعد	0.8694	0.8041
Word: على	0.8804	0.8067

neural network for handwriting generation compared to the result given by the ANN. For our proposed SNN we remark that the similarity degree between generated and initial scripts is of the order 0.87. Which confirms that the spiking neural network can generate scripts with a relatively good performance.

6 Conclusion

In this paper we presented a spiking neural network model for handwriting movements generation. In this model, script velocity is expressed as a Beta-elliptic parameters. This parameters was used as input for the spiking neural network to generate the script pen tip coordinates, $x(t)$ and $y(t)$. Our model which is a multilayer network with one hidden layer is successfully tested on MAYASTROUN database [15] composed with Latin and Arabic letters, digits and words. This is confirmed by the similarity degree measure. In deed, in all simulation examples this degree was never less than 0.85. The efficiency and the robustness of the proposed spiking neural network is also demonstrated by comparison with an artificial neural network proposed in previous work [11]. The proposed spiking neural network can also be trained on data from a specific individual.

References

1. Alimi, M.A.: Beta neuro-fuzzy systems. TASK Quarterly J. Spec. Issue Neural Netw. **7**(1), 23–41 (2003). Duch, W., Rutkowska, D. (eds.)
2. Bezine, H., Alimi, M.A., Sherkat, N.: Generation and analysis of handwriting script with the Beta-elliptic model. In: Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition IWFHR04, Tokyo, Japan, pp. 515–520 (2004)

3. Bezine, H., Kefi, M., Alimi, M.A.: On the Beta-elliptic model for the control of human arm movements. IJPRAI **21**(1), 5–19 (2007)
4. Bohte, S., Kok, J., Poutre, H.L.: Error backpropagation in temporally encoded networks of spiking neurons. Neurocomputing **48**, 17–37 (2002)
5. Brunel, N., Van Rossum, M.C.: Lapicques 1907 paper: from frogs to integrate-and-re. Biol. Cybern. **97**, 337–339 (2007)
6. Gerstner, W., Kistler, W.: Mathematical formulations of Hebbian learning. Biol. Cybern. **87**, 404–415 (2002)
7. Gerstner, W.: A framework for spiking neuron models: the spike response model. In: Moss, F., Gielen, S. (eds.) The Handbook of Biological Physics, vol. 4(12), pp. 469–516. Elsevier Science (2001)
8. Hodgkin, A.L., Huxley, A.F.: Currents carried by sodium and potassium ions through the membrane of the giant axon of Loligo. J. Physiol. **116**, 449–472 (1952)
9. Izhikevich, E.M.: Simple model of spiking neurons. IEEE Trans. Neural Netw. **14**(6), 1569–1572 (2003)
10. Ltaief, M., Njah, S., Bezine, H., Alimi, M.A.: Genetic algorithms for perceptual codes extraction. J. Intell. Learn. Syst. Appl. JILSA **4**, 256–265 (2012)
11. Ltaief, M., Bezine, H., Alimi, M.A.: A neuro-Beta-elliptic model for handwriting generation movements. In: International Conference on Frontiers in Handwriting Recognition ICFHR, pp. 799–804 (2012)
12. Ltaief, M., Bezine, H., Alimi, M.A.: A spiking neural network model for complex handwriting movements generation. Int. J. Comput. Sci. Inform. Secur. (IJCSIS) **14**(7), 319–327 (2016)
13. Maass, W.: Neural computation: a research topic for theoretical computer science? Some thoughts and pointers. In: Current Trends in Theoretical Computer Science, Entering the 21th Century, pp. 680–690 (2001)
14. Natschlaeger, T., Maass, W.: Spiking neurons and the induction of finite state machines. Theor. Comput. Sci. Spec. Issue Nat. Comput. **287**, 251–265 (2002)
15. Njah, S., Ben Nouma, B., Bezine, H., Alimi, M.A.: MAYASTROUN-a multilanguage handwriting database. In: International Conference on Frontiers in Handwriting Recognition ICFHR, Bari Italy, pp. 308–312 (2012)
16. Rusu, A., Govindaraju, V.: Handwritten CAPTCHA: using the difference in the abilities of humans and machines in reading handwritten words. In: 9th International Workshop on Frontiers in Handwriting Recognition, pp. 226–231 (2004)
17. Schomaker, L.R.B.: From handwriting analysis to pen-computer applications. Electron. Commun. Eng. J. **10**(3), 93–102 (1998)
18. Schomaker, L.R.B.: Simulation and recognition of handwriting movements: a vertical approach to modeling human motor behavior, Ph.D. Thesis, Nijmegen University. Netherlands (1991)
19. Teulings, H.L., Thomassen, A., Schomaker, L.R.B., Morasso, P.: Experimental protocol for cursive script acquisition: the use of motor information for the automatic recognition of cursive script. Report 3.1.2., ESPRIT Project, 419 (1986)
20. VanRullen, R., Guyonneau, R., Thorpe, S.J.: Spike times make sense. Trends Neurosci. **28**, 1–4 (2005)
21. Yanhong, L., David, L.O., Zheng, Q.: Similarity measures between intuitionistic fuzzy (vague) sets: a comparative analysis. Pattern Recogn. Lett. **28**, 278–285 (2007)

Using a Synthetic Character Database for Training Deep Learning Models Applied to Offline Handwritten Recognition

Jorge Sueiras, Victoria Ruiz, Ángel Sánchez, and José F. Vélez(✉)

Technical School of Computer Science, Universidad Rey Juan Carlos, c/Tulipan s-n, 28933 Mostoles, Madrid, Spain
{jorge.sueiras,victoria.ruiz,angel.sanchez,jose.velez}@urjc.es

Abstract. We present our current work on building a deep learning architecture for the offline handwritten character recognition problem. The proposed system is based on training a deep Convolutional Neural Network (CNN) to recognize handwritten characters, using a new synthetic character database derived from UNIPEN dataset. The presented approach is inspired in some successfully-used neural architectures for image classification, specially the VGG-CNN. Our system reads each word with the help of a sliding window in a similar way to how humans do. An innovative feature of our proposal is using a synthetic character database specifically built, in a optimized way, to identify the characters as component elements of the words. Experiments with this new training synthetic dataset produced recognition rates of 98.4% for uppercase and 96.3% for lowercase, respectively.

Keywords: Document analysis · Handwritten character recognition · Synthetic character database · Convolutional neural networks · Deep learning

1 Introduction

The offline handwritten text recognition problem [7], which consists in the automatic transcription of images of handwritten text, has been extensively studied from many years [14,17,19]. The significant differences in the writing of many individuals and the cursive nature of handwritten characters make this recognition task hard and still it remains as an open research problem.

In this work, we focus on the offline problem of recognizing handwritten characters. The main objective is building a character recognition model, general enough to be used as part of a word recognition system without needing to segment the words into its component characters. Thus, we use the proposed model over a sliding window, which runs along the words, in order to provide evidences of the occurrence of each character sequentially. So, the complete word is finally recognized.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_30

The task of building a character-based recognition model, that can be effective in characters included into a word without segmenting it, is difficult since most handwritten character databases contain only isolated characters [5,9]. For this reason, we built a new specific corpus of handwritten characters specially adapted to our problem where characters in a word can appear as joined.

Neural networks have been used along many years for the offline handwritten character recognition problem [12]. Good recognition results were achieved when training and evaluating over the same database. Different handwritten character databases were used for this problem. For example, Yuan et al. [19] used CNN on the UNIPEN offline database, Ciresan and collaborators [2] built ensembles of CNNs over the NIST database, and Van der Maaten [17] used neural networks from 2 to 5 layers over the TICH database. In all of the cases the authors offered their results by training and evaluating over a train-test partition from the same database. They did not offer cross-database evaluations to learn the models built over one corpus and to validate in another different one. We are interested in this type of evaluations since our method includes the use of a character-level model which is built on a different corpus of handwritten words.

We also applied the model presented in this paper to the handwritten word recognition problem. In previous papers, some authors used methods based on identifying characters [1], and other authors [10] used methods based on sliding windows for extracting features from the vertical lines of the image. In our future work, we are planing to combine these two approaches by applying a character-level model over a sliding window of the images of words in order to recognize the sequence of characters present in these words. At the end of this paper, we sketch a possible approach to perform this task.

In this paper, we built a new character recognition model and validate its generalization capacity using a different corpus. Our objective is building a new approach to handwritten word identification that avoids the character segmentation problem. To develop this approach, we have created a new offline handwritten character corpus using the UNIPEN database [9] and adding, in a synthetic way, artifacts to each character image. We aim to approximate the effect of the character shape when it is placed inside a word, basically by moving and truncating it. Also, we have added some artifacts to the left and right of each character. The main reason to create this new dataset is that in the character-level model we can find characters included as parts of words using a sliding window, and, therefore, these characters can appear as moved, truncated at left or right, and joined to other characters.

In order to build our model, we created a new architecture based on CNNs due to the recent impressive results produced by these networks in image classification problems, in general [11], and also in handwriting digit recognition [5] and character classification problems [20], in particular. Based on these advances, we have designed our CNN architecture to build the handwritten character image classification model which is based in the VGG network [16] including several layers of stacked convolutions of small size of 3×3 joint into two final dense layers. We provide our results on this model, which was applied for two different input image resolutions of 64×64 and 32×32 pixels, respectively.

Finally, we also present the first stages on how to apply the previous model to obtain a representation of word images as sequences of probabilities for each character, detected in word images using a sliding window. The initial results are promising and we are currently working in using this word representation to combine it with decoders as CTC [6] and with language models to develop an end-to-end word handwriting image classifier.

Although the goal of our current research is providing a general solution to the handwriting word recognition problem, this paper is mainly focused on a preliminary stage which includes: the creation of a new handwritten character database, the proposal of a classification model on this database and the evaluation of the whole solution on different public handwritten character corpuses. Our approach was applied over a dictionary of the 52 English uppercase/lowercase characters. In any case, the methodology presented can be used on other language characters with minimal changes and starting with a character corpus for the considered language.

2 Handwritten Character Database

In this section, we describe the steps followed to create the new dataset of handwritten characters. It is specially oriented to train a model that works with characters, but as a component of a more general system, to automatically recognize handwritten words. Our database was built in two steps:

- First, we extract isolated characters from the UNIPEN online handwritten database.
- Next, we perform several transformations on the initial character images to simulate how these characters can appear inside real words.

The UML activity diagram shown in Fig. 1 illustrates the stages involved for producing the new character database. The next two subsections describe in detail these stages.

2.1 Extraction of Characters from UNIPEN Online

By using this online database to extract the "base" characters, we avoid the problem of character segmentation in offline handwritten text and we also obtain perfect isolated characters. The isolated annotated characters are first identified. Next, these characters are retouched by hand so that, the x and y coordinates of pixels are connected in each stroke which shapes the character. Variable-thickness strokes are used depending on the original resolution of the characters to ensure that all final characters have similar thickness. The images from each available category (i.e. different types of alphabet characters) are generated: uppercase, lowercase, digits and punctuation marks. These generated images are resized to 64×64 pixels without changing their aspect ratio. Finally, the generated images are curated manually, one by one, to make sure that they are

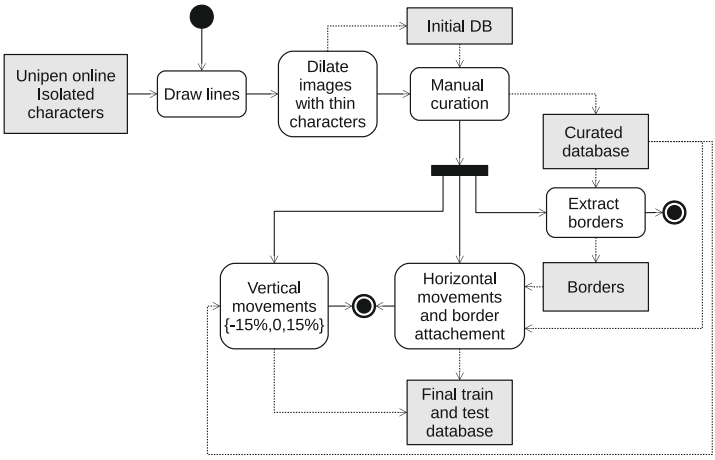


Fig. 1. UML activity diagram of the process of creation and transformation of the character database. Process are presented as white boxes and outputs as gray boxes.

assigned to their correct category and are human-legible. Our database contains 93 categories and a total of 68,382 image characters, where the average number of samples per category is 735. Table 1 illustrates the types of characters present in our dataset. In the experiments of this work, we only used the uppercase and lowercase image letters with a total of 46,102 image samples belonging to 52 classes. In order to facilitate the reproducibility of the results, our database of curated characters can be downloaded from: https://github.com/sueiras/handwritting_characters_database. We also provide here a Python code that implements the transformations described in the next paragraph.

2.2 Transformations Applied to Curated Original Images

A set of transformations (see Fig. 2) were applied to the original character images with two goals: augmenting the training set size in experiments and getting good results with the models when they are applied on parts of word images (i.e. at the word level), where the characters are not isolated.

The applied transformations are the following ones. First, we resize characters and translate them up or down. Specifically, we reduce the characters a 25% of its original size and the new image is moved up or down for each character. The main idea is that some characters only can appear up or centered in a word, while others can only appear down and centered, and finally other ones can appear in any place. The characters included in each case are:

- The letters: ‘a’, ‘c’, ‘e’, ‘i’, ‘m’, ‘n’, ‘o’, ‘r’, ‘s’, ‘u’, ‘v’, ‘w’, ‘x’, ‘z’ are placed up centered and down.
- The letters: ‘g’, ‘j’, ‘p’, ‘q’, ‘y’ are placed centered and down.

– Finally the letters: ‘b’, ‘d’, ‘f’, ‘h’, ‘k’, ‘l’, ‘t’, ‘A’, ‘B’, ‘C’, ‘D’, ‘E’, ‘F’, ‘G’, ‘H’, ‘I’, ‘J’, ‘K’, ‘L’, ‘M’, ‘N’, ‘O’, ‘P’, ‘Q’, ‘R’, ‘S’, ‘T’, ‘U’, ‘V’, ‘W’, ‘X’, ‘Y’, ‘Z’ are placed up and centered.

Second, move left or right the character image and include, on the border of each character, the beginning or the end of other character. So that, similar training examples are going to be available to those that are going to be found when applied the model on a piece of a word image. To achieve this, the borders of the characters (with 3 pixels of thickness) are extracted from the available sample. These borders are added to the extremes of the existing characters in order to generate a new sample. In this process, the no-overlapping between the characters and the added borders is guaranteed.

Table 1. Characters contained in the database.

| | | |

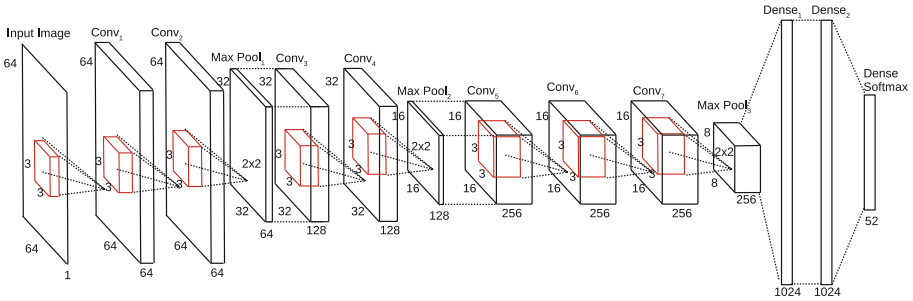


Fig. 3. Architecture model with the sizes of the layers and the sizes of the convolutional masks (in red). (Color figure online)

units (ReLU) [11]. Finally, we use the dropout regularization [15] in the first two fully-connected layers with a ratio of 0.5. We trained the model over 5 epochs with a learning rate of 0.01 and over other 5 epochs with a learning rate of 0.001. A momentum of 0.9 and the Nesterov technique [13] were used. The model obtained an accuracy of the 87.5% on the test data.

To determine the importance of the resolution of the input images, we built the previous model with the same architecture and the same training parameters but only changing the resolution of initial images to 32×32 pixels. A recognition accuracy in test images of 86.4% was achieved for this new image resolution.

To evaluate the generalization capacity of the presented model over other different datasets and compare our results with other published ones [19], we trained the model separately for uppercase and lowercase letters. Many public character benchmarks include separate error data for uppercase and lowercase [3]. Using the UNIPEN database, Yuan et al. [19] applied the LeNet-5 CNN for offline handwritten English character recognition. They reported recognition rates of 93.7% for uppercase and 90.2% for lowercase, respectively. In our experiments, on a test separated sample of the 20% of the original images of our corpus extracted from UNIPEN online, we achieved respective recognition rates of 98.4% for uppercase and 96.3% for lowercase.

Additional experiments were carried out over the TICH database [17]. Van der Maaten provided a best error recognition result of $82.77\% \pm 0.82\%$ in recognition using $10 - fold$ cross-validation procedure and a k-NN classifier. Our uppercase model (note that the TICH dataset contains only uppercase) trained with our database and test with TICH, produced a recognition result of 92.5% using the complete dataset.

There exists several benchmarks for using the NIST database (2nd edition) [8]. One of the most recent and accurate ones that uses deep networks is [3], which provided recognition rates of 98.17% for uppercase and 92.53% for lowercase. Our model, trained with our dataset and validated with this NIST benchmark, obtains recognition rates of 92.9% for uppercase (i.e. these results are very close to those obtained with the TICH database) and 71.5% for lowercase. In this case, the big differences between uppercase and lowercase characters can be explained

Table 2. Summary of character recognition results for different databases.

	LeNet-5 [19]	k-NN [17]	CNN [3]	Our proposal
UNIPEN upper case	93.7%	-	-	98.4%
UNIPEN lower case	90.2%	-	-	96.3%
TICH database	-	82.77%	-	92.5%
NIST database (2nd ed) upper case	-	-	98.17%	96.9%
NIST database (2nd ed) lower case	-	-	92.53%	90.4%

because our database has gray scale images, while the NIST database uses binary images. By training the same architecture using the NIST dataset (i.e. as train and test sets), we obtained recognition rates in NIST of 96.9% for uppercase and 90.4% for lowercase, respectively. All these results are summarized in Table 2.

4 First Steps for Word Image Recognition

In order to apply the character-level model to recognize word images, we define an algorithm that receives as input the image of a word. This input has a height of 64 pixels and a variable width of w pixels. By other way, the algo rithm returns as output an array of size $52 \times w$. Each column of this array will contain the likelihood of each character in the dictionary (26 uppercase and 26 lowercase letters) when it appears in the corresponding column of the original image (see Fig. 4). Basically, the algorithm applies the character model over the word image using several sliding windows and aggregates the model results in the final output. It has the following stages:

1. First, we attach to the left and right of the word image two 8-pixel-width bands of background in order to better include the start and end of the image in the sliding process. This can help to improve the detection of the final and initial characters of the word.
2. Next, we define several sliding windows to go over the word image from left to right. These windows have different widths (w). In particular, we use the following ones: 32, 40, 48, 56 and 64, respectively. The goal is to capture characters with approximately these widths.
   - For window widths less than 64 we need to complete them to a resolution 64×64 because this is the input shape expected by the character classification model. In order to do this, we complete the windows image with background bands to the left and right of it.
3. We apply the character-level model to the previous windows by sliding them over the word image. So, we obtain for each window a sequence of scores of the characters that the model can detect with this window.
   - For each initial image column, this appear in several windows with different sizes and receive several scores of the character model. In order to obtain a final prediction of that character to appear in this column, we

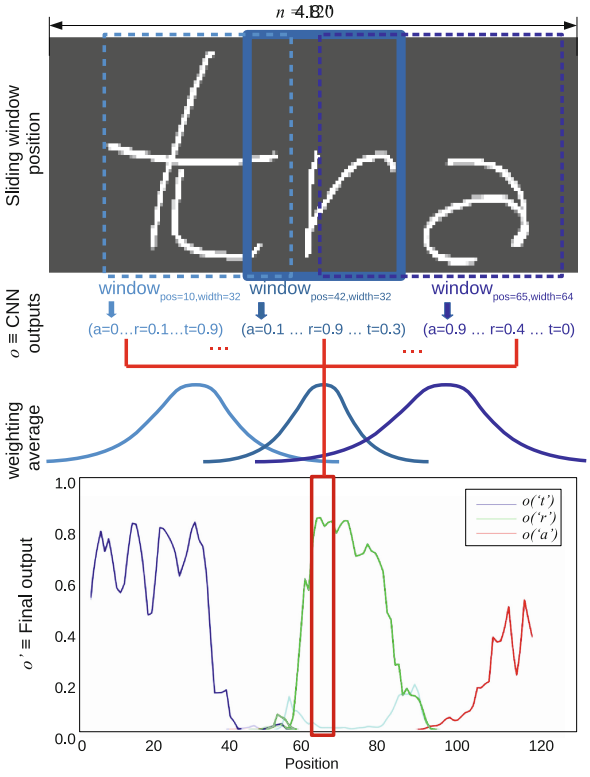


Fig. 4. Result of applying the proposed character recognition system over the image of a word. For simplicity, only three windows are averaged in the figure. Observe the different widths of the windows.

aggregate all the model scores that include this column. In the aggregation, we apply a weighted average using a normal distribution over each snapshot that includes this column to give more weight to the central columns of each window.

The previous algorithm can be described by Eq. 1 where n is the width of the input image, $o(i)$ corresponds to the CNN output when the input of the network is in position i, $N(x, i, w/2)$ is the value of a normal distribution in x with mean i and standard deviation $w/2$.

$$o'(x) = \frac{1}{|w|} \sum_{w \in \{32,40,48,56,64\}} \left(\sum_{i=0}^{n} o(i) \cdot N(x; i, w/2) \right) \tag{1}$$

5 Conclusion

This paper described a system designed to obtain a representation of any handwritten word image based on characters without segmenting it. It included a new

handwriting character image database and a new classification model to handwritten characters. We have shown how it is possible to use this representation to decode any word image. Finally, the proposed system does not need from any training process to be used with any specific English corpus.

As future work, we plan to improve the decoding process of the word images to build a complete and accurate handwriting offline word recognition system. It will be directly applicable over any corpus with English characters. Currently, we are working in the two following research lines:

- Apply a decoding model with the connectionist temporal classification (CTC) output layer [6] to improve the process of word decoding. This model can decode the representation of the word obtained by applying the character model (in the sliding window) to the correct word.
- Use the context of any given word in the corpus to improve the word recognition. To achieve this, a language model that considers the previous and the next words predictions to improve the current word recognition, will be included. In this point we are working with Recurrent Neural Networks (RNN) models using Long Short-Term Memory (LSTM) units [18].

Acknowledgements. This work was funded by the Spanish Ministry of Economy and Competitiveness project number TIN2014-57458-R and by the URJC-Banco de Santander Excellence Research Groups grant number 30VCPIGI09.

References

1. Bengio, Y., LeCun, Y., Nohl, C., Burges, C.: A NN/HMM hybrid for on-line handwriting recognition. Neural Comput. **7**(6), 1289–1303 (1995)
2. Ciresan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J.: Convolutional neural network committees for handwritten character classification. In: 11th International Conference on Document Analysis and Recognition (ICDAR), pp. 1250–1254 (2011)
3. Ciresan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
4. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
5. Deng, L.: The MNIST database of handwirtten digit images for machine learning research. IEEE Sig. Process. Mag. **29**(6), 141–142 (2012)
6. Graves, A., Fernandez, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: 23rd International Conference on Machine Learning (ICML), pp. 369–376 (2006)
7. Graves, A., Schmidhuber, J.: Offline handwriting recognition with multidimensional recurrent neural networks. In: 21th Advances in Neural Information Processing Systems (NIPS), pp. 545–552 (2008)
8. Grother, P.J.: NIST Special Database 19 - Handprinted Forms and Characters Database, 2nd edn. National Institute of Standards and Technology, Gaithersburg (2016). User's Guide

9. Guyon, I., Schomaker, L., Plamondon, R., Liberman, M., Janet, S.: Unipen project of on-line data exchange and recognizer benchmarks. In: 12th Internacional Conference on Pattern Recognition, vol. 2, pp. 29–33. IEEE (1994)
10. Kaltenmeier, A., Caesar, T., Gloger, J.M., Mandler, E.: Sophisticated topology of hidden Markov models for cursive script recognition. In: 2nd International Conference on Document Analysis and Recognition (ICDAR), pp. 139–142. IEEE (1993)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: 26th Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)
12. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
13. Nesterov, Y.: A method of solving a convex programming problem with convergence rate O(1/sqr(k)). Sov. Math. Dokl. **27**, 372–376 (1983)
14. Sayre, K.M.: Machine recognition of handwritten words: a project report. Pattern Recogn. **5**(3), 213–228 (1973)
15. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
16. Symonian, K., Zisserman, A.: Very deep convolutional networks for large scale image recognition. In: 3rd International Conference on Learning Representations (ICLR) (2015)
17. Van der Maaten, L.J.P. : A new benchmark dataset for handwritten character recognition. Tilburg University Technical Report, TiCC TR 2009-002 (2009)
18. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-Aware Neural Language Models, CoRR (2015)
19. Yuan, A., Bai, G., Jiao, L., Liu, Y.: Offline handwritten English character recognition based on convolutional neural network. In: International Workshop Document Analysis Systems, pp. 125–129 (2012)
20. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: 29th Advances in Neural Information Processing Systems (NIPS), vol. 28 (2015)

Cognitive Conversation Language - CCL

Wesley Willy Oliveira de Souza(✉) and Estevam Rafael Hruschka Júnior

Department of Computer Science, Federal University of Sao Carlos, Sao Carlos, Brazil
wesley_souza@dc.ufscar.br

Abstract. Enhancements in ways a conversational agent can participate in human-computer dialog tasks, as well as in the ability of such agents in answering questions has been explored in different research areas in the last years. But some limitations are still present in most of the current conversational agents' models. In this paper we present a new formal language, named Cognitive Conversation Language (CCL), designed to give support to conversational agents. CCL can be used to help such agents in a variety of tasks, including questions understanding and answer generation. Empirical results of the use of CCL are provided based on experiments using the knowledge base of a well-know never-ending learning system, called NELL (Never-Ending Language Learner). Results obtained in the experiments are empirical evidence that CCL can contribute in improving the quality of a question & answer task.

1 Introduction

Conversational agents [2,4] are computer programs designed to interact with humans through natural language. Such computer programs are usually used to assist, educate or entertain human users [4]. There are also some approaches where conversational agents are used to help computers to acquire knowledge from Humans. In such a scenario, a conversational agent automatically obtains human supervision, on specific topics, to help improving the assistant performance on specific tasks, such as in a *Conversing Learning* approach [15–17].

A current and relevant research topic is the search for improvement in the communication way and the ability to answer questions of these agents. Thus, different approaches were proposed in the literature in the last years, such the ones published in [7–9,13], focusing on allowing such agents to more naturally communicate, thus allowing them to go beyond a predefined framework with a fix (and predefined) set of questions and answers [5]. Most of the current successful approaches involve the integration of different reasearch areas, including Machine Learning [12], Natural Processing Language [10], Information Retrieval, Inference, Planning and Knowledge Representation [3,6,9,14,18].

One commonly described feature, in the literature [3,6,14], is the use of a conversational agent based on a specific knowledge base (KB) to find answers of questions asked by human users. The use of a KB may help in agent performance and, also in answers quality. Furthermore, in cases that the KB can

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_31

be continuously and automatically populated and extended (such as in NELL [1,11]), the agent can take advantage of the continuous addition of knowledge to try to answer questions that could not be answered in previous attempts.

Independently of the main goal of a conversational agent, one important issue when defining its architecture is related to the language model that the agent will be based on. As aforementioned, the agent can have a fixed set of predefined questions and answers to use during conversations, or it can use a more flexible language model that allows more intelligent behavior. In this paper we present a new language model designed to allow a conversational agent to have cognitive capabilities when keeping conversations with human users. Such new language model is named *Cognitive Conversational Language - CCL*. CCL is part of a bigger project of a question answers system using NELL's[1] KB for forums like *Yahoo Answers*, for example.

In the next sections we present a brief summarization on the state-of-the-art, followed by a formal definition of CCL, as well its basic structure, using examples. Results and results analysis of an application of CCL to a real scenario are also going to be presented.

2 Related Work

One of the most popular languages for development of conversational agents is *AIML* (Artificial Intelligence Markup Language) [19], which is known for being easy to use. AIML language has a limited KB, which is created by adding rules defining patterns to be used when answering questions. Thus, in an conversational agent based on AIML, the agent uses a KB only to answer previously programmed questions. Thus some research works such as [5] propose a way to automatically update the agent's KB through interaction with humans (using text).

Other approaches to develop conversational agents are also proposed in the literature, as in [3], where a graph-based system was developed to map and interpret questions sent by users and later, answer them. Approaches presented in [3,6,14] propose new techniques of interpretation and acquisition of knowledge through conversation agents.

3 The Language

In this section CCL is presented in two different ways, the first one is a formal definition of the language. CCL is formaly defined as a regular language, therefore its definition that can done in a easy way using regular expression.

The second view of CCL is given based on its writing structure. With such representation is possible to understand it in a convenient way, in other words, CLL's structure can reveal the function of each element, thus facilitating the language understanding and use.

[1] NELL: Never-Ending Language Learning, see more in: http://rtw.ml.cmu.edu/.

3.1 Formal Definition

Given the following alphabets:
$\sum_S = \{\backslash s, \backslash n, (null), (all)\}$
$\sum_{NDR} = \{(\$domain), (\$range)\}$
$\sum_{VDR} = \{(\$\$domain), (\$\$range)\}$
$\sum_{DR} = \{\sum_{NDR} \cup \sum_{VDR}\}$
$\sum_{VC} = \{(\$\$category)\}$
$\sum_T = \{(relation), (category)\}$
$\sum_{AS} = \{\sum_S \cup \sum_{DR} \cup \sum_{VC} \cup \sum_T\}$

$\Gamma_L = \{A, a, B, b, ..., Z, z\}$
$\Gamma_N = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$
$\Gamma_S = \{s | s \in$ set of all symbols recognized
by UTF-8 text format$\}$

$\sum_C = \{c | c \in$ set of categories names known by NELL$\}$
$\sum_R = \{r | r \in$ set of relations names known by NELL$\}$
we have:
$\sum_{CCL} = \{\sum_{AS} \cup \Gamma_L \cup \Gamma_N \cup \Gamma_S \cup \sum_C \cup \sum_R\}$

Given the following regular expression that uses $\sum_{CCL}$:
$r_1 = (\sum_C \cup \sum_R)[\backslash s] \sum_T [\backslash s] \sum_C [\backslash s](\sum_C \cup (null))[\backslash s]$
$r_2 = [*] \cup \sum_{DR}$
$r_3 = [\text{''}](\Gamma_L \cup \Gamma_N \cup \Gamma_S) * [\text{''}]$
$r_4 = (r_2 \cup r_3)([+]r_2 \cup r_3) * [\backslash s]$
$r_5 = [\backslash n] \cup \sum_{DR}$
$r_6 = [[](\sum_{VDR} \cup \sum_{VC})[;](\sum_{VDR} \cup [*])[;]((all) \cup (\Gamma_N + [\text{-}]\Gamma_N +) *))[]]$
$r_7 = (r_3 \cup r_5 \cup r_6)([+]r_3 \cup r_5 \cup r_6)*$

the regular expression that represents the CCL is:
$CCL(r) = (r_1 r_4 r_7)([\backslash n] r_1 r_4 r_7)*$

3.2 Structure of the Language

CCL's structure is easy to be programmed through a text file. In such an approach, each line of text document represents a rule and each rule has six fields to be sequentially filled, and separated by spaces. The six fields are:

Instances Set. It contains the prologue of instances set (i.e. the name that identifies) that are in NELL's KB[2] which is search guided.

Instances Set Type. Defines the type of instance set declared in first field, the acceptable values is *category* and *relation*.

[2] NELL' KB is freely available in the website of the Carnegie Mellon University project: http://rtw.ml.cmu.edu.

Domain Name. This field contains the prologue of the main category of a relation, in case the type of instances set is a relation. Otherwise the field needs to contain the same category prologue of the first field. For example, if the instance set is *actor_starred_in_movie* the domain name will be *actor*.

Range Name. This field contains the prologue of the second category of a relation, in case the type of instances set is a relation. Otherwise the field needs to contain the string *null*. For example, if the instance set is *actor_starred_in_movie* the range name will be *movie*.

An example of part of a CCL rule with the first four fields can be:

```
actor_sttared_in_movie relation actor movie
```

In the above example is showed an instance containing only with the already described four fields. When having such a rule the agent will be expected to search the KB to find answers in the relation *actor_starred_in_movie* and categories *actor* and *movie*. The next fields are necessary to complete this rule.

Question Pattern. This field is filled with a kind of mask, it will be compared with the user's question. If the pattern matches, the agent will use the next field (Answer Template) to send the answer. This field accepts six different types, namely: *Specific string*, *Undetermined string*, *Domain name*, *Domain value*, *Range name* and *Range value*. Where *Specific string* accepts any type of characters that are inside quotes. For example, if I want get only questions where the word *list* is present, I will use *"list"*.

The type *Undetermined string* is used to ignore a text interval of a undetermined size. In the same example of specific string, the user may use different words for the same question. The code below is a example of pattern to match with questions 1 and 2.

An example of part of a CCL question pattern using specific string and undetermined string is given below:

```
*+"list"+*
```

Examples of questions that match the above pattern:

- *Question 1:* Please, I need a list of movies.
- *Question 2:* Can you tell me a list of actors?

In *Question 1*, the both strings *Please, I need a* and *of movies.* will be ignored by the pattern verification, so the sentence will match with the question pattern described above. The same idea works on *Question 2*, in this case, the strings *Can you tell me a* and *actors?* will be ignored, but the string *list* will match the pattern.

Domain name, *Domain value*, *Range name*, and *Range value* are represented by reserved words. The *Domain name* is represented by string `$domain`, it guides

the conversational agent to check if there are the same characters inside the indicated position of the question pattern and in the domain name. In the example of *Question 1* and *Question 2*, if I want to get only questions about movies, I will use a question pattern like in the next example, considering that movies is the category instantiated in field *Domain Name*.

An example of a Question pattern with domain name, can be:

```
*+"list of "+$domain+*
```

As in *Domain name*, the *Range name* is represented by the reserved word, but in this case the reserved word is `$range` and represents the second argument of the relation used in the rule.

Domain value and *Range value* are represented by reserved words `$$domain` and `$$range` respectively. These reserved words represents instances names of an category instantiated in the rule. It will guide the agent to search strings in the NELL's knowledge base that match with the question. For example, if I want to get questions like: *"I need a list of movies starring Ben Stiller."* or *"I need a list of movies starring Al Pacino."*, I can use the question pattern in next example.

An example of Question pattern with range name and domain value

```
actor_sttared_in_movie relation actor movie
*+"list of "+$range+*+" starring "+$$domain+*
```

Answer Template. This is the last field of the rule and guides the conversational agent to answer the questions recognized by the *question pattern*. Its sub-fields are similar to the question pattern sub-field, but with some differences, its sub-fields are: *Specific string*, *Domain name*, *Domain value*, *Range name*, *Range value* and *Search Result*.

The *Specific string* in this case is used to guide the agent to write the exact text contained in the field. See an example of Answer Template with specific string:

```
"I hope you enjoy these movies:"
```

The sub-fields *Domain name* and *Range name* are used to write the exactly name of the categories in domain and range fields respectively.

Domain Value and *Range value* are used in the answer template to write the same value found in values of the question pattern or in the KB. Continuing the example of the question pattern code using specific string and undetermined string, the answer template can be assembled like in the following example of Answer Template with range name and domain value

```
"I think that this is a list of "+$range+ " starring "+$$domain
```

The last field accepted in the answer template is *Search Result*. This field is used to return a search result, this search is performed in NELL's knowledge base. Only results that satisfy the search condition are returned as part of the

answer. To perform the search, it is necessary to fill three parameters, they are: *search value*, *search condition* and *result interval*.

Search value is used to indicates which value will returned on knowledge base. This parameter accept `$$category`, `$$domain` and `$$range`.

Search condition is a parameter used to restrict the search results. The values accepted in this version are `$$domain` and `$$range`. It needs to be sent by the user and recognized in knowledge base to be validated.

Result interval is used to restrict the number of results that will be returned. For this, just insert the pattern *x-y* where *x* is an positive integer that indicates the begin of result interval and *y* is the end of interval. To return only the first interval just inset the value `1-1`, `1-5` for five first results and `all` for all results.

Giving these fields, is possible to understand the use the example showed below.

A complete Answer Template:

```
"I think that this is a list of "+$domain+" starred in "+$$range+":"
+\n+[$$domain;$$range;all]+\n+"I hope it helps you."
```

Where the field `[$$domain;$$range;all]` has as first parameter the value `$$domain`, indicating the agent to search by domain values in the knowledge base. The second parameter contains the value `$$range` that restricts the search in the knowledge base only in instances that have the range value given by the question. The last parameter has the value `all`, that indicates an result interval, in this case the agent will return all results founds in the knowledge base. In code below is showed an example of a complete rule

A complete Rule

```
actorstarrredinmovie relation actor movie
*+" list of "+$domain+*+" starred in "+*+$$range+*
"I think that is a list of "+$domain+" starred in "
+$$range+":"+\n+[$$domain;$$range;all]
+\n+"I hope it helps you."
```

With this rule the agent will recognize questions as: *"Please, I need a list of actors starred in Matrix"* and will perform a search in knowledge base about actors that starred in Matrix movie. Then, it will take the results and will elaborate the answer with the list of actors according the answer template.

4 Implementation

In this section is showed the results obtained in an implementation of an agent that uses the CCL to answer questions. The main aim of this implementation was to analyze whether CCL can help an agent to understand questions and whether the rules were described without ambiguity or interpretation failure. It was expected to have a chance to return wrong answers, because we used

only rules related to a specific scope. Another expected event is the agent not understanding part of the questions, by the same reason.

For the language validation a computational agent was developed in Java. It receives questions from Yahoo Answers![3] from actors and movies categories as input. The agent uses some CCL rules written in a text file to understand the questions and do searches in NELL's knowledge base when necessary.

The agent received as input: 122 questions; 1 CCL file with 16 rules containing all features presented in these release of CCL; All instances of "movie" and "actor" categories and of the "actor starred in movie" relation of the iteration 819 of NELL's knowledge base.

From those, 23 questions were answered by the agent. The rules file has been programmed to answer only questions consistent with the project scope, which is only to interpret the questions they would have to do some kind of consultation in NELL's knowledge base, although in CCL language, you can program the rules to answer all questions, including the incomprehensible.

5 Results and Analysis

The CCL-based agent has answered 23 questions. Through the results it was possible to identify the agent was able to recognize the questions, to search information on NELL's knowledge base and to elaborate successful answers. From the 23 answered questions, 9 were considered correct and 14 incorrect.

In seven of the fourteen incorrect answers the agent was able to interpret the questions correctly, was able to get the data from the knowledge base, but the information given in the answer was not enough to be considered a complete answer for the human user. The following is one of agent outputs along with the rule used.

```
Rule:
movie category movie null "what are "+*+\$domain+* "I hope you enjoy
  these movies:"+ n+[\$\$category;*;1-10]

Question:What are some good horror movies ?

Answer:I hope you enjoy these movies: aliens, an american in paris,
     beowulf, big lebowski, boogie nights, close encounters, collateral,
     dick tracy, die hard, doctor zhivago.
```

In 3 cases the lack of specification of information was given by the rules, see one of the outputs.

```
Rule:
actor category actor null *+"list of "+*+$domain "I hope this is the list
     of actors you're looking for:"+\n+[$$domain;*;1-10]

Question: How can a find a list of companies looking for actors?
```

[3] http://answers.yahoo.com.

```
Answer:I hope this is the list of actors you're looking for: al pacino,
    arnold schwarzeneggar, bruce willis, cary grant, emma thompson,  fay
    wray, frank sinatra, gary cooper, gary lockwood.
```

One answer was considered incorrect because of the two aforementioned problems. In those cases, both the lack of specification of information in the knowledge base as well as the generalization of the rule affected the agent's precision in answering.

```
Rule:
actor category actor null "who are"+*+$domain+* "what do you
     think of these:"+/n+[$$category;*;1-10]

Question: Who are a and b list actors?

Answer: What do you think of these: alan north, alistair findlay,
     al pacino, amy poehler, angie harmon, anika noni rose, anne
     bancroft, anthony mannino, arnold schwarzeneggar.
```

In one of the incorrect answers the problem was the agent that has not been programmed to handle missing data in the knowledge base. In such cases, the agent should not develop an answer even managing to interpret the question. Thus the agent returned the following result.

```
Rule:
actorstarredinmovie relation actor movie "which "+*+$domain+*+"starred in
     "+*+$$range+* "I think that "+[$$domain;$$range;1-1]+" starred in "+
     $$range+"."

Question: Which actor starred in the lethal weapon series ?

Answer: I think that  starred in lethal weapon.
```

The last two cases in which the answers were considered incorrect were a mixture of generalized rules, failure agent task in the missing information and the question asked. In such cases, users wrote a theme that should be a question, then the agent interpreted according to the rules and searched information in the knowledge base. As the problem was presented, there were all the information in the knowledge base and the agent returned an incomplete answer.

```
Rule:
actorstarredinmovie relation actor movie *+" cast "+*+$$range+* "I
      think this is the cast of "+$$range+":"+\n+[$$domain;$$range;
     all]

Question: STAR WARS OR STAR TREK?!?

Answer: I think this is the cast of star trek :
```

It is important to note that none of the issues was caused by problems in CCL, and that all presented features were used by the agent without errors.

6 Conclusion

We presented in this paper the formal model of the Cognitive Conversation Language (CCL) which aims to be used for conversational systems, not only in chats but also in forums. CCL, differently from other languages, such as AIML, allows to direct access a Knowledge Base (in this paper we've used NELL' KB) through reserved words that indicate where to find the information needed to answer questions.

Empirical evaluations were performed with only a few types of questions related to movies and actors. Results indicated empirical evidences that CCL helped in interpreting questions and to quickly access the knowledge base. However, this initial version of CCL does not allow access to more details of the knowledge base, as the prologues of the categories, information passed by specific components of NELL, such as SEAL, CPL and other relationships in which a category may be involved, among others. Such details would allow a chat system to answer other questions, enhancing the experience that the user may have with the system.

According to the results, there are three major problems. The first one was the generalization of some rules that affected at least seven answers. This can be solved by creating more specific rules for each type of question. Another issue that influenced in 4 of the wrong answers was the agent development, this problem will be solved in the next release. We can also research for specific information in the context of each instance found in NELL knowledge base, as well as suggest subcategories like "black_actor" for example.

CCL will be part of a larger project where this is only one module of a complete conversation system. Its structure allows the manipulation of rules by both humans and machines, so will be possible that a chat system (based on machine learning) create new rules, edit and delete previously created rules, towards improving its conversational ability. Before this conversational system is finalized, CCL will undergo through improvements in future work. One point of improvement of CCL in future work is the addition of new reserved words to cover the maximum access to the information in NELL's knowledge base. Thus enabling conversation systems to elaborate more complete answers to more variations of questions.

The advantage of having a language like CCL in a conversational system is mainly the possibility of avoiding the need to make interpretations during the dialog with the system. The idea of using CCL is to facilitate the conversation systems as well as other languages (such as AIML does), but with the advantage of allowing more information and flexibility than in other available approaches.

References

1. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr., E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: AAAI. vol. 5, p. 3 (2010)
2. Cassell, J.: Embodied Conversational Agents. MIT Press, Cambridge (2000)

3. Dali, L., Rusu, D., Fortuna, B., Mladenic, D., Grobelnik, M.: Question answering based on semantic graphs. In: Proceedings of the Workshop on Semantic Search (2009)
4. David., S.O.: Modern conversational agents. In: Wiesbaden, S.F. (ed.) Technologien für digitale Innovationen, pp. 63–84 (2014)
5. De Pietro, O., Frontera, G.: Tutorbot: an application aiml-based for web-learning. Adv. Technol. Learn. **2**(1), 29–34 (2005)
6. Fader, A., Zettlemoyer, L., Etzioni, O.: Open question answering over curated and extracted knowledge bases. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1156–1165. ACM (2014)
7. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., et al.: Building watson: an overview of the deepqa project. AI Mag. **31**(3), 59–79 (2010)
8. Johnston, M., Bangalore, S., Vasireddy, G., Stent, A., Ehlen, P., Walker, M., Whittaker, S., Maloor, P.: Match: an architecture for multimodal dialogue systems. In: Proceedings of the 40th ACL Annual Meeting, pp. 376–383. ACL (2002)
9. Lally, A., Prager, J.M., McCord, M.C., Boguraev, B.K., Patwardhan, S., Fan, J., Fodor, P., Chu-Carroll, J.: Question analysis: how watson reads a clue. IBM J. Res. Dev. **56**(34), 2:1–2:14 (2012)
10. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing, vol. 999. MIT Press, Cambridge (1999)
11. Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., Krishnamurthy, J., et al.: Never-ending learning. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pp. 2302–2310. AAAI Press (2015)
12. Mitchell, T.: Machine Learning. McGraw-Hill, New York (1997)
13. Moschitti, A., Quarteroni, S., Basili, R., Manandhar, S.: Exploiting syntactic and shallow semantic kernels for question answer classification. In: Annual Meeting-Association for Computational Linguistics, vol. 45, p. 776 (2007)
14. Pappu, A.K., Rudnicky, A.I.: Knowledge acquisition strategies for goal-oriented dialog systems. In: Proceedings of the SIGDIAL Conference (2014)
15. Pedro, S.D., Appel, A.P., Hruschka Jr., E.R.: Autonomously reviewing and validating the knowledge base of a never-ending learning system. In: Procedings of the 22nd International Conf on World Wide Web, pp. 1195–1204. ACM (2013)
16. Pedro, S.D., Hruschka Jr., E.R.: Collective intelligence as a source for machine learning self-supervision. In: Proceedings of the 4th International Workshop on Web Intelligence & Communities, p. 5. ACM (2012)
17. Pedro, S.D.S., Hruschka, E.R.: Conversing learning: active learning and active social interaction for human supervision in never-ending learning systems. In: Pavón, J., Duque-Méndez, N.D., Fuentes-Fernández, R. (eds.) IBERAMIA 2012. LNCS (LNAI), vol. 7637, pp. 231–240. Springer, Heidelberg (2012). doi:10.1007/978-3-642-34654-5_24
18. Ronald, B., Levesque, H.: Knowledge representation and reasoning. Elsevier, New York (2004)
19. Wallace, R.: The Elements of Aiml Style. Alice AI Foundation (2003)

A Genetic-Fuzzy Classification Approach to Improve High-Dimensional Intrusion Detection System

Imen Gaied(✉), Farah Jemili, and Ouajdi Korbaa

MARS Laboratory, ISITCom, University of Sousse, Sousse, Tunisia
gaiedimen@gmail.com, Jmili_farah@yahoo.fr,
Ouajdi.Korbaa@centraliens-lile.org

Abstract. With the increasing number of attacks and growing scalability of connected networks over the past few years, researchers are brought to find other alternatives to judge the relevance, severity and correlation of network attacks. The high-dimensional intrusion detection system seems a promising dynamic protection component in security fields. In this work we propose an optimized classification scheme that coordinates several techniques for generating fuzzy association rules based on a large data set. Our main task is to ameliorate the detection rate of attacks in a real-time environment by using the one-versus-one decomposition to minimize as much as possible the false alarm rate. In addition, we aim to reduce the loss of knowledge through a suitable n-dimensional overlap function in order to model the conjunction in fuzzy rules to provide enough classification accuracy. We can also opt for the aggregation method to obtain the final decision. To evaluate the performance of our approach, an experimental study is performed so as to achieve relevant results. The final outcome shows that our approach outperforms other classifiers by providing the highest detection accuracy, a low false alarm rate and time consumption.

Keywords: Intrusion detection system · OVO decomposition · n-dimensional overlap function · Fuzzy rules associations · Detection rate · False positives

1 Introduction

There has been recently an exponential growth of network attacks and an increase in their severity. Therefore, developing an efficient Intrusion Detection System (IDS) becomes an important open problem that is receiving a noteworthy attention from the research community. The main focus of this research work is to put forward an intelligent and accurate intrusion detection using linguistic Fuzzy-Rule-Based Classification Systems (FRBCS) in synergy between the preprocessing techniques, the pairwise learning and the n-dimensional overlap functions. This approach aims to explore the large search space, achieve a more efficient result accuracy, maintain a high confidence and a good coverage of the proposed database and to provide the user with high-quality rules. To our knowledge, this is the first research paper which suggests this approach for intrusion detection problems. To reach these goals, we have to deal at first with the

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_32

imbalanced distribution of classes in high-dimensional problems. In fact, the classification accuracy of a classifier is directly affected by the quality of the training data used to construct the final model. Basically, we apply the preprocessing of datasets to remove or correct the noisy pattern. Then, we use the One-Versus-One (OVO) decomposition to decrease the complexity and the noise effect of an original problem [1]. To better decrease the effects caused by noise, we consider the fuzzy association rule learning algorithm known as FARC-HD (Fuzzy Association Rule-based Classification model for High-Dimensional problems) [2] in order to obtain the most accurate and highest quality rules. Nevertheless, the usage of the product as a T-norm in a baseline FARC-HD algorithm with an OVO decomposition results in low-variation values, especially when we use a higher number of antecedents of the fuzzy rules. These values are utilized in the aggregation step to get the last decision. Therefore, some robust aggregation methods are affected by this undesirable condition and so can lead to a poorer classification than the original FARC-HD. To address this problem, we have to substitute the product T-norm by a suitable overlap function [3, 4]. The latter prior works defined n-dimensional overlap functions in order to obtain suitable outputs, which were in a wider range from the base classifiers. Hence, more information was maintained by a further aggregation process that could make a better improvement of the classification in the OVO scheme. To evaluate the performance of this new methodology in the intrusion detection problem, a comparison study was performed with another approach based on the original FARC-HD and OVO strategy [5]. Accordingly, we show the validity of such suggestions in ameliorating the accurate detection intrusion compared to other models. The remainder of this paper is organized as follows. Section 2 first describes the operating principle of FARC-HD algorithm and then the notion of classification via the pairwise techniques. After that, it specifies the binary tree of the SVM as an aggregation method, and finally draws a brief definition of the n-dimensional overlap functions. The proposed approach is presented in Sect. 3. The obtained experiments and results are discussed in Sect. 4. The conclusion and future work are given in Sect. 5.

2 Proposed Methodology

2.1 FARC-HD Algorithm

We use in this work the FARC-HD algorithm [2] in order to provide the high quality rules in a high dimensional problem. The model of the classification problem consists of m training examples with $x_p = (x_{p1}, \ldots, x_{pn}, C_p)$, $p = 1, 2, \ldots, m$ from m classes where x_{pi} is the ith attribute value ($i = 1, 2, \ldots, n$). The fuzzy rule is depicted as following:

$$\text{Rule } R_j : \text{If } x_1 \text{ is } A_{ji} \text{ and } x_n \text{ is } A_{jn} \text{ then Class} = C_j \text{ with } RW_j \tag{1}$$

Where R_j is the label of the jth rule, $x = (x_1, \ldots, x_n)$ is an n-dimensional pattern vector, A_{ji} is an antecedent fuzzy set, C_j is the class label and R_{wj} is the rule weight. The learning process consists of three main steps which are: Fuzzy association rule mining for classification, Candidate rule prescreening, Genetic rule selection and lateral tuning.

A search tree is built for each class to get the fuzzy rule base. We can reduce the search space by generating only the rules with a high support and a high confidence. To preselect the most strong rules, a rule assessment criterion is used by utilizing the "subgroup discovery" mechanism [8]. In order to decrease the computational costs, an evolutionary algorithm is carried out to obtain a lateral tuning of the fuzzy sets [9] so as to pick the final best rules from the rule base.

To classify a new example, the FARC-HD applies a fuzzy reasoning method, called additive combination, composed of four steps: Matching degree, Association degree, Confidence degree and Classification. In the first step the strength of activation of the if-part for all rules in the Rule Base (RB) with the pattern x_p is computed by means of using a conjunction operator (T-norm).

$$\mu_{A_j}(x_p) = T\left(\mu_{A_j}(x_{p1}), \ldots, \mu_{A_j}(x_{pn})\right), \mathrm{j} = 1, \ldots, \mathrm{L} \tag{2}$$

The association degree of the pattern x_p with each rule in the RB which has as a consequent class the rule R (k refers to class (R_i)), is computed by using a combination operator h to combine the matching degree with the Rule Weight (RW):

$$b_j^k = h\left(\mu_{A_j}(x_p), RW_j^k\right), \; K = 1 \ldots M; j = 1 \ldots L \tag{3}$$

In the third stage we use an aggregation function f (the sum in the case of FARC-HD) whose positive association degrees computed in the previous stage are combined.

$$y_k = f\left(b_j^k, \mathrm{j} = 1 \ldots \mathrm{L}\right), \mathrm{K} = 1 \ldots \mathrm{M}, \mathrm{b_j^k} > 0 \tag{4}$$

In the last step we apply a decision function F to determine the class that obtain the highest confidence degree is predicted.

$$F(y_1 \ldots y_M) = argmax(y_k), [K = 1 \ldots M] \tag{5}$$

2.2 Classification by Using Decomposition Strategies: OVO

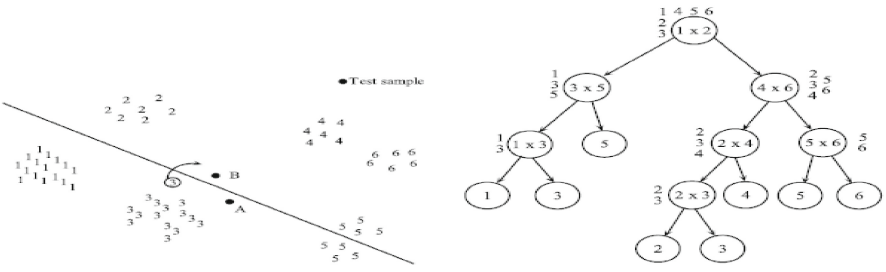
In the OVO, the original multi-class is divided into $\frac{m(m-1)}{2}$ binary sub-problems, which aim to distinguish a pair of classes C_i, C_j. When a new pattern is presented to each binary classifier, a pair of confidence degrees r_{ij}, r_{ji} is given in favor of two classes C_i, C_j (the class with the largest confidence is the output class of this classifier). All the outputs (confidence degrees) provided by all binary classifiers represented by a score matrix R are combined to make the final class prediction using aggregation models.

$$R = \begin{bmatrix} - & r_{12} & \cdots & r_{1m} \\ r_{21} & - & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & - \end{bmatrix} \tag{6}$$

In the literature [6] we distinguish different methods, which are the Max-Wins rule known as the Voting strategy (VOTE), the Winner Weighted Voting (WINW) whose validity was proven in [4], the methods based on preference relations (ND, LVPC). Please refer to [4, 6, 10] for more information. Otherwise, in this work we adopt an architecture of the Binary Tree of the SVM (BTS) [10] for the final decision.

The BTS is an architecture of tree structure which can be extend easily to any type of binary classifier aiming to construct a recursively binary tree. At the beginning, its root node considers all classes in the list of classes and a binary classifier is selected randomly for training in order to get a separating plane. Then, all the samples in the node are assigned to two subnodes derived from the two classes. To complete the binary tree structure, we have to apply this classification strategy in every node until a leaf node containing only one class is reached. The decision is made when using the binary classifier to discriminate between two classes in order to distinguish the remaining other classes simultaneously. Thereby, more than one class can be removed from the list. In addition, to avoid any false assumption, especially when the leaf contains more than one class, the output is computed with the voting strategy.

Figure 1 illustrates this concept applied on a six-class problem. The first node classifier discriminates classes 1 and 2. On the one hand, when class 1 is predicted, classes 4 and 6 are removed (when a testing sample is at position A, it does likely appertain to class 4 or 6). On the other hand, when class 2 is predicted, only class 1 is taken out. Therefore, classes 3 and 5 are maintained in the two next nodes by using probabilistic outputs, because class 3 is near the decision function (when a testing sample is at position B, it will have a chance to be classified to class 3) and because class 5 cannot be distinguished with the classifier in the root node.



(a) Multiclass problem (b) Architecture of BTS

Fig. 1. Six-class problem determined by BTS [10]

2.3 Classification by Using n-Dimensional Overlap Functions

As we mentioned above, the confidence degrees provided by the original FARC-HD have low variations, which is not advantageous for the aggregation process performed in the OVO. This negative effect is explained by the use of the product T-norm to model the conjunction. To provide a better synergy between the FARC-HD and the OVO decomposition, a solution is proposed consisting in substituting the product in an association degree by an n-dimensional overlap function so as to obtain results in a wider range and solve the problem of penalization of the rules that have a great number of antecedents. In such a way, the outputs of the base classifiers become more suitable for the subsequent aggregation step. In our case we use a MIN operation as an overlap function that satisfies the property of discrimination capability of the FARC-HD and preserves the idempotence criteria.

$$O^n\left(x_1 \ldots x_n\right) = \text{Min}\left(x_1 \ldots x_n\right) \tag{7}$$

3 Experimental Framework

In this section, we first describe the datasets picked for the experimental study (Sect. 3.1). Next we provide details of the base classifiers used in the study by describing their configuration parameters and aggregation strategies in Sect. 3.2. Finally, we present the measures employed to evaluate the performance of each classifier (Sect. 3.3).

3.1 Datasets: KDD CUP 99

The experiments were implemented step by step by utilizing the KDD99 benchmark dataset which was used by many researchers to evaluate their IDS [5, 7, 12]. It has been the most widespread and complete dataset opted for data mining applications. It is divided into data labelled for a training phase (about 5,000,000 records) and data unlabelled for the test base named "corrected KDD" (about 311,000, it includes 14 new types of attacks). Each record contains TCP/IP connections composed of 41 attributes and comprising four main classes of attacks, namely: Probe, R2L, U2R and DOS. Unfortunately, this dataset might have been some disadvantages: the redundant nature of alerts and a broad base containing 41 attributes with an imbalanced distribution. Those dilemma can affect the detection of rare attacks rather than dominant ones. To resolve this problem, some researchers have used different sizes of data sets prepared by a random selection [5, 7].

3.2 Configuration Parameters Used for Study

In this section, we briefly present the FARC-HD algorithm used as a base classifier and different combination methods for the OVO decomposition scheme. We chose 5 labels per attributes for the fuzzy sets which are in the form of triangular MFs. We considered five aggregation methods in addition to our proposed BTS method, which are the VOTE,

the Non-Dominance Criteria (ND), the Learning Valued Preference for Classification (LVPC), the WinWV. As a conjunction operator, we selected the Product (PROD), the Geometric Mean (GM), the Harmonic Mean (HM) and the Minimum operator (MIN). As an inference procedure we set up the additive combination. Furthermore, we have fixed the minimum support to 0.05, the minimum confidence to 0.8, the maximum depth level of the tree to 3, the k parameter for the pre-screening to 2, the maximum evaluation to 4,327 and the population size to 50 [2].

3.3 Performance Measure of Intrusion Detection

Unfortunately, it is not easy to design a satisfactory evaluation strategy to get a valuable conclusion. Indeed, the methods used previously show three major drawbacks: the use of a non-representative base of the test data, the lack of a rigorous evaluation methodology, and the utilization of improper metrics. We considered the following measures as an evaluation parameter for these reasons:

Classification Rate (CR): It is used as a classical metric, named also "overall accuracy". It defines the fraction of instances predicted correctly (TP+TN) by the total number of instances (n).
Cohen's kappa (Kappa) [11]: It is considered as an alternative measure to the CR. Since it scores the success instances independently for each class and aggregates them, compared to the CR that aims to score all the successes over all the classes. Kappa scoring is very convenient for measuring the classification's rate while it is less sensitive to the randomness caused by a different number of instances in each class.
Mean F-Score: It denotes the average of the F-Score of each class. It represents the trade-off between precision and recall. These measures are commonly used to evaluate the rare class prediction.
Detection Rate (DR): It is an important measure for the intrusion detection. It is computed as a fraction of the number of correctly detected attacks (TP) by the total number of attacks (TP+FN).
False Positive Rate (FPR): It is computed as a ratio between the number of normal examples detected as attacks (FP) and the total number of normal examples (FP+TN).
Receiver Operating Characteristic (ROC): It is used for comparing the detectors together. The ROC curve can be generated by plotting TP (sensitivity) against FP(1-specificty) for each utilized threshold or decision cutoff. A data point in the upper left corner corresponds to an optimal high DR with low FPR.
Area under ROC Curve (AUC): It is a performance measure used to compare the IDS, with an area of (0.1) representing a perfect test as it classifies all positive and negative cases correctly.

4 Experimental Study

In this section, we discuss some conducted experimentations to show the performance of our suggested approach on large datasets. All the experimentations were implemented

in Java language on an Intel Pentium IV personal computer with 2.40 GHz and 7.88GO RAM.

4.1 Datasets and Pre-processing

We proposed to apply three main preprocessing phases [7]. We removed all the repeated records in the entire KDD training and test set and retained only one copy of each record. In order to reduce the complexity of the large amount of a dataset, we processed the feature selection by applying the factorial multiple correspondence analysis [12] to extract the most relevant features of data. We used just 9 relevant attributes. To become adaptable to the fuzzy logic, a normalization step is necessary. It is determined by the Min-Max function.

4.2 Classifier Structure

We want to study here how the different overlap functions and the different aggregation strategy affect the rule base size on the one hand and the computation time of learning algorithm on the other hand. To do so, a 5-fold cross-validation model was considered. We used the whole of 10% KDD datasets and we split 90% random instances for training and 10% for testing. We considered the average number of rules and antecedents by rule for each overlap function. The results are presented in the Table 1. Table 2 shows the computation time of each proposition method. Here, we require developing an optimized IDS with low cost-rules, which is the case of the MIN conjunction function and the BTS aggregation method.

Table 1. Number of rules and antecedents by rule for each conjunction function

Conjunction function	Number of rules	Number of antecedents by rule
PROD	78	2.79
GM	99	1.93
HM	84	1.88
MIN	79	1.88

Table 2. Computation time for each conjunction function

Conjunction function	Computation time (min)				
	VOTE	ND	LVPC	WinW	BTS
PROD	113.34	95.38	84.45	96.70	85.93
GM	144.85	125.9	126.77	121.22	123.78
HM	118.47	146.55	89.77	93.69	92.11
MIN	112.01	106.36	90.62	133.47	90.10

4.3 Intrusion Detection Performance Comparison

This section aims to analyze which method is considered the most robust strategy, accounting for the IDS problems. An experimental study was achieved in three steps. First, we studied the affect of our proposed overlap function and aggregation method, which are considered the most appropriate to improve the final performance of our suggested IDS. Second we emphasize the effectiveness of our accurate IDS in contrast with the other approaches. Finally, we adopted the ROC graphs and the AUC measure, which are considered a convenient way in comparing detectors together. Table 3 shows the results for the test partition (10% of the datasets) of the baseline FARC-HD algorithm within the OVO strategies (FARCHD-OVO) for each aggregation method. Table 4, Tables 5 and 6 represent the results of the FARCHD-OVO within the GM overlap function, HM overlap function and MIN overlap function respectively. It is important to notice that the function MIN reaches a good trade-off between DR and FPR with 95.80% as DR and 1.41% as FPR compared to other overlap functions: the function GM and HM with 94.25% as DR and 0.67% as FPR; with 94.95% as DR and 0.65% as FPR respectively.

Table 3. Test evaluation of FARCHD-OVO-PROD approach

Approach	CR	Kappa	MFM	DR	FPR
FARCHD-ND-PROD [5]	97	93.99	64.75	95.40	1.01
FARCHD-VOTE- PROD	96.97	93.93	70.77	95.43	1.4
FARCHD-LPVC- PROD	93.89	87.72	44.61	93.39	5.48
FARCHD-WINW-PROD	91.59	83.68	52.34	89.76	4.41
FARCHD-BTS- PROD	96.94	93.82	64.95	95.6	1.41

Table 4. Test evaluation of FARCHD-OVO-GM approach

Approach	CR	Kappa	MFM	DR	FPR
FARCHD-ND-GM	96.39	92.79	69.10	93.98	0.66
FARCHD-VOTE-GM	96.62	93.24	60.72	94.09	0.68
FARCHD-LPVC-GM	90.30	80.49	49.82	93.38	13.48
FARCHD-WINW-GM	93.62	87.19	56.49	92.53	5.14
FARCHD-BTS-GM	96.54	93.08	63.48	94.25	0.67

Table 5. Test evaluation of FARCHD-OVO-HM approach

Approach	CR	Kappa	MFM	DR	FPR
FARCHD-ND-HM	96.79	93.55	64.5	94.87	0.84
FARCHD-VOTE-HM	96.99	93.96	64.67	95.01	0.86
FARCHD-LPVC-HM	86.49	75.72	42.45	87.65	13.12
FARCHD-WINW-HM	92.22	84.98	52.33	91.07	6.36
FARCHD-BTS-HM	96.93	93.85	64.15	94.95	0.65

Table 6. Test evaluation of FARCHD-OVO-MIN approach

Approach	CR	Kappa	MFM	DR	FPR
FARCHD-ND- MIN	96.97	93.92	64.60	95.64	1.40
FARCHD-VOTE- MIN	97.07	94.03	66.34	95.79	1.37
FARCHD-LPVC- MIN	93.86	87.65	49.89	94.07	6.39
FARCHD-WINW- MIN	87.77	80.11	52.87	81.73	4.54
FARCHD-BTS- MIN	97.05	94.08	68.24	95.80	1.41

When facing the results of our suggested FARCHD-BTS-MIN scheme, versus the other schemes, we see that there is a significant improvement in all instances, in particular the classification rate, the cohen's kappa and the detection rate. It be must also pointed out that our approach is considered a specific solution to correctly detect the boundaries for all classes, because the average of the F-Score measure, regards to the FARCHD-VOTE-PROD method, achieves a high score, compared with other methods. These schemes prove their performance, to detect a rare attack with other categories of attacks. Additionally, it provides the most optimal trade-off of two important metrics: the detection rate and the false alarm rate with 95.80% and 1.41% respectively. It's important to stress that a little existing method of comparison in the quoted reference addressed this policy in intrusion detection. Among the studies, we mentioned this work [5] which opted for FARCHD-ND-PROD scheme, it aimed to prove its effectiveness to detect a rare attacks with other classes of attacks. Our method outperforms the latest method obviously. It has proven its effectiveness to detect a rare attack with 68.24% compared to FARCHD-ND-PROD method with 64.75%. Figure 2 illustrates the performance of our method using the ROC graphs and the AUC measure, compared to other various detection method used MIN operator, to complete our experimental study. For calculating the multi-class AUC "Total AUC", we have calculated the average AUC of each class. Considering the ROC curve of the FARCHD-BTS-MIN classifier, we notice that this curve always outperforms the other classifiers with AUC = 0.9981. The rest of the measures are mentioned in Table 7.

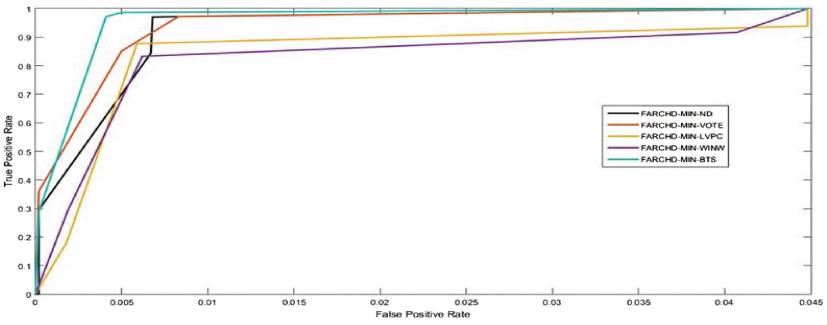


Fig. 2. ROC curve

Table 7. AUC value for each aggregation method using MIN function overlap

Aggregation method	ND	VOTE	LVPC	WINW	BTS
AUC	0.9964	0.9971	0.9928	0.9920	0.9981

5 Conclusion

In this work we put forward an intelligent misuse intrusion detection based on interpretable model by means of using some linguistic rules of the FARC-HD algorithm, the pairwise learning that is precisely founded on an architecture known as the BTS method, and the MIN operator as a conjunction function. Our new policies is characterized by a good interpretation, a high attack detection with a low false alarm rate and a high computational speed. This conclusion is supported by the experiments that have shown the effectiveness of our approach. Nevertheless, detecting rare attacks is still in need of improvement.

References

1. Saez, J.A., Galar, M., Luengo, J., Herrera, F.: Analyzing the presence of noise in multi-class problems: alleviating its influence with the one-vs-one decomposition. Knowl. Inf. Syst. **38**(1), 179–206 (2014)
2. Alcala-Fdez, J., Alcala, R., Herrera, F.: A fuzzy association rule based classification model for high-dimensional problems with genetic rule selection and lateral tuning. IEEE Trans. Fuzzy Syst. **19**(5), 857–872 (2011)
3. Elkano, M., Galar, M., Sanz, J., Bustince, H.: Fuzzy rule-based classification systems for multi-class problems using binary decomposition strategies: on the influence of n-dimensional overlap functions in the fuzzy reasoning method. Inf. Sci. **332**, 94–114 (2016)
4. Elkano, M., Galar, M., Sanz, J.A., Fernandez, A., Barrenechea, E., Herrera, F., Bustince, H.: Enhancing multiclass classification in FARC-HD fuzzy classifier: On the synergy between-dimensional overlap functions and decomposition strategies. IEEE Trans. Fuzzy Syst. **23**(5), 1562–1580 (2015)
5. Elhag, S., Fernandez, A., Bawakid, A., Alshomrani, S., Herrera, F.: On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on intrusion detection systems. Expert Syst. Appl. **42**(1), 193–202 (2015)
6. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. Pattern Recognit. **44**(8), 1761–1776 (2011)
7. Gaied, I., Jemili, F., Korbaa, O.: Intrusion detection based on neuro-fuzzy classification. In: International Conference on Computer Systems and Applications (AICCSA 2015), vol. 5, pp. 1–8, November (2015)
8. Kavsek, B., Lavrac, N.: APRIORI-SD: adapting association rule learning to subgroup discovery. Appl. Artif. Intell. **20**(7), 543–583 (2006)
9. Alcala, R., Herrera, F.: A proposal for the genetic lateral tuning of linguistic fuzzy systems and its interaction with rule selection. IEEE Trans. Fuzzy Syst. **15**(4), 616–635 (2007)

10. Fei, B., Liu, J.: Binary tree of SVM: a new fast multiclass training and classification algorithm. IEEE Trans. Neural Netw. **17**(3), 696–704 (2006)
11. Kohen, J.: A coefficient of agreement for nominal scale. Educ. Psychol. Measur. **20**, 37–46 (1960)
12. Jemili, F., Zaghdoud, M., Ben Ahmed, M.: Intrusion detection based on hybrid propagation in bayesian networks. In: Proceedings of the IEEE International Conference on Intelligence and Security Informatics, pp. 137–142, Dallas (2009)

A New Data Placement Approach for Scientific Workflows in Cloud Computing Environments

Hamdi Kchaou(✉), Zied Kechaou, and Adel M. Alimi

REsearch Groups in Intelligent Machines (REGIM),
National School of Engineers (ENIS), The University of Sfax,
BP 1173, 3038 Sfax, Tunisia
{hamdi.kchaou, zied.kechaou, adel.alimi}@ieee.org

Abstract. The reach of Cloud Computing technologies approved distributing with massive data applications such as Scientific Workflows, which processing huge scientific data in dispersed computing infrastructures. Among the characteristics of Cloud Computing, we mention the elasticity that allows workflows to dynamically stipulate necessary resources for tasks execution. The processing of massive data with scientific workflows increase the data transmission, rise execution delay and it request huge bandwidth cost. So, to reduce the execution cost of workflows and the data movements, data placement optimization technics must be taken into consideration. While placing datasets during execution of tasks for a job in a workflow, there are dependencies between datasets and between tasks. In this paper, we propose a data placement approach based on heuristic genetic algorithm which takes into accounts control and data flow dependency, in order to reduce data movements and so the utilization of resources in cloud environments.

Keywords: Cloud computing · Massive data · Scientific workflow

1 Introduction

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [1]. The cloud model comprises five essential features (On-demand self-service, Broad network access, Resource pooling, Rapid elasticity, Measured service, Resource usage), three service models (SaaS: Software as a Service, IaaS: Infrastructure as a Service and PaaS: Platform as a Service), and four deployment models (private cloud, community cloud, public cloud and hybrid cloud). The SaaS includes software applications, the PaaS contains platform applications and the IaaS take account of hardware resources as CPU, Memory, Disk and Bandwidth. Cloud computing and storage solutions provide users and enterprises with various capabilities to store and process their data in third-party data centers [2]. The cloud computing have the ability to exploit large amounts of computing and storage resources used to execute complex applications such as scientific workflow.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_33

A workflow is the automation of a process, during which data is processed by different logical data processing activities according to a set of rules [3]. Workflows are usually modeled as directed acyclic graphs (DAGs) such that workflow tasks are represented as graph vertices and the data flows among tasks are represented by graph edges. The direction of edges shows data flows among tasks [4]. The data required by scientific workflows are frequently big and distributed, that is why data placement becomes difficult. The adequate placement of data in data centers can reduce the data movements and equilibrium the load of data centers. It is ineffective to place all data sets in single data center and in order to achieve computing tasks; entirely requisite data sets must be placed in the same data center. So due to bandwidth limitations, we need to place datasets in one data center in the largest amount possible, to reduce data movements between storage nodes and compute nodes. Therefore, we can say as it is represented in Fig. 1, data placement is to place datasets in data centers in cloud computing environments. Each dataset has a size and each data center has a storage space. The data placement is a NP-hard problem. In order to solve this type of problems and as mentioned in [5–7], genetic algorithm have proven a better performances as an optimization technics for data placement.

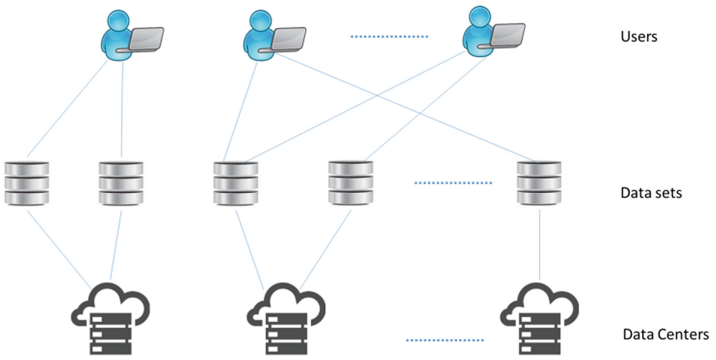


Fig. 1. Data placement on data centers under user's demands

To deal with massive data, MapReduce is one of the most way. Announced in 2004, MapReduce is a framework for large scale data processing using commodity clusters [8]. In the MapReduce programing model, a job is divided into two continuous phases which are named as the map phase and reduce phase, then the latter does not start until the former is complete [9]. The map and reduce tasks will be distributed to nodes and executed in parallel.

In this paper, we presented a novel approach for placement data with scientific workflows in cloud computing environments, which includes:

- Genetic algorithm for data placement.
- Dependencies between tasks of a MapReduce job and which illustrate the control flow in scientific workflow.
- Dependencies between datasets consumed and produced by MapReduce jobs.

The remainder of this paper is organized as follows. Section 2 illustrates the problem, describes the proposed approach and suggest a formulation of problem. Section 3 discusses the related work. Section 4 contains experimental simulation results. Section 5 summarizes our conclusions and points out the future work.

2 Proposed Approach

2.1 Description

A workflow is mainly imperative for applications in which data dependencies occur among the tasks. The dependencies represent the flow of data between workflow activities from data producer to data consumer. In control-driven workflows, or control flows, the connections between the activities in a workflow represent a transfer of control from the preceding task to the one that follows [10]. Control dependencies indicate a constraint on the relative order of execution of two tasks, without implying data transfer. They are only necessary for tasks with side effects, where data dependencies alone are not sufficient to determine the ordering [11]. Our proposed approach combines control flow and data flow, which called hybrid according to [10]. In this workflow, the tasks will be a MapReduce Jobs. Since MapReduce is data-driven, the hybrid workflow will be inclined on the road to data flow. In the other side, a job can be split into many map and reduce tasks, which are dependent to each other's. The map task is an input for the reduce task which can be an input for another map task and so on. This illustrates the control dependencies between tasks.

2.2 Problem Formulation and Proposed Approach

MapReduce is known as one of the most powerful framework to processes massive datasets. In order to solve problems of MapReduce platforms, researchers integrate MapReduce into scientific workflows like in [12]. We integrates MapReduce jobs in the model formulation which can be denoted as:

$$W = (N, J, E, D) \tag{1}$$

Where, W denotes the workflow, N is a set of Nodes, as vertices in a DAG (Directed Acyclic Graph), which contains a number of jobs. J is a number of MapReduce jobs which can be split into map tasks and reduce tasks. D represents a set of datasets. E is a set of directed edges. The data flow is controlled by the dependencies between datasets used by jobs. Each job ji, can be composed by one or several map tasks and reduce tasks. The control flow is measured by the dependencies between jobs of nodes specified by directed edges between nodes. A directed edge between N1 and N2 means that j4 depends on j1, j2 and j3. N1 produces d3 and j4 consumes it as an input.

In Fig. 2, N = {N1, N2, N3, N4, N5}; d1, d2 are the input data of t1; d3 is produced after executing task t1; d3 and d4 are the input data of t2; d5 is the input data of t3; d6 is the input data of t4 and t5; j1, j2, j3 compounds t1; j4 compounds t2; j5 compounds t4; j6 compounds t3; j7, j8 compounds t5. d1 and d2 belongs to DC1;

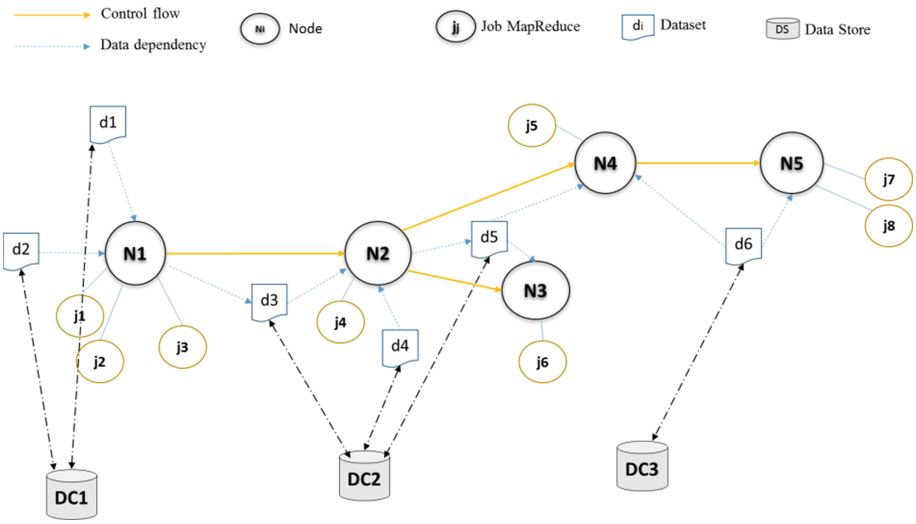


Fig. 2. Example of the proposed scientific application approach

d3, d4 and d5 have its place on DC2; d6 belongs to DC3. We can classify datasets into two groups: initial datasets and generated datasets like d3 and d5. So, we are dealing with two categories of data placement offline due to the initial datasets and online due to the generated datasets.

DC = set of data centers
DS = set of datasets
Dij = $|j_i \cap j_j|$ (i, j < n)

Which means the number of datasets required by both j_i and j_j forms a dependency matrix D. n is the total number of datasets.

A simple example, takes three datasets and two data centers. Job j1 requires two datasets ds1 and ds2, j2 requires the three datasets, and the dependency matrix is:

$$D = \begin{pmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

2.3 Heuristic Genetic Placement Algorithm

According to [5, 13], in order to solve the data placement problems, many optimization and evolutionary algorithms can be used, but the GA (Genetic Algorithm) (steps in Fig. 3) has been proven to have a better performance in many cases. GAs were first described by John Holland in the 1960s and further developed [14].

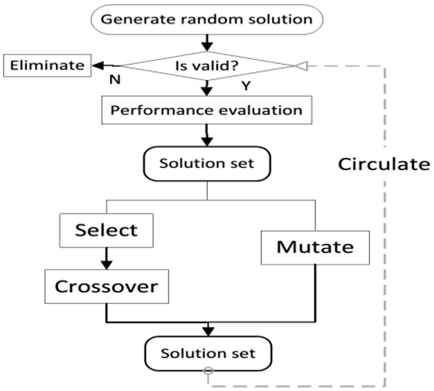


Fig. 3. Steps of the genetic algorithm [5]

A. Encoding Rules

To encode the problem in this paper to the genetic algorithm, we use integer-encoding rules and to each gene, there is a placement strategy. The placement of data in cloud is represented by two matrices. The first matrix is the data dependency and the second matrix is the MapReduce jobs control dependency matrix.

B. Individual and Population

A gene of individual is the placement of datasets on data centers. It means that the j^{th} dataset placed in the gj^{th} data center. For example there three data centers and 6 datasets. The gene code (311232) means that dataset 1 is placed in data center 3 and so on.

C. Fitness Function

The fitness function reveal the degree of dependency between datasets and the degree of control dependency among MapReduce jobs. We denoted Dt as the total degree of dependency, which is defined as:

$$D_t = \sum dsi, dsj + \sum ji, jj$$

The fitness function is: 1/ Dt

D. Optimization Placement Strategy

We propose below a heuristic to minimize data movement:

Steps for proposed heuristic algorithm

1. For i = 1, ….., n do
2. Calculate the dependencies of dataset dsi with each DC,
   Calculate the dependencies of job jj with each DC
3. Get the maximal dependency obtained in step 2
4. Place the dataset dsi to the suitable data center
   Place the job jj to the suitable node
5. If the size of dsi exceeds the storage capacity of the selected data center, find the next maximal dependency until this dataset can be placed in the suitable data center and the job jj to the suitable node;

3 Related Works

In our proposed approach, we investigate the combination of three concepts, which are the placement of massive data, the scientific workflow and the MapReduce model. We conducted our state of the art through the combination of these concepts as denoted in Table 1. To compare works in these domains we created three pairs that resulted in 'Placement in MapReduce', 'MapReduce with scientific workflows' and 'Placement of massive-data in scientific workflows'.

Table 1. Comparative study of data placement approaches

Authors	Placement in MapReduce		MapReduce with scientific workflows		Placement of massive data in scientific workflows	
	Jobs dependencies		With placement?		Total dependency	
	Yes	No	Yes	No	Yes	No
[Our approach]	*		*		*	
[Q. Zhao et al. 2016]						*
[K. Deng et al. 2015]					*	
[M. Ebrahimi et al. 2015]						*
[L. Cui et al. 2015]					*	
[Q. Zhao et al. 2015]						*
[L. Zeng et al. 2015]						*
[N. Mohamed et al. 2014]				*		
[M. Wang et al. 2014]						*
[J. Wang et al. 2014]		*				
[F. Ma et al. 2012]						*
[Z. Tang et al. 2012]				*		
[X. Fei et al. 2012]				*		
[N. Maheshwari et al. 2012]		*				
[E. Zhao et al. 2012]						*
[D. Yuan et al. 2011]						*
[Y. He et al. 2011]		*				
[P. Nguyen et al. 2011]				*		
[D. Yuan et al. 2010]						*
[X. Fei et al. 2009]				*		

3.1 Placement in MapReduce

In this section, we takes works that consider placement in MapReduce and then we classified them between two types to know which contains dependencies between MapReduce jobs and which not. Almost of papers [15–17] mentioned in this section do

not care about job dependency. We think as explained in Sect. 2, that dependencies between jobs is an important factor that affects the placement of massive data in such environments. Some works considers dependencies on data, like in [15]. There is no link between jobs in the case of multiple jobs with multiple datasets. The placement of jobs is treated separately. Therefore, in this work, it is more interesting to place jobs that have common interest together. In [16], authors proposes an energy efficient data placement algorithm that conserve energy for large data centers running MapReduce jobs. In addition, it concerns also data dependencies between jobs. There are some works like [17], which ameliorate the placement of data in MapReduce by proposing a placement storage structure for Hadoop and how to better store data in HDFS blocs. Each bloc is stored independently.

3.2 MapReduce with Scientific Workflows

In this part and according to Table 1, we cover works deploying scientific workflows, which integrates the MapReduce model, and we classified them into two types to know which presents placement algorithms and which not. All works founded [12, 18–21] do not holds placement solution. In [19], authors proposes a dataflow based scientific workflow in which the jobs are Map and Reduce. In [19, 20], authors considers dependencies between datasets. The authors in [12], proposes a MapReduce-enabled Scientific Workflow Framework with static optimization scheduling policy in heterogeneous clusters to get higher performance. However, the dependency between data is considered. In [21], authors proposes a workflow system for integrating structure, and orchestrating MapReduce jobs for scientific data intensive workflows. All cited works in this sector do not presents control dependencies in the workflow proposed.

3.3 Placement of Massive Data in Scientific Workflows

In this part and refers to Table 1, we consider works that concerns the placement of massive data in scientific workflow. Then we classified them into two parts, the first one contains total dependencies that means control and data dependencies. The second one only one type of workflow dependency. The paper in [23], proposes a data placement method, which could effectively reduce the data movement between the data centers and there is a data dependency between the nodes of datasets. In [24], authors proposes a Security-Aware and Budget Aware workflow scheduling strategy (SABA), which holds an economical distribution of tasks among the available CSPs (Cloud Service Providers) in the market, to provide customers with shorter make span as well as security services. The proposed approaches to cluster tasks based on data dependency. A matrix bases k-Means clustering strategy for data placement in scientific cloud workflows systems is done in [25]. The proposed strategy attempts to minimize the total data movement during the execution of workflows by placing and data dependency is reflected. In [26], authors proposes an algorithm that can calculate the minimum cost for intermediate dataset storage in scientific cloud workflows systems. It builds an intermediate data dependency graph (IDG) from the data provenances in

scientific workflows. Besides, another work proposes a data placement strategy based on data dependency clustering for scientific workflow in heterogeneous cloud in [27]. In [4], and [7], data dependency is taken into consideration. At the best of our knowledge, there are two works presenting a total dependency (control and data flow), one presents a K-cut graph-partitioning algorithm to minimize the volume of data transfer [22]. Besides, another propose a genetic algorithm based data replica placement strategy to reduce data transmissions in cloud [5].

4 Simulation Results

In the experiment, we run the simulation with test workflows on twenty data centers and we evaluate the number of data movements. We compare our proposed strategy to the random strategy where datasets and jobs are distributed to data centres and nodes randomly. The random strategy is the random data placement strategy used in Hadoop Distributed File System [28]. In Fig. 4, the number of data movements is indicated with diverse number of datasets. We assume that there is one job MapReduce per node in the proposed approach and in a future work we will adopt more than one job in each node. We execute a MapReduce job on the Ambari console TechSandBox lunched on my private cloud space offered by IBM demo cloud, as shown in Fig. 5.



Fig. 4. Data movements

The experiment is taken on a laptop with Intel Core i3 380 M 2.53 GHz, RAM 6 GB, DISK 500 GB, Bandwidth 100 MB.

From Fig. 4, with the increase of the number of data sets, the number of data movements is also increasing. The results shown in Fig. 4 proves that our proposed placement strategy can reduce the number of data movements better than the random strategy.

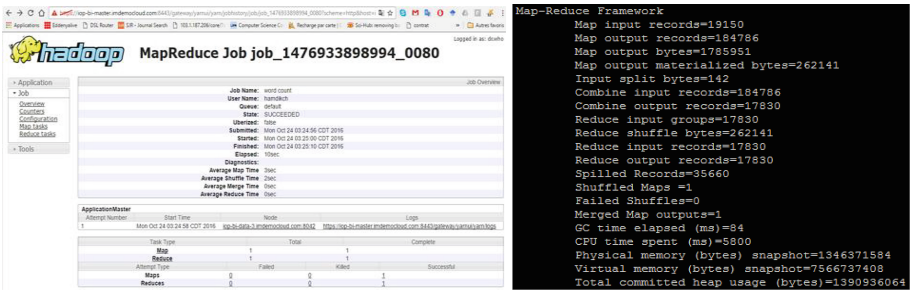


Fig. 5. Running a MapReduce job with Ambari

5 Conclusion

This paper outlines the current issues for data placement in scientific applications. It suggests an approach that can reduce the number of data movements. The proposed approach in this paper integrates the MapReduce model on the scientific workflow in Cloud Computing environments. Precisely, it proposes a placement of multiple MapReduce jobs on scientific workflow with control and data dependencies between jobs. In the future, we will use an online-shared environment for data science like the Ambari Console for testing multiple Hadoop jobs per node on real cloud environment.

Acknowledgments. The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program.

References

1. Mell, P., Grance, T.: The NIST definition of cloud computing recommendations of the National Institute of Standards and Technology. NIST Special Publication, vol. 145, p. 7 (2011)
2. Haghighat, M., Zonouz, S., Abdel-Mottaleb, M.: Expert systems with applications CloudID: trustworthy cloud-based and cross-enterprise biometric identification. Expert Syst. Appl. **42**, 7905–7916 (2015)
3. Coalition, W.M.: Workflow management coalition terminology and glossary
4. Ebrahimi, M., Mohan, A., Kashlev, A., Lu, S.: BDAP: a big data placement strategy for cloud-based scientific workflows. In: 2015 IEEE First International Conference on Big Data Computing Service and Applications, pp. 105–114 (2015)
5. Cui, L., Zhang, J., Yue, L., Shi, Y., Li, H., Yuan, D.: A genetic algorithm based data replica placement strategy for scientific applications in clouds. Trans. Serv. Comput. **1374**, 1–13 (2015)
6. Er-Dun, Z., Yong-Qiang, Q., Xing-Xing, X., Yi, C.: A data placement strategy based on genetic algorithm for scientific workflows. In: 2012 Eighth International Conference on Computational Intelligence and Security, pp. 146–149 (2012)

7. Ebrahimi, M., Mohan, A., Lu, S., Reynolds, R.: TPS : a task placement strategy for big data workflows, pp. 523–530 (2015)
8. Dean, J., Ghemawat, S.: MapReduce : simplified data processing on large clusters. In: OSDI 2004: Proceedings of 6th Symposium Conference on Operating Systems Design and Implementation, pp. 6, 1–13 (2004)
9. Song, J., He, H., Wang, Z., Yu, G., Pierson, J.-M.: Modulo based data placement algorithm for energy consumption optimization of MapReduce system. J. Grid Comput. (2016)
10. Deelman, E., Gannon, D., Shields, M., Taylor, I.: Workflows and e-Science: an overview of workflow system features and capabilities. Futur. Gener. Comput. Syst. **25**, 528–540 (2009)
11. Kelly, P.M.: Applying Functional Programming Theory to the Design of Workflow Engines. Science (2011)
12. Tang, Z., Liu, M., Li, K., Xu, Y.: A MapReduce-enabled scientific workflow framework with optimization scheduling algorithm. In: Proceedings of Parallel and Distributed Computing, Applications and Technologies. PDCAT, pp. 599–604 (2012)
13. Mitchell, M.: Genetic algorithms: an overview. Complexity **1**, 31–39 (1995)
14. Atay, Y., Kodaz, H.: Intell. Evol. Syst. **5**, 43–55 (2016)
15. Wang, J., Shang, P., Yin, J.: DRAW: a new data-gRouping-aware data placement scheme for data intensive applications with interest locality. In: Li, X., Qiu, J. (eds.) Cloud Computing for Data-Intensive Applications, pp. 149–174. Springer, New York (2014)
16. Maheshwari, N., Nanduri, R., Varma, V.: Dynamic energy efficient data placement and cluster reconfiguration algorithm for MapReduce framework. Futur. Gener. Comput. Syst. **28**, 119–127 (2012)
17. He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: RCFile: A fast and space-efficient data placement structure in MapReduce-based warehouse systems. In: Proceedings of International Conference on Data Engineering, pp. 1199–1208 (2011)
18. Mohamed, N., Maji, N., Zhang, J., Timoshevskaya, N., Feng, W.C.: Aeromancer: a workflow manager for large-scale MapReduce-based scientific workflows. In: Proceedings of 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications. TrustCom 2014, pp. 739–746 (2015)
19. Fei, X., Lu, S.: A dataflow-based scientific workflow composition framework. IEEE Trans. Serv. Comput. **5**, 45–58 (2012)
20. Fei, X.F.X., Lu, S.L.S., Lin, C.L.C.: A MapReduce-enabled scientific workflow composition framework. In: 2009 IEEE International Conference on Web Services, pp. 663–670 (2009)
21. Nguyen, P., Halem, M.: A MapReduce workflow system for architecting scientific data intensive applications. In: Proceedings of the 2nd International Workshop on Software Engineering for Cloud Computing, pp. 57–63 (2011)
22. Deng, K., Ren, K., Zhu, M., Song, J.: A data and task co-scheduling algorithm for scientific cloud workflows. IEEE Trans. Cloud Comput. **7161**, 1 (2015)
23. Ma, F., Yang, Y., Li, T.: A Data placement method based on bayesian network for data-intensive scientific workflows. In: 2012 International Conference on Computer Science & Service System, pp. 1811–1814 (2012)
24. Zeng, L., Veeravalli, B., Li, X.: SABA: a security-aware and budget-aware workflow scheduling strategy in clouds. J. Parallel Distrib. Comput. **75**, 141–151 (2015)
25. Yuan, D., Yang, Y., Liu, X., Chen, J.: A data placement strategy in scientific cloud workflows. Futur. Gener. Comput. Syst. **26**, 1200–1214 (2010)
26. Yuan, D., Yang, Y., Liu, X., Chen, J.: On-demand minimum cost benchmarking for intermediate dataset storage in scientific cloud workflow systems. J. Parallel Distrib. Comput. **71**, 316–332 (2011)

27. Zhao, Q., Xiong, C., Zhao, X., Yu, C., Xiao, J.: A data placement strategy for data-intensive scientific workflows in cloud. In: Proceedings of IEEE/ACM 15th International Symposium on Cluster, Cloud and Grid Computing. CCGrid 2015, pp. 928–934 (2015)
28. Hadoop. http://hadoop.apache.org/. Accessed 10 Oct 2016

A Novel Simulated Annealing-Based Learning Algorithm for Training Support Vector Machines

Madson L. Dantas Dias and Ajalmar R. Rocha Neto(✉)

Department of Teleinformatics, Federal Institute of Ceará (IFCE),
Av. Treze de Maio, 2081, Benfica CEP, Fortaleza, Ceará 60040-215, Brazil
madson.dias@ppgcc.ifce.edu.br, ajalmar@ifce.edu.br

Abstract. A theoretical advantage of large margin classifiers such as support vector machines (SVM) concerns the empirical and structural risk minimization which balances the complexity of the model against its success at fitting the training data. Metaheuristics have been used to work with SVMs in order to select features, tune hypeparameters or even achieve a reduced-set of support vectors. In spite of such tasks being interesting, metaheuristics such as simulated annealing (SA) do not play an important role in the process of solving the quadratic optimization problem, which arises from support vector machines. To do so, well-known methods such as sequential minimal optimization, kernel adatron or even classical mathematical methods have been used with this goal. In this paper, we propose to use simulated annealing in order to solve such a quadratic optimization problem. Our proposal is interesting when compared with those aforementioned methods, since it is simple and achieved similar (or even higher) accuracy and high sparseness in the solution.

Keywords: Support vector machines · Simulated annealing · Learning methods

1 Introduction

Pattern classification methods aims at providing functions that defines the relation between input vectors and their class labels. Differently from artificial neural networks (ANNs) and decision trees training process, the training process of large margin classifiers such as support vector machines (SVM) relies on minimizing the empirical and structural risk [19]. Minimizing the empirical risk means to reducing the classification error on training dataset and minimizing the structural risk is related to reducing the classification error on unseen patterns. Since learning algorithms are mostly based on minimizing only the empirical risk, this is an important characteristic of SVM over other classifiers.

The authors would like to thank the IFCE and CAPES for supporting their research.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_34

In order to train SVMs, one must solve a quadratic programming problem. SVMs have two formulations for its quadratic programming problem, the primal and the dual. The dual optimization problem is presented in terms of Lagrange multipliers and the bias. There are a few methods, which have been used to obtain the Lagrange multipliers and the bias, such as sequential minimal optimization (SMO) [13], kernel adatron (KA) [1] or even classical mathematical methods based on numerical optimization [5]. SMO solves the SVM quadratic programming problem in an analytical way by decomposing it into quadratic programming sub-problems and solving the smallest possible optimization problem, which involves two Lagrange multipliers, at each step. KA is an online method that solves the SVM quadratic programming problem by using first order information. Classical mathematical methods, henceforth called QP, solve the quadratic programming problem by numerical optimization, which unfortunately demands for the manipulation of large matrices leading to more numerical precision errors.

Metaheuristics such as simulated annealing (SA) and genetic algorithms (GAs) are an alternative to methods based on numerical optimization, such as those ones included in classical mathematical libraries. SA also aims at generating useful solutions to search problems. Due to the underlying features of these metaheuristics, some optimization problems can be solved without supposing linearity, differentiability, continuity or convexity of the objective function. Unfortunately, these desired properties are not found in several classical mathematical methods found in literature. Metaheuristics have been used with SVMs in classification tasks mainly to tune their parameters, select features and obtain reduced-set of support vectors [14,16].

The first attempt to use metaheuristics (in fact, GAs) in order to solve the primal problem of SVMs for classification tasks has demanded some change on its primal problem formulation [3,18]. A similar approach in terms of modeling was proposed but to train SVMs for regression tasks [17]. Similarly, the proposed regression method also models a solution for the primal optimization problem. Besides these attempts, a GA-based proposal for the dual formulation was presented in [10]; however, the problem was changed in order to leave one of the constraints out and the bias was defined as zero.

Another attempt to train SVMs based on PSO, called Linear PSO, was applied [11,12] and extensions [8,15,20]. However, in such a proposal the Karush-Kuhn-Tucker (KKT) conditions are not fully satisfied. Therefore, to the best of our knowledge, there is not in the literature an metaheuristics-based method, which deals with the quadratic optimization problem of SVM in its default dual formulation and, as consequence, also keeping the primal problem as initially proposed by Vapnik [19].

In this paper, we introduce a novel learning approach based on simulated annealing for training easily (SATE) support vector machines. To do so, we model an instance of SA to handle the dual optimization problem and its constraints in order to obtain the Lagrange multipliers and the bias for the decision function. In terms of SA, we propose a solution modeling, as well as

a neighborhood and an energy function. The remaining part of this paper is organized as follows. In Sect. 2 we review the fundamentals of support vector machines. After that, in Sect. 3, we briefly introduce simulated annealing and then in Sect. 4 is presented our proposal. In Sect. 5, we present our simulations and, after all, the paper is concluded in Sect. 6.

2 Support Vector Machines

Consider a training dataset $\{\mathbf{x}_i, y_i\}_{i=1}^{l}$, so that $\mathbf{x}_i \in \mathbb{R}^p$ is an input vector and $y_i \in \{-1, +1\}$ are the corresponding class labels. For soft margin classification, the SVM [19] primal problem is defined as

$$\min_{\mathbf{w},\xi_i} \left\{ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \right\}, \tag{1}$$

$$\text{s.t.} \quad y_i[(\mathbf{w}^T\mathbf{x}_i) + b] \geq 1 - \xi_i, \tag{2}$$

$$\xi_i, \geq 0. \tag{3}$$

where $\{\xi_i\}_{i=1}^{l}$ are slack variables, b is the bias and C is a cost parameter that controls the trade-off between allowing training errors and forcing rigid margins.

The SVM dual optimization problem can also be presented as

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i d_j \mathbf{x}_i^T\mathbf{x}_j \right\}, \tag{4}$$

$$\text{s.t.} \quad \sum_{i=1}^{l}\alpha_i y_i = 0, \tag{5}$$

$$0 \leq \alpha_i \leq C. \tag{6}$$

After solving the problem and having the Lagrange multipliers and the bias by Sequential Minimal Optimization (SMO) [13], Kernel Adatron (KA) [1] or classical mathematical optimization methods (QP) [5], the output can be computed based on the decision function described as

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{l}\alpha_i y_i \mathbf{x}^T\mathbf{x}_i + b \right). \tag{7}$$

where $\mathbf{x}^T\mathbf{x}_i$ can be replaced here and in Eq. (4) by a kernel function $k(\mathbf{x}, \mathbf{x}_i)$.

3 Simulated Annealing

Simulated Annealing (SA) is an adaptation of the Metropolis-Hastings algorithm and was independently described in [4,7]. SA is a popular local search meta-heuristic used to deal with discrete and continuous optimization problems. The key feature of simulated annealing is to provide a means to escape from local

optimum by allowing hill-climbing moves towards worse objective function values in hope that a global optimum is found.

The parameters for SA algorithm are an initial solution ($\boldsymbol{\alpha}_0$), an initial temperature (T_0), a final temperature (T_f), a temperature cooling schedule and a desired value of internal energy (S_f). Besides that, some functions must be defined, such as the neighbor function NEXT($\boldsymbol{\alpha}$), which generates solutions at the neighborhood of $\boldsymbol{\alpha}$; a repetition schedule function F(T), that defines the number of iterations executed at each temperature T; and the system energy function H($\boldsymbol{\alpha}$). As output for the SA algorithm, we have a solution $\boldsymbol{\alpha}^*$, which minimizes the internal energy H($\boldsymbol{\alpha}$) of the system.

4 Our Proposal: SATE

Our proposal, henceforth called SATE (Simulated Annealing for Training Easily support vector machines), relies on simulated annealing algorithm, which maximizes the expression presented in Eq. (4) and aims at obtaining the Lagrange multipliers and the bias.

In order to describe our proposal in terms of a SA-based optimization problem, we have to explain in details the solution modeling, the neighborhood function NEXT($\boldsymbol{\alpha}$) and the energy function H($\boldsymbol{\alpha}$). Thus, we present these requirements in the next subsections.

4.1 Solution and Initialization

The SA solution we propose is an vector of real values as follows

$$\boldsymbol{\alpha} = [\alpha_1 \;\; \alpha_2 \;\; \ldots \;\; \alpha_i \;\; \ldots \;\; \alpha_l], \tag{8}$$

where α_i is the value of the i-th Lagrange multiplier for the i-th training pattern. In this encoding, α_i different from zero is support vectors and, therefore, each solution has as many support vectors as the number of components not equal "zero".

In the beginning, we only have two Lagrange multipliers with non-zero values so that each class has only one support vector. Since the Lagrange multipliers are from different classes and have the same value, the solution complies with the constraint presented in Eq. (5). In order to maintain the constraint shown in Eq. (6), the value to be assign for both Lagrange multipliers is generated at random in $[0, C]$. The idea is to achieve a sparse solution from the sparsest one.

4.2 Neighborhood Function

The neighborhood function NEXT($\boldsymbol{\alpha}$) is designed to provide a new solution $\boldsymbol{\alpha}^{new}$ from a previous one. To do so, we propose three functions in order to generate solution candidates, named modify, add and swap. The entire neighborhood function is presented below.

NEXT($\boldsymbol{\alpha}$)
1 $\boldsymbol{\alpha}^{mdf} \leftarrow$MODIFY($\boldsymbol{\alpha}$)
2 $\boldsymbol{\alpha}^{add} \leftarrow$ADD($\boldsymbol{\alpha}$)
3 $\boldsymbol{\alpha}^{swp} \leftarrow$SWAP($\boldsymbol{\alpha}$)
4 **if** RAND $[0, 1]$
5 **then return** $\underset{\boldsymbol{\alpha}}{\text{argmin}}\ \mathrm{H}(\boldsymbol{\alpha}) = \{\boldsymbol{\alpha} \mid \boldsymbol{\alpha} \in \{\boldsymbol{\alpha}^{swp}, \boldsymbol{\alpha}^{mdf}, \boldsymbol{\alpha}^{add}\}\}$
6 **else return** $\underset{\boldsymbol{\alpha}}{\text{argmin}}\ \mathrm{H}(\boldsymbol{\alpha}) = \{\boldsymbol{\alpha} \mid \boldsymbol{\alpha} \in \{\boldsymbol{\alpha}^{swp}, \boldsymbol{\alpha}^{mdf}\}\}$

Modify Function: in the modify function two positions from a certain class (positive or negative) in a solution $\boldsymbol{\alpha}$ are selected at random and then a linear combination of the values at these positions are made for new Lagrange multiplier values. Firstly, two integer values i and j are randomly generated in range of 1 and l such that $i \neq j$ and $y_i = y_j$. After that, a real value γ in $[0, 1]$ is generated at random in order to perform the linear combination of α_i^{new} and α_j^{new}; so the function generates soluction as follows

$$\alpha_i^{new} = \gamma \cdot \alpha_i^{old} + (1 - \gamma) \cdot \alpha_j^{old}, \tag{9}$$

$$\alpha_j^{new} = \gamma \cdot \alpha_j^{old} + (1 - \gamma) \cdot \alpha_i^{old}. \tag{10}$$

Add Function: in order to include new Lagrange multipliers in the actual solution, we select at random two positions (i and j) belonging to different classes from those with zero values and then we assign a certain real value τ in $[0, C]$ also chosen randomly to them.

Swap Function: the swap function works by randomly selecting two positions in the solution $\boldsymbol{\alpha}$ and then swapping the values at these positions. Since this procedure can violate the constraint presented in Eq. (5), we must repair the candidate solution by an adjustment algorithm (this strategy is common in metaheuristics [9]).

Adjustment Function: to do so, we propose an algorithm to repair the solution in order to make it feasible. First of all, let us re-write the constraint presented in Eq. (5) as

$$\sum_{i=1}^{l} \alpha_i y_i \Delta_i^+ + \sum_{i=1}^{l} \alpha_i y_i \Delta_i^- = 0, \tag{11}$$

where

$$\Delta_i^+ = \begin{cases} 1, & \text{if } y_i = +1; \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \Delta_i^- = \begin{cases} 1, & \text{if } y_i = -1; \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

By swapping values we obtain a sum $s = \sum_{i=1}^{l} \alpha_i y_i \Delta_i^+ + \sum_{i=1}^{l} \alpha_i y_i \Delta_i^-$, which most likely violates such a constraint (i.e., $s \neq 0$). Taking this into account, the main idea of our adjustment algorithm for a certain solution $\boldsymbol{\alpha}$ is to iteratively reduce an amount of a certain Lagrange multiplier selected at random until

complying with the constraint presented in Eq. (11). If the sum is larger than zero (i.e., if $s^+ > s^-$, where $s^+ = \sum_{i=1}^{l} \alpha_i y_i \Delta_i^+$ and $s^- = \sum_{i=1}^{l} \alpha_i y_i \Delta_i^-$), we reduce an amount of a certain Lagrange multiplier with $y_i = +1$, but this amount must not be larger than the current Lagrange multiplier value (recall that $0 \leq \alpha_i \leq C$). We perform this process until the constraint be complied. However, when the sum is lower than zero (i.e., $s^- > s^+$), we reduce an amount of a certain Lagrange multiplier with $y_i = -1$. Notice that if the sum equals zero ($s = 0$), the solution complies with the constraint and the adjustment is not required.

It is important to highlight that, if the sum s is larger than zero, a certain amount could be added up to any Lagrange multiplier value with $y_i = -1$ instead of reducing it; however, reducing is a better choice, since we can achieve sparser solutions (i.e., classifiers with less patterns as support vectors).

Our adjustment algorithm ADJUSTMENT($\boldsymbol{\alpha}$) for a certain solution $\boldsymbol{\alpha}$ is shown below. Besides that, we present an example for constraint violation and its adjustment such an algorithm in Fig. 1. In this figure, the algorithm steps from an unfeasible solution to a feasible one are presented in details.

ADJUSTMENT($\boldsymbol{\alpha}$)
1 $\boldsymbol{\alpha}^{new} \leftarrow \boldsymbol{\alpha}$
2 $s \leftarrow \sum_{i=1}^{l} \alpha_i y_i$
3 **while** $s \neq 0$ **do**
4 $\quad s^+ \leftarrow \sum_{i=1}^{l} \alpha_i^{new} y_i \Delta_i^+$
5 $\quad s^- \leftarrow \sum_{i=1}^{l} \alpha_i^{new} y_i \Delta_i^-$
6 $\quad$ **if** $s^+ > s^-$ **then**
7 $\quad\quad k \leftarrow$ RAND$[1, l]$ such that $y_k = 1$;
8 $\quad$ **else**
9 $\quad\quad k \leftarrow$ RAND$[1, l]$ such that $y_k = -1$;
10 $\quad$ **end if**
11 $\quad$ **if** $\alpha_k^{new} > s$ **then**
12 $\quad\quad \alpha_k^{new} \leftarrow \alpha_k^{new} - s$
13 $\quad$ **else**
14 $\quad\quad \alpha_k^{new} \leftarrow 0$
15 $\quad$ **end if**
16 $s \leftarrow \sum_{i=1}^{l} \alpha_i^{new} y_i$
17 **end while**
18 **return** $\boldsymbol{\alpha}^{new}$

4.3 Energy Function

The energy function for our proposal is based on the SVM dual optimization problem presented in Eq. (4). As we are working with an optimization problem in terms of minimizing, our energy function is given by

$$\mathrm{H}(\boldsymbol{\alpha}) = \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i d_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{i=1}^{l} \alpha_i. \tag{13}$$

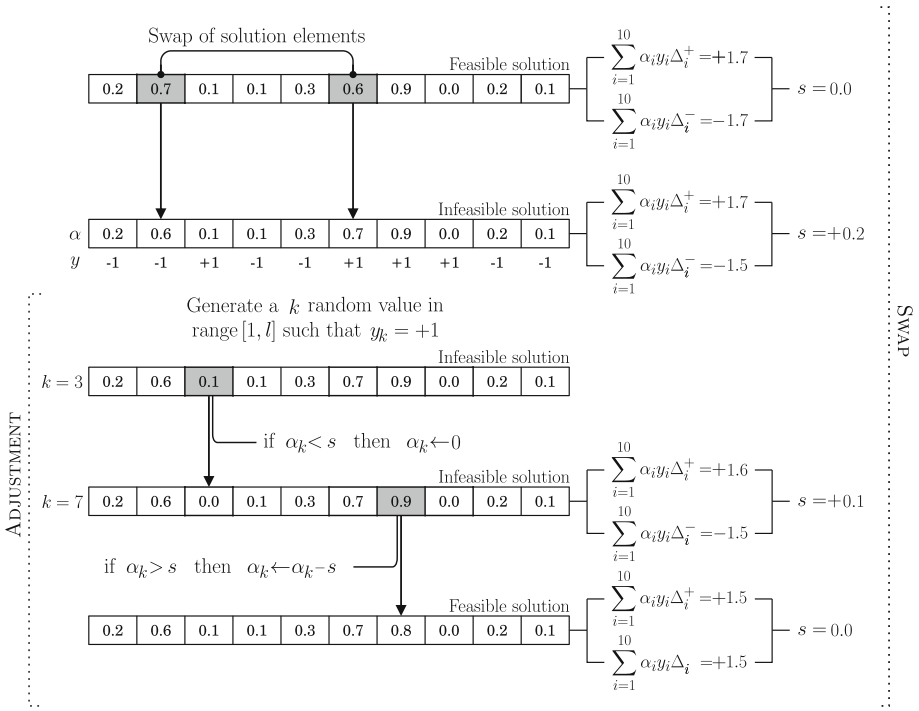


Fig. 1. Constraint violation example and its repairing.

4.4 Bias Computation

After having the Lagrange multipliers by the SATE algorithm, we can compute the bias b for Eq. (7) as follows

$$b = \frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{j=1}^{l} y_j \alpha_j \mathbf{x}_i^T \mathbf{x}_j \right). \tag{14}$$

5 Simulations and Discussion

We carried out some simulations on real world datasets and present some results for SVM trained with SATE, LPSO, SMO, QP and KA in this section. As for datasets, we used some UCI binary datasets [2]; namely, Australian Credit Approval (AUS) with 690 patterns, Haberman's Survival (HAB) with 306 patterns, Pima Indians Diabetes (PID) with 768 patterns and Vertebral Column Pathologies dataset (VCP) with 310 patterns.

In our simulations, 80% of the data examples were randomly selected for training purposes and so the remaining 20% of the examples were used for

assessing the classifiers' generalization performance. We carried out 30 executions on each dataset. All experiments were performed with linear kernel and tuned parameter C of classifiers were tuned by applying grid search with 5-fold cross-validation over the training dataset. We have done this in order to follow the common strategy use in SMO, KA and QP for parameter tuning.

In Table 1, we report performance metrics (mean value and standard deviation of the Accuracy and Number of Support Vectors (#SVs)) for SVM trained by SATE, LPSO, SMO, KA and QP in aforementioned methodology 80%–20% on testing set averaged over 30 independent runs. We also show results of applying the Wilcoxon rank-sum test [6]. By analyzing the Table 1, one can conclude that the performances for SATE were equivalent or even superior to those achieved by the LPSO, SMO, KA and QP when applied to the aforementioned datasets. Moreover, one can also see that our proposal achieves sparse solutions, as well as the SATE was sparser than the other classifiers for AUS and HAB. In addition to these results, SATE achieved the highest accuracy for AUS, PID and VCP.

Table 1. Performance metrics for SATE, LPSO, SMO, QP and KA; #SVs and results for Wilcoxon rank-sum test (WRST), so that ✓ and ✗ mean equivalent and inferior results.

Dataset	Algorithm	Accuracy (%)	# SVs	WRST
AUS	SATE	87.1 ± 2.8	321.6 ± 25.9	
	LPSO	85.6 ± 2.4	524.6 ± 20.5	✗
	SMO	86.8 ± 3.1	402.8 ± 5.0	✓
	QP	87.1 ± 2.3	530.9 ± 25.0	✓
	KA	87.1 ± 2.4	490.5 ± 20.0	✓
HAB	SATE	72.5 ± 5.0	165.1 ± 7.9	
	LPSO	74.3 ± 4.6	193.7 ± 21.8	✓
	SMO	72.2 ± 5.8	197.9 ± 7.7	✓
	QP	71.6 ± 5.8	244.8 ± 0.5	✓
	KA	63.0 ± 9.0	241.6 ± 1.4	✗
PID	SATE	76.6 ± 3.4	393.3 ± 9.3	
	LPSO	70.4 ± 3.2	560.5 ± 20.4	✗
	SMO	75.4 ± 3.3	351.2 ± 8.9	✓
	QP	75.4 ± 3.6	605.0 ± 3.4	✓
	KA	71.1 ± 2.9	551.1 ± 9.9	✗
VCP	SATE	86.4 ± 5.4	89.7 ± 5.9	
	LPSO	61.5 ± 6.0	131.3 ± 16.1	✗
	SMO	83.7 ± 3.9	86.0 ± 8.0	✓
	QP	84.5 ± 3.8	124.9 ± 5.7	✓
	KA	67.1 ± 5.9	240.9 ± 5.1	✗

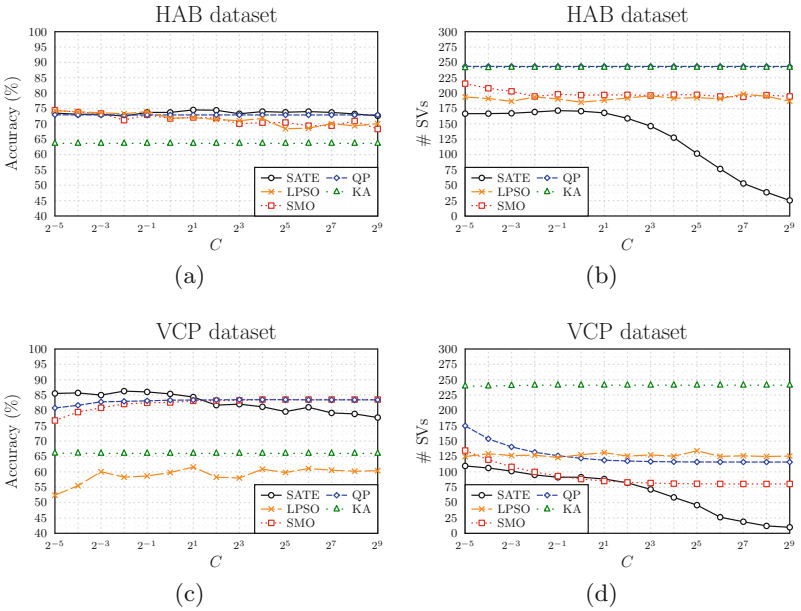
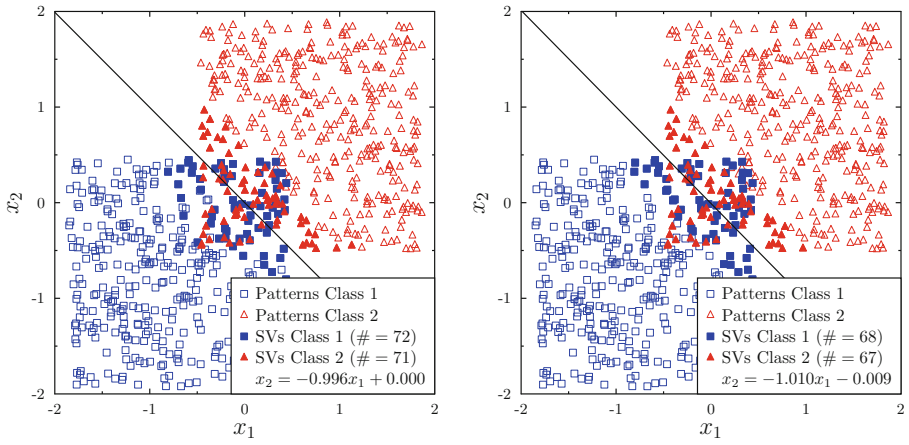


Fig. 2. (a) and (c) Accuracy versus hyperparameter C. (b) and (d) #SVs versus hyperparameter C of SATE, LPSO, SMO, KA and QP for HAB and VCP datasets.

In Fig. 2(a) and (c) are depicted the accuracy versus the hyperparameter C for SATE, LPSO, SMO, QP and KA when applied to HAB and VCP datasets, respectively. Besides that, we present in Fig. 2(b) and (d) the #SVs versus the trade-off C. By analyzing these figures, one can see that SATE has higher accuracies than the other algorithms (except for LPSO) at any value of C for HAB dataset. Moreover, the #SVs for SATE is lower than the #SVs for the other algorithms at any value of C for dataset HAB.

As for dataset VCP, SATE has higher accuracies than LPSO, SMO, QP and KA when C in $[2^{-5}, 2^{1}]$ and similar accuracies out of this range. Nevertheless, the #SVs for SATE are lower than LPSO, SMO, QP and KA when C in $[2^{-5}, 2^{-1}]$ and $[2^{2}, 2^{9}]$; but, close to the other algorithms out of these ranges. Overall, based on theses figures and on the Table 1, one can see that our proposal is very competitive in terms of accuracy, since SATE has similar (or even superior accuracies) when compared with SMO. Along with this interesting achievement, SATE is also competitive with respect to #SVs.

For a qualitative analysis, we have also applied SATE and SMO to an artificial problem presented in Fig. 3. Based on the Fig. 3, we can infer that SATE produced decision surfaces similar to the other algorithms. Besides that, one can see in the Fig. 3 that the number of SVs for SATE is almost the same as the number of SVs for SMO.



(a) SATE (b) SMO

Fig. 3. Decision surface and #SVs for SVM trained with (a) SATE and (b) SMO when applied to an artificial problem.

6 Conclusion

We propose an algorithm to solve the quadratic optimization problem that comes up with soft-margin support vector machines through simulated annealing. Our proposal called SATE is able to handle the dual optimization problem and then to obtain the Lagrange multipliers and the bias for SVMs. The constraints for SVMs were successfully embedded into a version of simulated annealing algorithm and the achievements indicate that the SATE works very well, providing a competitive classifier while maintaining its simplicity and sparseness. We have successfully modeled the solution in the simulated annealing framework, as well as the neighborhood function and the energy function. Moreover, SATE was able to train support vector machines and also to maintain or even to improve the classifier's performance. Along with these achievements, our proposal is interesting when compared with LPSO, SMO, QP and KA because it is simple with respect to modeling.

References

1. Anlauf, J.K., Biehl, M.: The adatron: an adaptive perceptron algorithm. Europhys. Lett. **10**(7), 687 (1989)
2. Blake, C., Merz, C.J.: UCI repository of machine learning databases (1998)
3. Dumitrescu, D., Preuss, M., Stoean, C., Stoean, R.: Evolutionary support vector machines and their application for classification (2006)
4. Černý, V.: Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. J. Optim. Theor. Appl. **45**(1), 41–51 (1985)
5. Gill, P.E., Murray, W., Wright, M.H.: Practical Optimization. Academic Press, London (1981)

6. Hollander, M., Wolfe, D.A., Chicken, E.: Nonparametric Statistical Methods. Wiley, New York (2013)
7. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science **220**(4598), 671–680 (1983)
8. Li, Y., Tong, Y., Bai, B., Zhang, Y.: An improved particle swarm optimization for svm training. In: Third International Conference on Natural Computation (ICNC 2007), vol. 2, pp. 611–615. IEEE (2007)
9. Michalewicz, Z., Janikow, C.Z.: Handling constraints in genetic algorithms. In: ICGA, pp. 151–157 (1991)
10. Mierswa, I.: Evolutionary learning with kernels: a generic solution for large margin problems. In: Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, pp. 1553–1560. ACM (2006)
11. Paquet, U.: Training support vector machines with particle swarms (2007)
12. Paquet, U., Engelbrecht, A.P.: A new particle swarm optimiser for linearly constrained optimisation. In: The 2003 Congress on Evolutionary Computation, CEC 2003, vol. 1, pp. 227–233. IEEE (2003)
13. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel Methods - Support Vector Learning. MIT Press, Cambridge, MA, USA (1999)
14. Neto, A.R.R., Barreto, G.A.: Opposite maps: vector quantization algorithms for building reduced-set SVM and lSSVM classifiers. Neural Process. Lett. **37**(1), 3–19 (2013)
15. Silva, A., Gonçalves, T.: Training support vector machines with an heterogeneous particle swarm optimizer. In: Tomassini, M., Antonioni, A., Daolio, F., Buesser, P. (eds.) ICANNGA 2013. LNCS, vol. 7824, pp. 100–109. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37213-1_11
16. Silva, D.A., Silva, J.P., Neto, A.R.R.: Novel approaches using evolutionary computation for sparse lssvms. Neurocomput. **168**, 908–916 (2015)
17. Stoean, R., Dumitrescu, D., Preuss, M., Stoean, C.: Evolutionary support vector regression machines. In: Eighth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2006, pp. 330–335 (2006)
18. Stoean, R., Preuss, M., Stoean, C., Dumitrescu, D.: Concerning the potential of evolutionary support vector machines. In: IEEE Congress on Evolutionary Computation, CEC 2007, pp. 1436–1443. IEEE (2007)
19. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience, New York (1998)
20. Yuan, H., Zhang, Y., Zhang, D., Yang, G.: A modified particle swarm optimization algorithm for support vector machine training. In: 2006 6th World Congress on Intelligent Control and Automation, vol. 1, pp. 4128–4132. IEEE (2006)

Fish School Search Variations and Other Metaheuristics in the Solution of Assembly Line Balancing Problems

I.M.C. de Albuquerque(✉), J.B.M. Filho, F.B.L. Neto, and A.M.O. Silva

Department of Computer Engineering,
Polytechnical School of Pernambuco, Recife, Brazil
{imca,jbmf,fbln,amos}@ecomp.poli.br

Abstract. Problems involving assembly lines optimization have been target of many researchers due to its relevance in practical applications. The Assembly Line Balancing Problem is directly related to the productivity of assembly lines. However, its NP-Hard nature makes the problem solution becomes non-trivial for exact procedures, which makes room for metaheuristic approaches. In this work, we applied six recent variations of the Fish School Search algorithm in the solution of the Simple assembly Line Balancing Problem Type 1 and compared either solution quality and convergence speed against the results obtained by other three metaheuristic procedures. The different approaches were compared through statistical tests and the results give an indication of which procedure is most suitable for this class of combinatorial optimization problems.

Keywords: Metaheuristics · Fish school search · Assembly line balancing problem · Genetic algorithms · Ant colony optimization · Particle swarm optimization

1 Introduction

Assembly line is a flow based production system in which the work units, denominated workstations, are disposed in a serial manner [1]. The work pieces travel along the line moved by a transport system. Assembly lines constitute the main production paradigm of the contemporary manufacturing industry and one of the main elements related with its productivity is the distribution of assembly tasks on each workstation, *i.e.* balancing the assembly line. The set of problems regarding balancing assembly lines is referred as Assembly Line Balancing Problem (ALBP).

All the relevant problem versions of the assembly line balancing family of problems are known to be NP-Hard [2]. Hence, different metaheuristic procedures have been studied in order to solve them in practicable time. Some of the mentioned approaches include Differential Evolution [3], Genetic

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_35

Algorithms [4], Genetic Programming [5], Particle Swarm Optimization [6] and Ant Colony Optimization [7].

The Fish School Search (FSS) algorithm is a population based procedure first proposed in the work of Bastos-Filho and Lima-Neto [8]. This algorithm was recently applied in the solution of Assembly Lines Balancing Problem [9]. Moreover, some variations of FSS were proposed in order to improve the search mechanisms employed [10,11].

This work aims to compare the performance of the Fish School Search algorithm against other metaheuristic procedures in the solution of a specific version of ALBP known as Simple Assembly Line Balancing Problem (SALBP). A non-parametric statistical test is employed for the production of reliable conclusions. Quality of the solutions provided by each approach is evaluated in terms of both: Objective Function and Convergence Speed.

This paper is organized as follows: in Sect. 2 the FSS algorithm as well as its variations are described. In Sect. 3, we provide some definitions and a brief description of the utilized SALBP setup. Section 4 presents the tests performed and the results obtained in all metaheuristic approaches applied.

2 Fish School Search Algorithm

FSS is a population based search algorithm inspired in the behavior of fishes while looking for food. Each fish n-dimensional position represents a possible solution to the problem. Fishes embed an innate memory of their successes: their weights. FSS is composed of feeding and movement operators (divided into three sub-components). In the following we provide a brief description of the aforementioned search procedure. The interested reader is recommended to refer to the work of Bastos-Filho and Lima-Neto [8] in order to have access to a complete description of the original algorithm.

Individual Component of Movement: In this movement component, every fish performs a local search through a random move. The new position will be accepted if the fitness of the fish improves with the position change. If it is not the case, the position will not change.

Feeding Operator: Each fish can increase its weight depending on its success in the search for food during the Individual movement.

Collective-instinctive Component of Movement: This movement is based on an average of the displacements performed by every fish within the Individual movement. The mean displacement is calculated taking into account the fitness improvements so that successful fishes have higher influence in the resulting direction of movement.

Collective-volitive Component of Movement: This operator is used in order to regulate the automatic switching of exploration/exploitation of the school during the search process. First, it is necessary to calculate the barycenter of the school. If the total school's weight has increased from last to current iteration, fishes are attracted to the barycenter. If the total school's weight has not improved, fishes are spread away from the barycenter.

2.1 FSS Variations

Some modifications were proposed in the original version of FSS, referred from now on as FSS-Vanilla, in order to improve its search mechanisms. In that initial version, fishes only move according to the Individual movement operator if they improve their fitness. However, in a very smooth search space, there would be many moving trials with no success and the algorithm could fail to converge. Moreover, the Collective-volitive movement was designed to regulate the exploration/exploitation ability of the algorithm along the search process, but this behavior depends on the possibility of the school total weight to reduce. If it does not happen, the ability of attracting fishes to school barycenter in order to exploit the search space will always predominate with relation to the ability of spreading the school away from the barycenter in order to allow exploration. To solve these issues, in the version of FSS proposed by Monteiro et al. [10] and denominated as FSS-SAR (Stagnation Avoidance Routine), authors introduce a parameter α decaying exponentially along with the number of iterations. This parameter measures the probability of a worsening allowance. It means that, every time a fish tries to move to a position that does not improve its fitness in the Individual Component of the movement, a random number is chosen and if it is smaller than α the movement is allowed.

Furthermore, two other variations of FSS-Vanilla were proposed recently [11]. The first one, referred as FSS-NF (Normalized Feeding), aims to provide a reliable weight parameter calculation in order to guarantee that if fish a has a higher weight when compared to fish b, the fitness of fish a is better than that of fish b.

The new feeding operator is defined according to:

$$W_i(t+1) = 1 + (W_{scale} - 1) * \frac{(F_i - F_{min})}{(F_{max} - F_{min})}, \tag{1}$$

where W_i and Fi represent fish i weight and fitness, respectively. F_{min} and F_{max} are the maximum and minimum fitness values found during all the search process. W_{scale} is a user-defined parameter indicating the maximum possible weight a fish can reach.

In addition to single FSS-NF, an elitist behavior was introduced in FSS originating one more variation referred as FSS-NPSS (Not Penalizing Static Success) in a sense that fishes which were not able to move within the Individual movement due to the fact that they have already reached a good fitness region can still be able to contribute in the Collective-instinctive movement operator.

Originally in FSS-Vanilla, fishes that do not improve cannot contribute to the Collective-instinctive movement operator. However, in FSS-NPSS, these fishes will have their displacement $\Delta \mathbf{x}_i$ and ΔW_i replaced by:

$$\Delta \mathbf{x}_i^* = step_{vol} \frac{\Delta \mathbf{x}_i^t}{step_{ind}^t}, \tag{2}$$

and:

$$\Delta W_i^* = max(\Delta W_i) \frac{W_i - 1}{W_{scale} - 1}, \tag{3}$$

where $\Delta \mathbf{x}_i^*$ and ΔW_i^* are fish i "fake" displacement and weight increase to be used within the Collective-instinctive movement. $step_{vol}$ is the current step (parameter indicating the maximum displacement that a fish can perform) of the Collective-volitive movement operator. $\Delta \mathbf{x}_i^t$ is the last displacement of fish i which occurred in iteration t and $step_{ind}^t$ is the step of the Individual movement operator in iteration t.

Either FSS-NF and FSS-NPSS can be combined with FSS-SAR raising two other variations known as FSS-NF-SAR and FSS-NPSS-SAR. In summary, six different versions of the Fish School Search algorithm are being considered in this work: FSS-Vanilla, FSS-SAR, FSS-NF, FSS-NF-SAR, FSS-NPSS and FSS-NPSS-SAR.

3 Simple Assembly Line Balancing Problem

Assembly Line Balancing Problem (ALBP) is the term used to refer to the decision problem of assigning tasks to workstations in a serial production system [12]. Each workstation is assigned to a worker or robot, with identical cycle time or production rate, given a group of elementary tasks with predefined processing times [13]. The tasks consist of elemental operations required to convert raw material into finished goods. ALBP must assign tasks to an ordered sequence of workstations such that the precedence relations among the tasks are satisfied and some performance measure is optimized [14].

The Simple Assembly Line Balancing Problem (SALBP) is a member of ALBP in which:

- *There is no product variety*: only one variant of a single homogeneous product is manufactured;
- *There is no process alternatives*: all operations are performed in a predetermined manner;
- *Each workplace belongs to a single station*: the assembly line consists of stations arranged in a serial manner and being independent of each other;
- *Static and deterministic task times*;
- *Indivisibility of tasks*: an operation cannot be split up between multiple workplaces;
- *One workplace per workstation*: each workstation contains exactly one workplace.

3.1 SALBP-1 with FSS

SALBP-1 is the version of SALBP aiming to minimize the number of workstations. The main issues regarding the solution of SALBP-1 using FSS are related to solution representation scheme and treatment of capacity and precedence constraints. We applied the task-oriented representation [15] in order to represent solutions. However, in FSS a fish position is composed by real numbers. For this reason, the random-keys [16] procedure was utilized to map a fish position vector into a task indexes vector.

As mentioned, two types of constraints have to be considered in order to solve SALBP-1 instances. A correction procedure [16] was employed in order to correct task sequences presenting precedence relations problems. Moreover, the capacity constraints are taken into account during the assignment procedure once it only allows tasks to be allocated if they do not extrapolate the cycle time. A detailed explanation of capacity constraint treatment in task-oriented approaches is provided by Scholl [15]. A pseudocode for the SALBP-1 solution with FSS can be seen in [10].

SALBP-1 utilizes the number of workstations as objective function. This fact brings a disadvantage in the use of FSS in order to tackle this problem once this kind of objective function implies that the search space is composed of many plateaus, which mainly has effect in the Individual movement operator. Once it only allows fishes to move if the fitness improve, if some fish is on a big plateau in the search space, it will get trapped there. To solve that, instead of the common use of the number of workstation as objective function, it was utilized:

$$minimize \left(m \times \sqrt{\sum_{k=1}^{m} (c - t_k)^2} \right), \tag{4}$$

where m is the number of stations, c is the cycle time and t_k is the workload at workstation k. This objective function simultaneously improves the number of workstations and the smoothness of the line balancing. Furthermore it changes the search space allowing more variation in the fitness when fishes move.

4 Tests and Results

4.1 Metaheuristic Approaches for Comparison

Genetic Algorithm: Four different Genetic Algorithms were considered in this comparison. The main difference between the chosen versions rely on the scheme utilized within the selection operator and in the treatment of precedence constraints. The selection schemes applied were *Roulette* and *Tournament*. The treatment of precedence constraints was performed in two different ways. In the first one, a random initial population was employed and standard crossover and mutations operators [17] were applied. Once the output of the operators is infeasible, Hamta et al. [16] approach, the same applied in FSS, was used in

order to correct it. The alternative approach is based on crossover and mutation operators designed to generate only feasible individuals [18]. In order to do so, a feasible (treated) initial population must be provided.

In summary, the four versions generated by the combination of the mentioned features are: (i) *GA-Roulette Treated Initial Population* (RTIP), (ii) *GA-Roulette Random Initial Population* (RRIP), (iii) *GA-Tournament Treated Initial Population* (TTIP) and (iv) *GA-Tournament Random Initial Population* (TRIP).

The crossover and mutation rates applied in the simulations in this work were 80% and 10%, respectively. An elitism rate of 20% was also employed.

Ant Colony Optimization: Three different versions of ACO were chosen to be tested: (i) *Ant System* (AS) [19], (ii) *Elitist Ant System* (EAS) [20] and (iii) *Ant Colony System* (ACS) [19].

The heuristic information η_j of task j employed in order to guide the search process was defined according to [21]:

$$\eta_j = \frac{t_j}{c} + \frac{|F_j^*|}{max|F_j^*|}, \tag{5}$$

where c is the cycle time and $|F_j^*|$ is the number of successors of task j.

α and β parameters used in tests performed were $\alpha = 3$ and $\beta = 1$. Pheromone evaporation was employed with $\rho = 0.1$.

Particle Swarm Optimization: In the PSO version chosen [22], for particle i position $\mathbf{x}_i$, its position in iteration $t + 1$ is defined by:

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \chi[\mathbf{v}_i + c_1 r_1(\mathbf{pb}_i - \mathbf{x}_i(t)) + c_2 r_2(\mathbf{gb}_i - \mathbf{x}_i(t))], \tag{6}$$

$$\chi = \frac{2}{|2 - (c_1 + c_2) - \sqrt{(c_1 + c_2)((c_1 + c_2) - 4)}|}, \tag{7}$$

where χ is known as constriction factor and r_1 and r_2 are uniformly distributed random numbers in the interval $[0; 1]$. In this version c_1 and c_2 must satisfy $c_1 + c_2 \geq 4$. For this work we have chosen $c_1 = c_2 = 2.1$. The solution of SALBP-1 with PSO follows the same flow as for FSS described in Sect. 3.1.

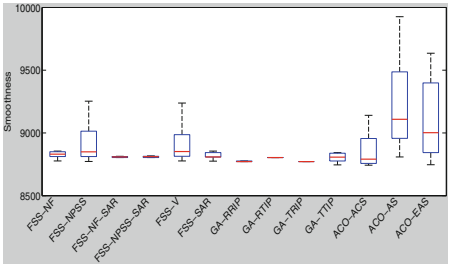
4.2 First Comparison: FSS, GA and ACO

In order to chose the best version of each technique in the solution of SALBP-1, all algorithms were used to solve three different test instances. All test cases were selected from the repository: http://www.assembly-line-balancing.de. Different problem sizes were chosen in order to evaluate the performance of each approach in problems with different complexity: (i) a large instance with 100 tasks referred in the data base as $n = 100_34$, (ii) a medium instance containing 50 tasks ($n = 50_25$) and (iii) a small problem with 20 tasks ($n = 20_26$). The minimum number of workstations is known *a priori* by the solution with an exact procedure

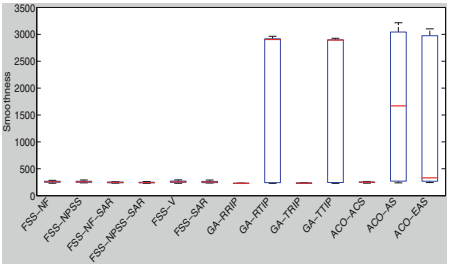
provided in the data base: 12, 6 and 15 for the small, medium and large instances, respectively.

The parameters used by all versions of FSS were [9]: $W_{scale} = 10000$, $Step_{ind} = 2$ and $Step_{vol} = 0.2$. For SAR algorithms, the decay mode of α occurs according to $\alpha = 0.8e^{-0.007t}$, where t is the current iteration.

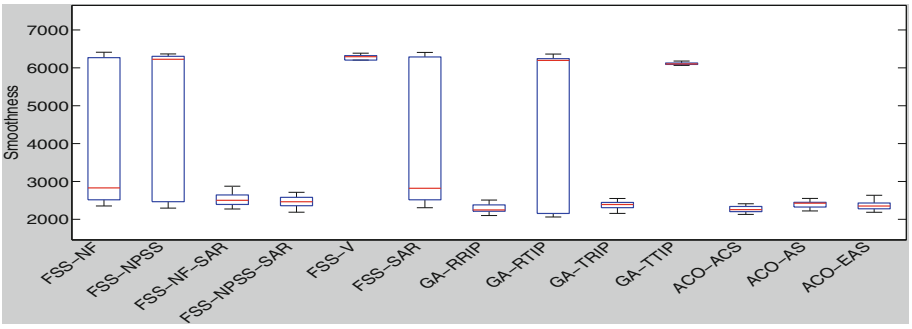
The boxplots of the best Objective Function values obtained by each algorithm within 30 runs is presented in Fig. 1(a), (b) and (c), for the small, medium and large data sets, respectively.



(a) Small Data Set



(b) Medium Data Set



(c) Large Data Set

Fig. 1. Objective function boxplots

One can notice that all procedures considered are able to achieve similar results with low variability in the small test set. Some performance difference can be seen when the medium data set is tackled. Some of the Genetic Algorithms and Ant Colony Optimization versions present a high variability. All FSS versions considered in both of these two instances present very similar performance. Tests performed with the large data set cause more variability in the results obtained by the differences in the search mechanisms.

The approaches presenting the lowest mean output in each data set within each metaheuristic family, *i.e.* FSS, GA and ACO, will be selected in order to be compared. Hence, the chosen versions, due to the achievement of the lowest Objective Function values in all the data sets, were: FSS-NPSS-SAR, GA-RRIP and ACO-ACS.

4.3 Second Comparison: Best Results and PSO

To determine which technique has the best performance, a non-parametric statistical significance test, namely the Wilcoxon Signed Rank test [23], was performed for the evaluation of the best versions provided previously. The Particle Swarm Optimization (PSO) was included in this phase. Either Objective Function values and number of iterations until convergence (IUC) were used as comparison criteria. IUC is the last iteration in which an improvement of more than 10^{-4} on the Objective Function was reached by the algorithm.

In the statistical test employed, if the resulting p-value is less than the significance level of 0.05, we are able to conclude that there is a difference in the outputs of the algorithms considered. Therefore, the one presenting minimum mean value in each criterion will be selected as best. If p-value is not higher than the significance level, no performance difference can be concluded between the two compared approaches. Table 1 summarizes the results of the comparison carried out. The names of the best performance algorithm on specifics data sets and criteria are written. Only the main identifier of each technique is used (FSS, GA, ACO and PSO).

Table 1. General comparison

Algorithms	Objective function			IUC		
	Small	Medium	Large	Small	Medium	Large
FSS - to - PSO	PSO	PSO	Indifferent	Indifferent	FSS	Indifferent
FSS - to - GA	GA	GA	GA	Indifferent	FSS	FSS
FSS - to - ACO	Indifferent	Indifferent	ACO	ACO	ACO	ACO
PSO - to - GA	GA	GA	GA	Indifferent	Indifferent	PSO
PSO - to - ACO	Indifferent	PSO	ACO	ACO	ACO	ACO
GA - to - ACO	Indifferent	GA	Indifferent	ACO	ACO	ACO

5 Conclusion

In this work, a performance comparison was carried out in order to evaluate which version of Fish School Search algorithm is most suitable for tackling the Assembly Line Balancing Problem. Moreover, a comparison with other metaheuristic procedures was performed.

Two criteria were used as evaluation factors and a nonparametric statistical test was employed in order to reach reliable conclusions regarding the observed performance differences among the techniques studied.

Regarding specifically FSS versions, FSS-NPSS-SAR, a version including elitist behavior as well as boosted exploratory abilities observed when compared to the original version of the algorithm, produced the best results in all three problem instances used for evaluation. However, FSS-NPSS-SAR was not able to reach the best results as those obtained by GA, ACO and PSO in some of the tests performed. When compared to PSO, FSS-NPSS-SAR was able to reach

equivalent outputs in the large instance, but a faster convergence in the medium instance indicates the algorithm gets trapped in local optima. The same can be observed in the comparison with GA. It is noteworthy that in a combined perspective (algorithms, outputs and test problems), FSS matched six and outperformed three out of eighteen comparisons. FSS converged faster but was not able to reach equivalent Objective Function results. Another conclusion from tests performed is that GA-RRIP was the best performer being able to achieve better results for the Objective Function than all the other techniques in at least one of the data sets.

As future works, this comparison should be extended to more metaheuristic approaches in a wider set of problems as well as problem versions in order to provide a reliable benchmark data base for the Assembly Line Balancing Problem research community. Furthermore, an investigation regarding whether the causes of the early convergence of FSS-NPSS-SAR are related to assumptions tried out in this paper, such as α decay mode, should be performed.

Acknowledgments. The authors thank to CAPES (Coordination for the Improvement of Higher-Education Personnel), Brazil, for the partial financial support for this paper.

References

1. Boysen, N., Fliedner, M., Scholl, A.: Assembly line balancing: Which model to use when? Int. J. Prod. Econ. **111**(2), 509–528 (2008)
2. Becker, C., Scholl, A.: A survey on problems and methods in generalized assembly line balancing. Eur. J. Oper. Res. **168**(3), 694–715 (2006)
3. Pitakaso, R.: Differential evolution algorithm for simple assembly line balancing type 1 (SALBP-1). J. Ind. Prod. Eng. **32**(2), 104–114 (2015)
4. Sikora, C.G.S., Lopes, T.C., Silva, H., Magat, L.: Genetic algorithm for type-2 assembly line balancing. In: 2nd Latin American Congress on Computational Intelligence (LA-CCI). vol. 41, pp. 1–6 (2015)
5. Baykasoglu, A., Ozbakir, L.: Discovering task assignment rules for assembly line balancing via genetic programming. Int. J. Adv. Manuf. Technol. **76**(1–4), 417–434 (2014)
6. Dou, J., Li, J., Su, C.: A novel feasible task sequence-oriented discrete particle swarm algorithm for simple assembly line balancing problem of type 1. Int. J. Adv. Manuf. Technol. **69**(9–12), 2445–2457 (2013)
7. Zheng, Q.X., Li, Y.X., Li, M., Tang, Q.H.: An improved ant colony optimization for large-scale simple assembly line balancing problem of type-1. Appl. Mech. Mater. **159**, 51–55 (2012)
8. Bastos Filho, C.J., de Lima Neto, F.B., Lins, A.J., Nascimento, A.I., Lima, M.P.: A novel search algorithm based on fish school behavior. In: 2008 IEEE International Conference on Systems, Man and Cybernetics, SMC 2008, pp. 2646–2651. IEEE (2008)
9. Albuquerque, I.M.C., Monteiro, J.B., Neto, F.B.L., Oliveira, A.M.: Solving assembly line balancing problems with fish school search algorithm. In: IEEE-Symposium Series on Computational Intelligence (2016)

10. Monteiro, J.B., Albuquerque, I.M.C., Neto, F.B.L., Ferreira, F.V.S.: Optimizing multi-plateau functions with FSS-SAR (Stagnation Avoidance Routine). In: IEEE-Symposium Series on Computational Intelligence (2016)
11. Monteiro, J.B., Albuquerque, I.M.C., Neto, F.B.L., Ferreira, F.V.S.: Improved search mechanisms for the fish school search algorithm. In: 16th International Conference on Intelligent Systems Design and Applications (2016)
12. Kumar, D.M., et al.: Assembly line balancing: a review of developments and trends in approach to industrial application. Glob. J. Res. Eng. **13**(2), 1–23 (2013)
13. Bautista, J., Mateo, M., Ferrer, R., Pereira, J., Companys, R.: The assembly line balancing problem solved by hybrid heuristic procedures and driven exploration. POMS, Sevilla (2000)
14. Toksarı, M.D., İşleyen, S.K., Güner, E., Baykoç, Ö.F.: Simple and u-type assembly line balancing problems with a learning effect. Appl. Math. Model. **32**(12), 2954–2961 (2008)
15. Scholl, A., Scholl, A.: Balancing and sequencing of assembly lines. Physica-Verlag Heidelberg (1999)
16. Hamta, N., Fatemi Ghomi, S.M.T., Jolai, F., Akbarpour Shirazi, M.: A hybrid PSO algorithm for a multi-objective assembly line balancing problem with flexible operation times, sequence-dependent setup times and learning effect. Int. J. Prod. Econ. **141**(1), 99–111 (2013)
17. Engelbrecht, A.P.: Computational Intelligence: An Introduction. Wiley, New York (2007)
18. Breginski, R., Cleto, M., Junior, J.S.: Assembly line balancing using eight heuristics. In: 22nd International Conference on Production Research (2013)
19. Dorigo, M., Birattari, M., Stutzle, T.: Ant colony optimization. IEEE Comput. Intell. Mag. **1**(4), 28–39 (2006)
20. Ataie-Ashtiani, B., Ketabchi, H.: Elitist continuous ant colony optimization algorithm for optimal management of coastal aquifers. Water Resour. Manage. **25**(1), 165–190 (2011)
21. Bautista, J., Pereira, J.: Ant algorithms for a time and space constrained assembly line balancing problem. Eur. J. Oper. Res. **177**(3), 2016–2032 (2007)
22. Clerc, M., Kennedy, J.: The particle swarm-explosion, stability, and convergence in a multidimensional complex space. IEEE Trans. Evol. Comput. **6**(1), 58–73 (2002)
23. Richardson, A.: Nonparametric statistics for non-statisticians: A step-by-step approach by Gregory W. Corder, Dale I. Foreman. Int. Stat. Rev. **78**(3), 451–452 (2010)

Improved Search Mechanisms for the Fish School Search Algorithm

João Batista Monteiro Filho, Isabela Maria Carneiro Albuquerque(✉), Fernando Buarque Lima Neto, and Filipe Vieira Silva Ferreira

Department of Computer Engineering,
Polytechnical School of Pernambuco, Recife, Brazil
{jbmf,imca,fbln,fvsf}@ecomp.poli.br

Abstract. In this work we introduce two new mechanisms for the Fish School Search algorithm in order to improve the search ability of its original and niching versions. Two modifications in the usual operators are proposed aiming to increase weight parameters reliability and also to include elitist behavior. Five benchmark optimization problems were employed to evaluate the effectiveness of the modifications proposed. We analyze the convergence curves and also the minimum mean fitness obtained by each version. The results show that the proposed mechanisms improved the convergence of the niching version of the Fish School Search algorithm.

Keywords: Fish school search · Metaheuristics · Swarm intelligence · Optimization

1 Introduction

Several real-world optimization problems involve a great number of variables and complex search spaces. In some cases, those problems are characterized as NP-Hard and classical methods are not able to tackle them in practical time. In this context, non-exact approaches, as metaheuristics, turn out to be efficient and effective methods to solve this sort of problems due to the fact they are able to find at least sub-optimal solutions in an acceptable amount of time [1]. In literature, it is possible to find a great variety of metaheuristics approaches aiming to solve problems with different features, such as combinatorial and continuous optimization. Many of those algorithms are known to be bio-inspired, that is, their operators simulate behaviors found in nature. Genetic Algorithms [2], Particle Swarm Optimization [3] and Ant Colony Optimization [4] are examples of very often applied bio-inspired metaheuristics.

The Fish School Search (FSS) algorithm, developed in 2008 [5], arouse as a bio-inspired metaheuristic that incorporates a new mechanism in the search process: the capability of automatically to adapt the search entities to explore or exploit the search space. Along the years, researchers proposed many modifications in the original version of FSS, referred as FSS-Vanilla, in order to tackle

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_36

different issues such as multi-objective optimization [6], multi-modal optimization [7] and binary search [8]. Among those, it is important to highlight the weight-based FSS (wFSS) [9], a niching and multi-modal version of FSS able to tackle some issues of the original version, and the FSS-SAR [10], proposed in order to increase FSS ability to converge to good results in very smooth search spaces and also applied to solve real-world problems [11].

Despite the fact many aspects of FSS-Vanilla have been tackled during recent years, there still are two important issues to consider in order to improve FSS algorithm's performance. The first one is related with the normalization process used in the Feeding operator. In the Vanilla version there are cases in which the weight of a fish do not reflect how well this fish performed during the search. In this work we propose some modifications in the Feeding operator in order to guarantee that the heaviest fishes performed better during the search. This version is named FSS-Normalized Feeding (FSS-NF). We also incorporate this change in the wFSS algorithm, creating wFSS-NF. The second issue is related to a fish that reaches promising areas during the search process. Due to the fact FSS is based on improvements in fitness values, it is harder for this fish to improve its fitness and it is not able to influence the school in the Collective-instinctive movement anymore. In order to overcome that, we propose a version with a modified Collective-instinctive movement operator which includes a mechanism to take those fishes into account. This new version is named FSS-Not Penalizing Static Success (FSS-NPSS). Details of these modifications are further described in the present work.

The remaining of this article is organized as follows: Sect. 2 presents an overview of FSS-Vanilla and wFSS. In Sect. 3, we describe the variations of FSS proposed in this work. Section 4 describes the tests performed and results obtained from the comparison between variations of the algorithm. Section 5 provides conclusions and future works.

2 Fish School Search Algorithm

2.1 FSS-Vanilla

The Fish School Search simulates the behavior of a fish school while looking for food. In this algorithm, each fish moves independently and then the school moves as a group, contracting and expanding. A fish is represented by a n-dimensional position vector that is a solution for the optimization problem. In addition, each fish has a weight value, a variable that is associated with the performance of a fish during the search.

Here, we provide an overview of the main mechanisms used by this search procedure. The interested reader should refer to the complete and detailed description of the algorithm search operators available in the work of Bastos-Filho and Lima-Neto et al. [5]. FSS operators are: (i) **Individual movement:** In this operator every fish performs a random local search. A fish moves if the new position has a better fitness when compared to the original position. (ii) **Feeding:** Feeding in FSS means to update the weight parameter of every fish in

the school, indicating accumulated success within the search. (iii) **Collective-instinctive movement:** This operator has as goal moving the school to the best directions obtained in the local search. The mean displacement vector of the school, referred as vector **I**, is obtained utilizing the fitness improvement as weight factors. (iv) **Collective-volitive movement:** This operator intends to modify the school exploration/exploitation ability during the search. First, the algorithm calculates the school barycenter **B** based on each fish position ($\mathbf{x}_i$) and weight (W_i). Then, in case the school weight $\sum_{i=1}^{N} W_i$ has increased, all fishes move towards the barycenter. Otherwise, the school will spread away from the barycenter.

2.2 Weight-Based FSS

First introduced in the work of Lima-Neto et al. [9], wFSS is a weight-based niching version of FSS which intends to find multiple solutions. The niching ability is supported by an operator called Link Formation. This operator creates sub-schools by linking fishes to leaders that guide the search.

Furthermore, some modifications in the original operators of FSS need to be done. In the Collective-instinctive movement, every fish has its own vector **I**, which is calculated taking into account the fish and its leader positions and fitness increment. In the Collective-volitive movement, for every fish a barycenter is calculated considering its position and weight and its leader position and weight.

3 New Search Mechanisms Proposed

3.1 Normalized Feeding Version

In FSS-Vanilla and wFSS the Feeding operator is responsible for mapping fitness variation values in the interval $[1; W_{scale}]$, where W_{scale} is an user-defined parameter. As the normalizing scale of weights increments varies along the iterations, the cumulative process of computation for this parameter may lead to inconsistent results, once there will be cases where a fish with a greater fitness is represented by a smaller weight. In order to deal with this, we propose the FSS-NF version that incorporates the fitness normalization in the Feeding operator presented in Eq. 1 in order to guarantee that the heaviest fish is the one which presented the best result during the search process.

$$W_i(t) = 1 + (W_{scale} - 1)\left(\frac{f_i(t) - f_{min}(t)}{f_{max}(t) - f_{min}(t)}\right), \tag{1}$$

$f_i(t)$ represents the fitness of a fish i. $f_{min}(t)$ and $f_{max}(t)$ represent the minimum and maximum fitness values found during the entire search process, then, they have to be updated in every iteration.

Considering that after the Feeding operator the weight of each fish is a reliable representation of how good is the performance of it during the search, we also

propose a modification in the Collective-instinctive movement operator. Instead of using the fitness variation to weight the displacements, this new operator considers weights variation values. Thus, for a school with N fishes, vector $\mathbf{I}$ is computed as follows:

$$\mathbf{I} = \frac{\sum_{i=1}^{N} \Delta \mathbf{x}_i \Delta W_i}{\sum_{i=1}^{N} \Delta W_i}. \tag{2}$$

Due to the fact wFSS niching procedure is very related to fishes weights, we also include this modifications in the wFSS original version, originating the wFSS-NF version.

3.2 Not Penalizing Static Success Version

In FSS-Vanilla, a fish that does not move in the Individual movement is not taken into account in the computation of vector $\mathbf{I}$ during the Collective-instinctive movement. Hence, even in the case that a fish is in a good position, it will not attract other fishes to its position. This feature can be understood as a lack of elitism. In some cases, this feature retards the algorithm convergence and FSS-NPSS intends to deal with this.

The Collective-instinctive movement proposed for the NF version uses information regarding position and weight variations and fishes which did not move during the Individual movement have these variations equal to 0. Thus, in the FSS-NPSS these fishes will have their displacement $\Delta \mathbf{x}_i$ and ΔW_i replaced by:

$$\Delta \mathbf{x}_i^* = step_{vol} \frac{\Delta \mathbf{x}_i^t}{step_{ind}^t}, \tag{3}$$

and:

$$\Delta W_i^* = max(\Delta W_i) \frac{W_i - 1}{W_{scale} - 1}, \tag{4}$$

where $\Delta \mathbf{x}_i^*$ and ΔW_i^* are fish i "fake" displacement and weight increase to be used within the Collective-instinctive movement. $step_{vol}$ is the current step (parameter indicating the maximum displacement that a fish can perform) of the Collective-volitive movement operator. $\Delta \mathbf{x}_i^t$ is the last displacement of fish i which occurred in iteration t and $step_{ind}^t$ is the step of the Individual movement operator in iteration t.

It is important to highlight that the "fake" displacement and weight are used just in the case a fish did not move during the Individual movement. By doing that, the algorithm takes into account all fishes in the school during the Collective-instinctive movement and then those which are already in good positions are able to contribute, introducing elitism. In addition, the Feeding operator applied in FSS-NPSS is the same as the one proposed for FSS-NF, represented by Eq. 1.

4 Tests and Results

4.1 Experimental Setup

Several tests were performed in order to analyze the performance of each version of FSS. All variations described in this work were tested against five optimization benchmark problems. Furthermore, we consider the FSS-SAR and wFSS-SAR and also included the SAR routine [10] in the versions proposed, creating FSS-NF-SAR, FSS-NPSS-SAR and wFSS-NF-SAR. We selected problems with different features in order to test the algorithms under different conditions. The benchmark functions utilized and the respective search spaces are: 1. Ackley, $[-32.768; 32.768]^{30}$; 2. Sum of Different Powers, $[-100; 100]^{30}$; 3. Griewank, $[-600; 600]^{30}$; 4. Rastrigin, $[-5.12; 5.12]^{30}$; 5. Rosenbrock, $[-100; 100]^{30}$. For all problems the fitness value in the global optimum is equal to 0.

One simulation of an algorithm is composed by 30 runs with 10000 iterations. The considered output is the best fitness value achieved in all iterations for the 30 runs. Our methodology for determining the values for W_{scale}, $Step_{ind}$ and $Step_{vol}$ parameters is inspired on the Factorial Planning design of experiments technique [12]. We selected to test each one of the three factors (parameters) with a high and low level, which gives $2^3 = 8$ different combinations of values, specified in Table 1. For the W_{scale} parameter, values in bold are used in the weight-based versions. Each pair version-objective function was tested with all eight different parameter sets.

Table 1. High and low level values for each parameter

Parameter	Low	High
W_{scale}	10^4/**100**	10^5/**500**
$Step_{ind}$	1%	10%
$Step_{vol}$	0.1%	0.1%

In order to simplify the results we specify for all functions the versions that presented better results with the values corresponding to the high level of each parameter. The versions omitted presented better results using the low values. *Ackley*: (i) W_{scale}: FSS-NPSS, wFSS-NF and wFSS-NF-SAR; (ii) $Step_{ind}$: FSS-NF, FSS-NF-SAR, FSS-NPSS, FSS-NPSS-SAR, wFSS, wFSS-SAR, wFSS-NF and wFSS-NF-SAR; (iii) $Step_{vol}$: FSS-Vanilla, FSS-SAR and FSS-NPSS. *Sum of Different Powers*: (i) W_{scale}: FSS-NF-SAR and FSS-NPSS-SAR; (ii) $Step_{ind}$: FSS-Vanilla and wFSS-NF; (iii) $Step_{vol}$: FSS-Vanilla and wFSS. *Griewank*: (i) W_{scale}: FSS-Vanilla, FSS-NF, FSS-NPSS, FSS NPSS-SAR, wFSS, wFSS-NF and wFSS-NF-SAR; (ii) $Step_{ind}$: wFSS-SAR, wFSS-NF and wFSS NF-SAR; (iii) $Step_{vol}$: FSS-NPSS and FSS-NPSS-SAR. *Rastrigin*: (i) W_{scale}: FSS-Vanilla, FSS-SAR, FSS-NF, FSS-NF-SAR, FSS-NPSS-SAR, wFSS and wFSS-SAR; (ii) $Step_{ind}$: FSS-NF, FSS-NF-SAR, FSS-NPSS, FSS-NPSS-SAR,

wFSS, wFSS-SAR and wFSS-NF. (iii) $Step_{vol}$: FSS-Vanilla, FSS-SAR and wFSS-NF-SAR. *Rosenbrock*: (i) W_{scale}: FSS-Vanilla and FSS-NPSS; (ii) $Step_{ind}$: FSS-SAR; (iii) $Step_{vol}$: FSS-SAR. In addition, for the SAR versions, we used $\alpha = 0.8e^{-0.007*t}$, where t is the current iteration number, as in the original work [10].

4.2 Results

In this section we present the results obtained with the best parameter set for each pair version-objective function. In order to understand the behavior of each version we plot the convergence curves obtained considering the mean best fitness of the 30 runs for all functions. In addition, in Tables 2 and 3 we present the mean minimum fitness obtained by each version in each problem.

Table 2. Results obtained with FSS versions

Function	FSS	FSS-SAR	FSS-NF	FSS-NF-SAR	FSS-NPSS	FSS-NPSS-SAR
Ackley	0.0155	0.0085	1.2701	1.2475	18.5662	1.2458
Diff. Powers	0.1671	0.0002	0.0413	0.0413	0.0408	0.0386
Griewank	0.0119	0.0111	0.0771	0.0778	0.1145	0.0484
Rastrigin	60.6568	46.6265	88.0146	96.8236	273.0029	97.1778
Rosenbrock	39.2723	515.9823	29.47	28.87	35.33	28.44

Table 3. Results obtained with wFSS versions

Function	wFSS	wFSS-SAR	wFSS-NF	wFSS-NF-SAR
Ackley	1.6743	0.0000	0.0000	0.0000
Diff. Powers	0.0133	0.0261	0.0000	0.0000
Griewank	0.0832	0.0000	0.0000	0.0000
Rastrigin	110.3514	0.0000	0.0000	0.0000
Rosenbrock	28.9279	34.2009	28.5988	28.7549

Ackley: Figure 1 shows the convergence curves obtained by FSS and wFSS versions, respectively. Considering FSS versions, it is possible to notice that both FSS-Vanilla and FSS-SAR converged faster than the others. In addition, from Table 2, we observe that both versions achieved better results. Thus, we conclude that, for this function, the proposed modifications are not effective, as they do not improve convergence nor fitness results. Furthermore, it is important to highlight that the FSS-NPSS got stuck in a local optimum. We attribute this fact to the use of elitism in this version. For the weight-based algorithms, we observe that the proposed mechanisms were able to improve original wFSS convergence and fitness results. It can be explained by the fact that the weight-based versions are more dependent of weight values, thus the NF is more effective for them. Comparing wFSS-NF with wFSS-SAR it is possible to notice that the former improved convergence speed.

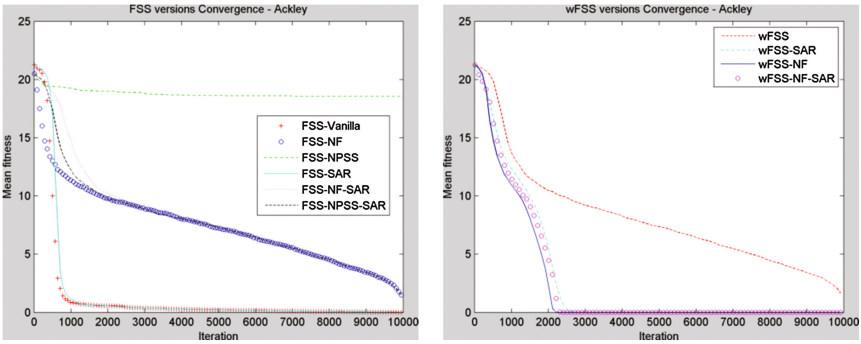


Fig. 1. Convergence curves for FSS and wFSS versions in Ackley function

Sum of Different Powers: Results for convergence of the Sum of Different Powers function are presented in Fig. 2. For the non weight-based versions, it is possible to notice that the FSS-NF and FSS-NPSS improved convergence and fitness values in comparison with FSS-Vanilla, though the same can not be observed for FSS-SAR, which is the version that presented faster convergence and smaller minimum mean fitness value. For the weight-based versions, we observe that wFSS-NF-SAR improved wFSS and wFSS-SAR convergence speed and both wFSS-NF and wFSS-NF-SAR converged to better fitness values. Hence, it is possible to conclude that the NF was effective for the case of a bowl-shaped function.

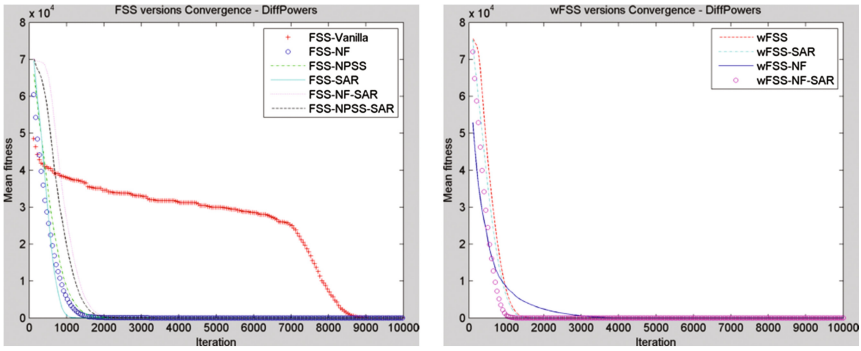


Fig. 2. Convergence curves for FSS and wFSS versions in sum of different powers function

Griewank: Results obtained for the Griewank function can be seen in Fig. 3. For the non weight-based versions, it is possible to notice that FSS-Vanilla and FSS-SAR presented similar results and both converged faster than the NF and NPSS versions. For the weight-based versions, we observe that the proposed mechanism to improve the Feeding operator is effective for multi-modal functions, such as

Griewank. wFSS-NF converged faster than all other versions and was able to achieve the optimum value for minimum mean fitness. wFSS-SAR and wFSS-NF-SAR also achieved the optimum value for minimum mean fitness, although wFSS-NF-SAR converged slower due to its greater exploratory capability.

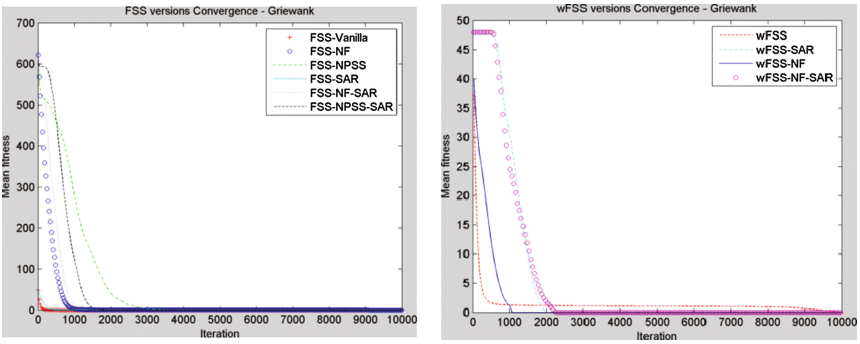


Fig. 3. Convergence curves for FSS and wFSS versions in Griewank function

Rastrigin: Results obtained for the Rastrigin function are shown in Fig. 4. For the non weight-based versions, we observe that both mechanisms proposed in this work achieved worse results for convergence and minimum mean fitness in comparison with FSS-Vanilla and FSS-SAR. Furthermore, it is possible to notice that the FSS-NPSS version got stuck in a local optimum, the same behavior presented by this algorithm for the Ackley function, which is also multi-modal with a non-smooth valley. For the weight-based versions, we observe that wFSS-NF and wFSS-NF-SAR were able to converge faster than wFSS and wFSS-SAR. Regarding the minimum mean fitness, all versions but the original were able to achieve the optimum value.

Rosenbrock: Results obtained for the Rosenbrock function are shown in Fig. 5. Considering FSS versions, NF and NPSS mechanisms did not show improvements in the convergence in comparison with FSS-Vanilla. For the weight-based versions, it is possible to notice that wFSS-NF-SAR converged faster than all other versions and both wFSS-NF and wFSS-NF-SAR achieved minimum mean fitness close to wFSS. Thus, the inclusion of the NF mechanism improved wFSS convergence speed and allowed wFSS-SAR to converge to good results.

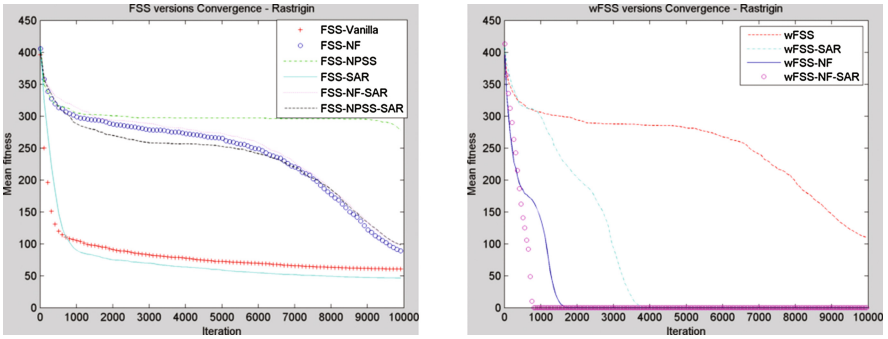


Fig. 4. Convergence curves for FSS and wFSS versions in Rastrigin function

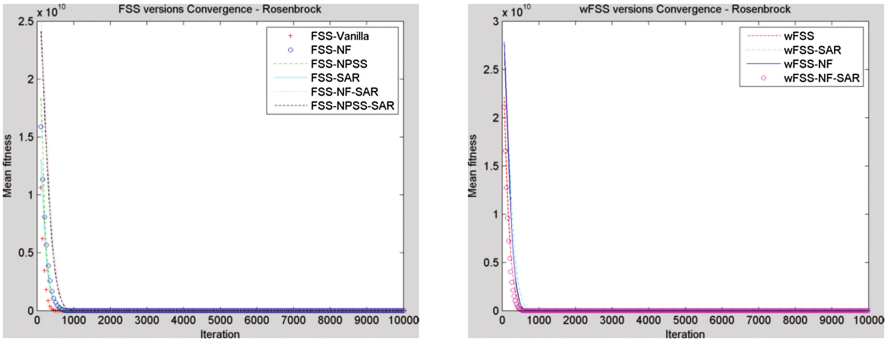


Fig. 5. Convergence curves for FSS and wFSS versions in Rosenbrock function

5 Conclusion

In this work, we proposed two new search mechanisms for the Fish School Search algorithm. The first one intended to modify the Feeding operator in order to guarantee that a greater weight always represents a better solution. This version is called Normalized Feeding. The second mechanism intends to increase the influence in the Collective instinctive movement of fishes which already reached good positions in the search space. Versions that incorporate this mechanism are called Not Penalizing Static Success. We included both modifications in the original version of FSS and also in the niching version, named wFSS.

In order to evaluate all algorithms proposed, we selected 5 benchmark functions and tested each with 8 different parameters set. The best set for each pair version-function was considered for convergence analysis. Regarding the weight-based versions, the mechanisms proposed in this work were able to allow the achievement of better results for fitness in all test functions and also for convergence speed. This fact can be associated with the increase in the consistency of fishes weights computation, which now better indicate the most successful fishes. On the other hand, for the non weight-based versions, we can conclude

that the new mechanisms improved the results just for the Sum of Different Powers function. For the other test problems considered, the new versions decrease convergence speed, as we expect, due to the fact they have more exploratory capability, and were not able to achieve better results for fitness.

For future works, we suggest the application of all versions to solve even more complex test instances and also real-world problems in order to verify whether the behaviors observed in this work remain. In addition, to evaluate the effects caused by different decay modes of the worsening allowance variable α.

References

1. Glover, F.W., Kochenberger, G.A.: Handbook of Metaheuristics, vol. 57. Springer, Heidelberg (2006)
2. Holland, J.H.: Genetic algorithms. Sci. Am. **267**(1), 66–72 (1992)
3. Eberhart, R.C., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, New York, vol. 1, pp. 39–43 (1995)
4. Dorigo, M., Maniezzo, V., Colorni, A.: Ant system: optimization by a colony of cooperating agents. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **26**(1), 29–41 (1996)
5. Filho, C.J.A.B., De Lima Neto, F.B., Lins, A.J.C.C., Nascimento, A.I.S., Lima, M.P.: A novel search algorithm based on fish school behavior. In: Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, pp. 2646–2651 (2008)
6. Bastos-Filho, C.J.A., Guimarães, A.C.S.: Multi-objective fish school search. Int. J. Swarm Intell. Res. **6**(1), 23–40 (2015)
7. Madeiro, S.S., de Lima-Neto, F.B., Bastos-Filho, C.J.A., Nascimento Figueiredo, E.M.: Density as the segregation mechanism in fish school search for multimodal optimization problems. In: Tan, Y., Shi, Y., Chai, Y., Wang, G. (eds.) ICSI 2011. LNCS, vol. 6729, pp. 563–572. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21524-7_69
8. Sargo, J.A.G., Vieira, S.M., Sousa, J.M.C., Bastos-Filho, C.J.A: Binary fish school Search applied to feature selection: Application to ICU readmissions. In: IEEE International Conference on Fuzzy Systems, pp. 1366–1373 (2014)
9. De Lima-Neto, F.B., Pereira, G., de Lacerda, M.: Weight based fish school search. In: 2014 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 270–277. IEEE (2014)
10. Monteiro, J.B., Albuquerque, I.M.C., De Lima-Neto, F.B., Ferreira, F.V.S.: Optimizing multi-plateau functions with FSS-SAR (stagnation avoidance routine). In: IEEE-Symposium Series on Computational Intelligence (2016)
11. Albuquerque, I.M.C., Monteiro, J.B., De Lima-Neto, F.B., Oliveira, A.M.: Solving assembly line balancing problems with fish school search algorithm. In: IEEE-Symposium Series on Computational Intelligence (2016)
12. Hicks, C.R.: Fundamental Concepts in the Design of Experiments. Holt, Rinehart and Winston, New York (1963)

A Survey on Outlier Detection in the Context of Stream Mining: Review of Existing Approaches and Recommadations

Imen Souiden[1(✉)], Zaki Brahmi[2,3], and Hajer Toumi[3]

[1] Higher Institute of Computer Science and Management, Kairouan, Tunisia
imen.sui@gmail.com
[2] Higher Institute of Computer Science and Communication Techniques, Hammam Sousse, Tunisia
[3] RIADI-GDL Lab, Manouba University, Manouba, Tunisia

Abstract. Generally, extracting only expected knowledge from data is not sufficient since unexpected ones can hide useful information concerning the data behavior. These information can be further used to optimize the current state. This has lead to the outlier detection. It refers to the data mining task that aims to find abnormal points or sequence of data hidden in the dataset. In fact, due to the emergence of new technologies, applications often generate and consume data in form of streams. This data differs from the static one. Therefore, traditional techniques cannot be used. Hence, convenient ones suitable to the data stream nature must be applied. In this paper, we will review different techniques of outlier detection in the data streams. In addition, we shall describe different approaches based on these techniques in order to establish a comparative study based on different criterion. This study aims to help users and facilitates the choice of the appropriate algorithm for a certain context.

Keywords: Data stream mining · Outlier detection · Data stream

1 Introduction

Recently, data has become the new human fortune. It contains information serving to reveal many facts, observations and achieving predictions. Furthermore, with the emergence of new technologies and the advance of hardware, data has witnessed a big revolution. Many new applications have appeared, such as communication and social network [1], etc. They are generating continuous, dynamic and infinite volume of data known as data stream. This data differs from traditional one. It's massive, unbounded, generated with a varying rate, temporally ordered and evolving over time. So, it cannot fit into memory and scanned multiple times as traditional data. In order to elaborate a comprehensive data analysis and mining, neglecting the rare events and the abnormal data is not

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_37

recommended. Although their frequency is ranging from 5% to less than 0.01%, depending on the application, they may threaten the system [2]. So, outlier detection has gained a big attention recently. It aims to detect objects having abnormal behavior according to the normal data. Thus, this can improve the data analysis and avoid bad conclusions. Outlier detection in data stream is an important and challenging task. It's used widely in different application domain, such as security and telecommunication [3]. Applying the traditional techniques to detect outliers is not suitable due the difference between the stream and the static data. So, the best solution is to employ techniques adaptable with their characteristics and demands. A large number of approaches for outlier detection in data streams have been proposed [6,7,23,25,27,35]. Choosing the appropriate technique and approach to be applied in a certain context represents a challenging task. In fact, many requirements must be satisfied and taken into consideration according to the input type and the intended result. In this paper, we aim to introduce a comparative study between a set of approaches based on their used technique (distance, clustering, etc.) in addition to a set of criterion (the effectiveness of the detection, computational cost, etc.). Furthermore, we will propose some recommendations in order to assist researchers and industrials to choose the appropriate approach according to the characteristic of their context.

The paper is structured into 5 sections: Sect. 2 defines the fundamental concepts. Section 3 presents a set of outlier detection approaches in data stream organized by the technique type. Section 4 first describes the comparison parameters used, then illustrates the general comparison and the recommendations. Finally, Sect. 5 concludes the paper with a summary and future work.

2 Fundamental Concepts

This section depicts the definitions of the fundamentals of the outlier detection in data streams.

2.1 Outlier Detection

An outlier may appear in the data due to an abnormal behavior in the data generation process. Therefore, it often contains utile information concerning the abnormal characteristics of the systems and entities influencing the data generation process. So, detecting those unusual characteristics provides useful knowledge specific to the application. The outliers can be classified into three categories: point, contextual and collective [4]. Point outlier is the data point which significantly deviates from the rest of the data set. Contextual outlier is a data point which deviates significantly based on a selected context. Collective outliers are subset of objects collectively deviate significantly from the whole data set, even if the individual data objects may not be outliers. We can define the outlier detection as the data mining task aiming to identify patterns that do not conform to the expected and usual behavior. This task is used extensively in a wide variety of applications, such as fraud and intrusion detection, etc.

2.2 Data Stream Mining

Data stream mining refers to the process of analyzing the data stream and finding useful patterns for decision making. The data streams are voluminous, continuous, unbounded, ordered sequences of data arriving in a high rate and having an evolving distribution. As an example, web searches, sensor data, etc. Extracting knowledge from data streams is considered as an important need because of the voluminous quantity of information it contains. Applying traditional algorithms for data streams showed its limitations due to the features of streaming data. A data stream mining algorithm must ensure he following requirements: scan the data only once in real time. In addition, it must be able to handle the change and the evolution of the data distribution. Further, a limited memory space and time must be taken into account along with the above considerations.

2.3 Outlier Detection in Data Stream

Revealing information from abnormal infrequent data has become an excessive need in many applications generating streaming data. To efficiently detect outliers in data stream, the algorithms must consider the different requirements and constraints provided by the data stream mentioned in the precedent section [5].

3 Study of the Approaches

The following section is devoted to present a set of outlier detection techniques in data stream. In addition, for every technique different approaches are introduced.

3.1 Distance Based Outlier Detection

This technique decides the "outlierness" of an object based on the distances to its neighbors in the dataset [4]. If less than k objects in the input data are located at a distance R from this point, then it is considered as outlier. This definition is the mostly used. [9] uses a range query function to compute the object's neighbors in order to detect outliers in a sliding window. To reduce the cost of the outlier detection provided by [9,10] was proposed. It continuously keeps the number of neighbors of an object for all window slides until its expiration. Later, [11,12] have been proposed to combine the best of the resource consumption of [9,10]. They reduce the number of expensive rang queries by the use of an event based approach to avoid unnecessary calculations. Two additional definitions are introduced: The top n outliers with the highest distance and highest average distance to their respective kNN [8]. In [8], a faster and more general framework for high volume and high dimensional data streams have been proposed. It handles the three types of distance-based outliers. In addition, it gives a better performance in terms of the computational cost by optimizing the search space.

3.2 Density Based Outlier Detection

This technique is based on the estimation of the density around the points. Contrary to normal points, outliers are surrounded by a density dissimilar to their local neighbor. The density is evaluated using an outlier score called LOF [13]. This score represents the ratio of the local density of the point and the local density of its nearest neighbor. Points having the highest LOF are declared as outliers [4]. In [14] an adapted version of the static LOF [13] to the stream setting was proposed. Its performance is equivalent to the performance of [13]. However, it's able to detect the changes in the data behavior. [2] has been proposed to tackle the problem of distinguishing between outliers and new data behavior by the use of a sliding window. A point is declared as outlier if it does not form a dense cluster with their subsequent data points in the window. In addition, it provides a reduction in the computation cost.

3.3 Clustering Based Outlier Detection

Clustering is an unsupervised data mining technique that aims to group similar data objects in a cluster. Outliers either do not belong to any cluster, or situated far away from the center of their nearest cluster, or belong to small or sparse one. Different approaches based on clustering and its variant methods have been proposed to detect outliers. Hierarchical approaches such as Birch and partitioning approaches like K-means, CURE, Clarans, Clara, E-lacarans [15] were used either together or separated. In fact, [15] is the most efficient among all of them [16]. Furthermore, [16] have been introduced to improve E-clarans by reducing the temporal complexity which represents its major drawback by choosing the proper nodes based on the maximum cost instead of the arbitrary selection. In [17], outliers are detected based on k-means and the attribute weighting. [18] consists of an online and offline phases working in parallel. In [19], AnyOut a fast algorithm that solves the anytime outlier detection problem for data streams with variable and constant arrival rate was introduced. Concerning the density based clustring, [20] can distinguish between normal data and outliers by using the concept of micro-clusters. The incoming data are assigned to either an outlier or a potential core micro-cluster according to the distance between it and their centers. To guarantee a better clustering accuracy in the presence of a large quantity of outliers [20] was extended on [21]. However, the high computational cost represents its major drawback. Further, [22] improved [20] by extracting the boundary points in the overlapped micro-clusters. Improving the clustering accuracy will improve the outlier detection accuracy. Other approaches have been proposed using both density and partitioning methods such as [23] which represents an extension of [17]. It differs in the integration of DBSCAN after the data pre-processing phase. Also, [24] which is processed in the same way as [23]. However, it uses the feature extraction method which gives the ability to treat high dimensional data but this contributes to an increase in the computational cost.

3.4 Statistical Based Outlier Detection

This technique uses a data distribution model to declare points as outliers or inliers. The outliers have the lowest probability generated from the global distribution. It can be classified in tow categories: (i) the parametric method which assumes the distribution model priory. (ii) the non-parametric method which does not make any assumption about the data distribution. Therefore, it's adaptable to the data stream. The histogram and kernel density estimation are among the non-parametric one. For instance, in [25] authors employed the histogram technique together with a modified version of Chi-square test to detect outliers in VMWare cloud data centers. This technique is efficient but it's destined for only one dimension data. The Kernel density estimation is used in [26] to approximate the sensor data distribution in order to find malicious nodes that deviates from the normal behavior. In addition, in [27] it was destined for the outlier detection in the power grid. [27] outperform [26] since it uses a window model in addition to the clustering algorithm which improves the estimated bandwidth.

3.5 Frequent Pattern Mining Based Outlier Detection

This unsupervised data mining task aims to find the frequent patterns which model the normal data behavior. They are items occurring together and having a support greater than a minimum threshold [30]. The outliers are the patterns that do not conform to the established normal behavior. In general, an outlier is any data point in the data stream that has short and/or small number of frequent patterns [31]. Some metrics are used to measure the deviation from the normal behavior. [28] uses an FPCOF metric that measures the contractiveness between objects in a sliding window with an arbitary size. [29] uses MFPOF factor that indicates the outliers by reporting the intersection between a data object and the set of maximal frequent patterns. It is devoted to detect outliers in high dimensional time series stream. [30] detect anomalies in network by employing a maximum frequent pattern outlier factor. [31] is proposed in order to detect outliers in offline data streams. It's able to adapt to the concept drift and has good detection effectiveness. Contrary to [28,29] which cannot handle concept drift. In fact, the detection based on maximal frequent pattern [29–31] are more effective in term of the computational cost.

3.6 Classification Based Outlier Detection

The classification is the data mining task that uses a training data to learn a classification model and classify the unseen instances into one of the learned normal or outlier class. In case where there is a lack of training data labels, the model will be learned to fit the majority of data during training. Furthermore, in the one-class unsupervised techniques a boundary is learned around the normal data and points falling outside this boundary are considered as outliers [32]. In most cases, the outlier detection is considered as an intermediate step in the novelty detection problem.

For instance, [33] uses a KNN classifier to determine the decision boundary, the discrete Gini Coefficient to detect novel classes and a graph based approach to detect multiple novel class simultaneously and separate them from others. [34] uses two decision boundaries: a kNN followed by a SVM polynomial kernel classifier. A point is termed an outlier if at least one classifier declare it as outlier. Finally, a q-NSC (neighborhood silhouette coefficient) is used to distinguish novel class from outlier. A recent example used directly for the outlier detection is [35]. It employs an online sparse least squares one class SVM (OS-LS-SVM). For any new data, the distance to the obtained hyper plane can be used to measure it resembles to the training set. This approach has less resource consumption because it used a LS-SVM which consumes less resources than SVM. Generally, all of these approaches can deal well with concept drift in addition to the high dimensional data.

3.7 Angle Based Outlier Detection

Outliers are detected by calculating the angles variations between each target instance and the rest of the data points. An outlier produces a smaller angle variance than the normal points. This idea will reduce the impact of the "curse of dimensionality" on mining high-dimensional data [37]. [36] represents an adopted version of the static ABOD algorithm [37] to the stream setting.

4 Comparison

This section is devoted to compare the different approaches presented in Sect. 3 based on a set of criterion presented in the following. In the end, a set of recommendation will be offered.

4.1 Comparison Parameters

Detection effectiveness (D.E): Represents the algorithm's capability to achieve a good detection. This is manifested by good measurement values such as high detection rate and low false alarm rate or good precision, recall, ROC (Receiver Operating Characteristic) and AUC (Area Under Roc Curve) values. Contrary, a moderate or bad detection are characterized by a medium or bad values.

Time complexity (T.C): Refers to the amount of time taken by the algorithm to run and detect outliers.

Space complexity (S.C): Represents the amount of memory required for the execution of the algorithms and the storage of data.

Adaption to the concept drift (A.C.D): The concept drift refers to the change in the data distribution over time. This contributes to the change in the "outlierness" of the objects. A good outlier detection algorithm can adapt its

model to these changes [38]. The ability of the approach to adapt its model to the different types of drift often depends on the window or the block size or the decay factor value. For instance, a shorter window allows the algorithm to react faster to changes while a larger one may take too long to adapt to concept drift.

Instance management technique (I.M.T): Different instance management techniques are used to monitor the change of concept in data stream such as windowing and weighting [39]. The windowing is used to limit the number of processed data. The choice of the window size affects the ability of the algorithm to react and adapt to the concept drift. Weighting is similar to the windowing (a window corresponds to a binary weighting). The examples which serve to learn the model are weighted over time. Those weights quantify its relative importance compared to others. Recent data will have greater weight compared to the oldest one. Different weighting functions can be used (linear forgetting, exponential, ..).

Data processing approach (D.P.A): According to the order of appearance of data defined by a time stamp associated with each object, three data processing approaches can be defined [39]: (i) agglomerative where the data is processed one after another. (ii) by packet: data are grouped into packets of the same fixed size then they will be processed one after another. (iii) by sliding window: the data is processed in a window with a fixed size.

Arrival rate (A.R): The stream has a rapid arrival rate in addition it can be variable or fixed. If it's variable, then the data points to which the current data point is compared to detect the "outlierness" should be adjusted dynamically based on the available processing time [5]. In addition, if the stream uses a fixed window, then it's more suitable for fixed stream and the adaptation to the speed depends on the window size. Moreover, if the stream is treated by batch and buffer, then it can be used for fixed or varying stream. In this case, if the size is fixed, then the adaptation to the speed depends also on that size.

Anytime outlier detection (A.O.D): It refers to the problem of determining within any period of time whether an object in a data stream is outlier or not [19]. The decision is proportional with the available time. The more time is available, the more reliable the decision should be [12]. So we can say that an approach supports anytime outlier detection if it obeys to this two criterion.

Dimensionality (Dim): The data stream can be of different dimensions. Some algorithms are able to deal with high dimensional data (H) while other can only deal with one, low (L) or moderate (M) dimension.

Outlier rate (O.R): The outlier detection algorithm performance can be affected by the outlier rate. Some algorithms give an acceptable or consistent performance in the presence of a considerable amount of outliers while others are damaged in a remarkable or minor manner.

Nature of the input data (N.I.D): The attributes of the input data can be of different nature such categorical, numeric, mixed (numeric and categorical).

Outliers Type (O.T): An outlier can be **global** if it is an outlier with respect to all the available data points (Example: the current window). However, it's **local** if it is outlier with respect to its nearest neighbors [40].

Output type (Op.T): The output can be a label indicating whether an object is an outlier or not. Or a score depending on its degree of "outlierness" [5].

4.2 General Comparison

Table 1 illustrates the comparison between the best of the approaches presented. Their performances vary according to the satisfied criterion. All of them are effective in terms of the outlier detection. However, only [19] supports the anytime outlier detection. In terms of the computational cost, these approaches have low [8,19,25,30,31] or acceptable [2,16,19,25,35,36] requirements except in [21,24,27,33]. Concerning the input type, only the numeric data is supported in [2,8,16,21,23,24] while [25,31] support categorical data only. Regarding the adaptation to the concept drift, the majority of the algorithms [2,8,19,21,23–25,27,31,33–36], can handle different types of changes in data with respect to the window or the block size or according to the decay factor. However, it's not the case in [16]. Furthermore, the sensitivity to the outlier rate differs from one approach to another. In fact, [21] is destined for the applications having a big number of outliers. Besides, the performance of [8,23] is regular with the increase of the percentage of outliers. Among all of these approaches, only [8,24,30,31,36] are cable to treat the high dimensional data. The arrival rate represents an important factor. The majority of the approaches can handle varying as well fixed data stream's arrival rates [19,21,23–25,27,30,31,36]. On the other side [2,8,16,33,35] deal with fixed arrival rates only. We can conclude the efficiency of the adaptation to the stream speed basing on the block or the window size. For example, [19] prefers large windows for either slow or rapid stream.

Typically, the outlier detection technique gives important information about the approaches characteristics. The statistical, distance, density, clustering based techniques have generally low performance for high dimensional data due to the effect of the curse of dimensionality. This inefficiency is more manifested in the distance based method. Moreover, this technique is not destined for regions with varying density. This problem is overcome by the density based technique. However, even if it's more efficient than the distance based, it suffers from high computational cost. The auto regression technique [4] is used specially in time series streams but it's not mentioned here because of its limitation with dynamic data due to concept drift. The clustering, frequent pattern mining and classification are data mining building blocks. The clustering is the most used one but it's not optimized directly for the outlier detection problem. Yet, recently some approaches are used to improve the outlier detection [21]. The classification techniques are not widely used since the data stream is often unlabeled. However, it's convenient in the presence of some known types of outliers.

Table 1. General comparison

	[2]	[8]	[16]	[21]	[23]	[24]	[19]	[27]	[25]	[30]	[31]	[33]	[35]	[36]
S.C	A	L	-	H	C	H	-	-	-	L	L	H	A	-
T.C	A	L	A	H	A	H	L	C	L	L	L	A	A	A
A.C.D	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
A.R	F	F	F	F-V	F-V	F-V	F-V	F-V	F-V	F-V	F-V	F	F	F-V
A.O.D	No	No	No	No	No	No	Yes	No	No	No	No	No	No	No
D.E	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
Dim	L-M	L-M	L-M-H	L-M	L-M	H	L-M	L-M	1 dim	L-M-H	L-M-H	L-M-H	L-M-H	H
O.R	-	Good	-	Good	Consistent	-	-	Consistent	-	-	-	-	-	-
T.I.M	wn	wn	-	wt	wt	wn	wn	wn	wn	-	wn	-	-	wn
A.S.P	SW	SW	SW	-	P	P	SW	SW	SW	P	Pt	Agge	SW	Agg
N.I.D	N	N	N	Nc	N	N	-	N	C	-	C	-	-	-
OT	l	G-l	G	G	G	G	G	G-l	G	G	G	G	G	G
Op.T	S	L-S	L	L	L	L	S	L	L	S	S	S	S	S

Notes: SW: Sliding window, L: Low, H: High, C: Considerable, A: Acceptable, F: Fixed, V: Varying, wn: Windowing, wt: Weighting, P: Packet, Agg: Agglomerative, S: Score, L:Label, G: Global, l: Local.

4.3 Recommendations

Generally, there is no best approach but there is an appropriate one for a certain context. The choice operation needs a deep understanding about the problem context and the approach characteristics. The outlier detection often depends

Table 2. Recommendation

Problem characteristics	Approach to use	Comments
Result provided at any time and can be improved by more time	[19]	Low time complexity/Problem of scalability and high dimensionality
-Local outliers -Region with varying densities	[2]	Good performance/Problem of high dimensionality
High percentage of outliers in the data and the effectiveness of the outlier detection is the measure of interest	[21]	Sacrifices of the space and time to give more accurate result/Not efficient in high dimensional data
The percentage of outlier is not prefixed and can increase	[23–25]	[23] is suitable for 1 dimension problem
High dimensional data	[24,30,31,36]	[24,36] are destined specifically for high dimensional data
Offline streams such as reports based on web log streams etc.	[31]	The effectiveness of the outlier detection is more important than the timing of the result
Detection of novel classes	[33]	The aim is to discover new classes
High volume and high speed stream	[8]	If a fixed threshold distance is not available or changing over time it's appropriate to use the KNN distance outliers

on the characteristics of the input and the output data. Furthermore, an important factor is the resource requirements due to the limited available capacity. Consequently, the approach must conform to the specification of the problem to treat. In our study, we have chosen a set of criterion to expose the approach performance. This will help readers to match it with the characteristics of their problem. As a result, it will facilitate the choice operation and conserve time. To be more helpful we have provided some recommendations illustrated by Table 2. Furthermore, in order to benefit from the best of each approach, some hybridization can be elaborated. This will improve the results in different terms. For instance, the clustering approaches can be mixed with other techniques such as the distance, density, frequent pattern mining, classification. This will reduce the calculation due to the concept of clusters. But the choice of the clustering algorithm must be convenient. In addition, we can benefit from the advantage of anytime outlier detection by applying this concept in other approaches. Also, different approaches belonging to the same technique can be mixed.

5 Conclusion

Recently data streams have gained a big attention. The outlier detection is one of the challenging areas in data stream mining. Therefore, several approaches have been proposed in order to solve this problem. The choice of the correct algorithm to be applied in a certain context is not that easy. In fact, it depends on different parameters concerning the characteristic of input and the intended results. In this paper, we have discussed different approaches based on several used techniques of outlier detection in data stream. Then, a comparative study is achieved in order to approximate the best algorithms in each method. The comparison is based on different parameters concerning the effectiveness of the outlier detection, the computational cost and other criterion such as the nature, the type of the incoming data in addition to the adaptation to the concept drift, etc. According to this comparison we have established some recommendations in order to be more helpful. This work contains a very interesting approaches of outlier detection in data stream but not all of them. So a further work can be done by adding other approaches and treating other problems. In addition, we are willing to improve the results of our theoretical study by some experimentation to effectively compare the different approaches. Moreover, we are going to choose a suitable algorithm and try to apply it in the context of the detection of the abnormal behaviors in cloud infrastructure in term of CPU, memory.

References

1. Aggarwal, C.C.: Data Mining: The Textbook. Springer, Switzerland (2015)
2. Karimian, S.H., Kelarestaghi, M., Hashemi, S.: I-inclof: improved incremental local outlier detection for data streams. In: 16th CSI International Symposium on Artificial Intelligence and Signal Processing (2012)

3. Beigi, M.S., Ebadollahi, S., Chang, S.F., Verma, D.C.: Anomaly detection in information streams without prior domain knowledge. IBM J. Res. Dev. **55**, 550–560 (2011)
4. Thakkar, P., Vala, J., Prajapati, V.: Survey on outlier detection in data stream. Int. J. Comput. Appl. **136**, 13–16 (2016)
5. Sadik, M.S., Gruenwald, L.: Research issues in outlier detection for data streams. ACM SIGKDD Explor. **15**, 33–40 (2014)
6. Stevanovic, D., Vlajic, N.: Next generation application-layer DDoS defences: applying the concepts of outlier detection in data streams with concept drift. In: 2014 13th International Conference on Machine Learning and Applications (2014)
7. Miller, Z., Deitrick, W., Hu, W.: Anomalous network packet detection using data stream mining. J. Inf. Secur. **2**, 158–168 (2011)
8. Cao, L., Yang, D., Wang, Q., Yu, Y., Wang, J., Rundensteiner, E.A.: Scalable distance-based outlier detection over high-volume data streams. In: 30th International Conference on Data Engineering (2014)
9. Angiulli, F., Fassetti, F.: Distance-based outlier queries in data streams: the novel task and algorithms. Data Min. Knowl. Disc. **20**(2), 290–324 (2010)
10. Yang, D., Rundensteiner, E., Ward, M.: Neighbor-based pattern detection for windows over streaming data. In: EDBT 2009, pp. 529–540 (2009)
11. Kontaki, M., Gounaris, A., Papadopoulos, A.N., Tsichlas, K., Manolopoulos, Y.: Continuous monitoring of distance-based outliers over data streams. In: The 27th International Conference on Data Engineering (ICDE), pp. 135–146 (2011)
12. Kontaki, M., Gounarisn, A., Papadopoulos, A.N., Tsichlas, K., Manolopoulos, Y.: Effcient and flexible algorithms for monitoring distance based outliers over data streams. Inf. Syst. **55**(C), 37–53 (2016)
13. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: identifying density based local outliers. In: 2000 ACM SIGMOD International Conference on Management of Data, vol. 29(2), pp. 93–104 (2000)
14. Pokrajac, D., Lazarevic, A., Latecki, L.J.: Incremental local outlier detection for data streams. In: IEEE Symposium on Computational Intelligence and Data Mining, pp. 504–515 (2007)
15. Christopher, T., Divya, M.T.: A comparative analysis of hierarchical and partitioning clustering algorithms for outlier detection in data streams. Int. J. Adv. Res. Comput. Commun. Eng. (2015)
16. Mathur, N., Tiwari, M., Khandelwal, S.: Increased performance factor for the best clustering algorithm. Int. J. Eng. Tech. Res. **3** (2015)
17. Yogita, T.D.: A framework for outlier detection in evolving data streams by weighting attributes in clustering. In: 2nd International Conference on Communication Computing and Security (2012)
18. Koupaie, H.M., Ibrahim, S., Hosseinkhani, J.: Outlier detection in stream data by clustering method. Int. J. Adv. Comput. Sci. Inf. Technol. **2**, 25–34 (2013)
19. Assent, I., Kranen, P., Baldauf, C., Seidl, T.: AnyOut: anytime outlier detection on streaming data. In: Lee, S., Peng, Z., Zhou, X., Moon, Y.-S., Unland, R., Yoo, J. (eds.) DASFAA 2012. LNCS, vol. 7238, pp. 228–242. Springer, Heidelberg (2012). doi:10.1007/978-3-642-29038-1_18
20. Cao, F., Ester, M., Qian, W., Zhou, A.: Density-based clustering over an evolving data stream with noise. In: 2006 SIAM Conference on Data Mining, pp. 328–339 (2006)
21. Li-xiong, L., Jing, K., Yun-fei, G., Hai, H.: A three-step clustering algorithm over an evolving data stream. In: IEEE International Conference on Intelligent Computing and Intelligent Systems, pp. 160–164 (2009)

22. Kumar, M., Sharma, A.: Mining of data stream using DDenStream clustering algorithm. In: 2013 International Conference in MOOC on Data Engineering, pp. 315–320 (2013)
23. Yogita, T.D.: Unsupervised outlier detection in streaming data using weighted clustering. In: 12th International Conference on Intelligent Systems Design and Applications, pp. 160–164 (2012)
24. Gurav, R.B., Rangdale, S.: Hybrid approach for outlier detection in high dimensional dataset. Int. J. Sci. Res. **3** (2014)
25. Solaimani, M., Iftekhar, M., Khan, L.: Statistical technique for online anomaly detection using spark over heterogeneous data from multi-source VMware performance data. In: IEEE International Conference on Big Data (2014)
26. Kumar Samparthi, V.S., Verma, H.K.: Outlier detection of data in wireless sensor networks using kernel density estimation. Int. J. Comput. Appl. **5** (2010)
27. Uddin, M.S., Kuh, A., Weng, Y., Ili'c, M.: Online bad data detection using kernel density estimation. In: IEEE Power and Energy Sociaty and General Meeting (2015)
28. Tang, X., Li, G., Chen, G.: Fast detecting outliers over online data streams. In: International Conference on Information Engineering and Computer Science, pp. 1–4 (2009)
29. Lin, F., Le, W., Bo, J.: Research on maximal frequent pattern outlier factor for online high dimensional time-series outlier detection. J. Convergence Inf. Technol. **5**, 66–71 (2010)
30. Dominic, D.D., Said, A.M.: Network anomaly detection approach based on frequent pattern mining technique. In: International Conference on Computational Science and Technology (2014)
31. Said, A.M., Dominic, P.D.D., Faye, L.: Data stream outlier detection approach based on frequent pattern mining technique. Int. J. Bus. Inf. Syst. **20**, 55–70 (2015)
32. Zhang, Y., Meratnia, N., Havinga, P.: Outlier detection techniques for wireless sensor networks: a survey. IEEE Commun. Surv. Tutorials **12**, 159–170 (2010)
33. Kale, A., Ingle, M.D.: SVM based feature extraction for novel class detection from streaming data. Wireless Pers. Commun. J. **110**, 1–3 (2015)
34. Masud, M.M., Gao, J., Han, J., Khan, L., Thuraising-ham, B.M.: Classification and adaptive novel class detection of feature-evolving data streams. IEEE Trans. Knowl. Data Eng. **25** (2013)
35. Uddin, M.S., Kuh, A.: Online least-squares one-class support vector machine for outlier detection in power grid data. In: IEEE International Conference on Acoustics Speech and Signal Processing (2016)
36. Ye, H., Kitagawa, H., Xia, J.: Continuous angle-based outlier detection on high-dimensional data streams. In: 19th International Database Engineering and Applications Symposium, pp. 162–167 (2015)
37. Kriegel, H.P., Hubert, M.S., Zimek, A.: Angle based outlier detection in high-dimensional data. In: 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008)
38. Salperwyck, C.: Apprentissage incrémental en-ligne sur flux de données. Ph.D. thesis, University Charles de Gaulle (2012)
39. Marascu, A.: Extraction de motifs séquentiels dans les flux de données. Ph.D. thesis, Université de Nice Sophia Antipolis (2009)
40. Salehi, M., Leckie, C., Bezdek, J., Vaithianathan, T.: Local outlier detection for data streams in sensor networks: revisiting the utility problem invited paper. In: 10th International Conference on Intelligent Sensors, Sensor Networks and Information Processing (2015)

Heuristic for Scheduling Intrees on m Machines with Non-availability Constraints

Khaoula Ben Abdellafou[1(✉)], Hatem Hadda[2], and Ouajdi Korbaa[1]

[1] MARS Laboratory, ISITCom, University of Sousse, Sousse, Tunisia
khawlaabdellafo@gmail.com
[2] Ecole Nationale d'Ingenieurs de Tunis, OASIS,
Universite de Tunis El Manar, 1002 Tunis, Tunisia

Abstract. This paper considers the problem of scheduling n tasks subject to intree-precedence constraints on m identical machines under non-availability constraints. The objective is to minimize the makespan. This problem is known to be NP-hard. We propose and test several heuristic variants based on different selection and dispatching rules.

Keywords: Scheduling · Resumable · Heuristic · Makespan · Intree

1 Introduction

The popularity of the Internet and the availability of powerful computers and high-speed networks are changing the way we use computers today. These technical opportunities have led to the possibility of using geographically distributed and multi-owner resources to solve large-scale problems in science, engineering, and commerce. Recent research on these topics has led to the emergence of a new paradigm known as Grid computing [1].

In this respect, task scheduling problems have attracted great attention from researchers [2,3]. This article, will consider non-preemptive scheduling problem on parallel processors, where each of them can be interrupted only by a non-available period. In literature of machine scheduling with non-availability constraints, the general model is known as the semi-resumable case, where a non finished task must be partially restarted for the portion that has been processed. This model has two particular variants: resumable and non-resumable. A task is said to be resumable if in the case when the task is interrupted by a non-available period, the task can continue without any penalty after the machine is available again. Alternatively, a task is said to be non-resumable if it has to restart, rather than continue [4]. According to the three-field notation introduced by Graham et al. [5], the problem considered in this article is denoted $P|intree, sr-a|C_{max}$. Moreover, we consider that the tasks are partially ordered by precedence constraints as an intree. We recall that a graph G is an intree, if each vertex except the root has exactly one immediate successor. Particularly, if each node has also a single predecessor the graph is said a chain.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_38

Several polynomial-time algorithms have been developed for scheduling problems with graph constraints. In non preemptive scheduling and without non-availability constraints, the problem denoted $P_m|prec|C_{max}$ is easily if the number of machines is illimited [6]. For the case where the processing time of tasks is unitary and the number of machines is limited to 2 ($P_2|p_j = 1, prec|C_{max}$), the problem can be polynomially solved by coffman and Graham [7]. If the tasks of the graph takes the form of an intree ($P_m|intree, p_j = 1|C_{max}$), Hu algorithm can solve the problem.

In non preemptive scheduling on machines subject to non-availability constraints ($P_m|nr - a|C_{max}$), the problem of minimizing the makespan for a set of n independent tasks is known to be NP-hard [8–10]. In [8] Lee considers this last problem knowing that some machines may not be available at time zero and has shown that the makespan obtained by applying the longest processing time (LPT) do not exceed the optimal one by $\frac{3}{2}$. Then in a second work [9], Lee tackles the general problem assuming that there is a machine that is always available. By adopting this hypothesis, Lee has shown that the makespan obtained by LPT algorithm can not exceed the optimal makespan by $\frac{m+1}{2}$.

In the literature, there are many works that have dealt with problems of scheduling independent tasks on processors with non-availability constraints. However, the problem of scheduling intrees with limited machine availability has not been very studied. Among these few works we can cite Garey and Johnson [11] who have shown that the scheduling of intrees on machines subject to a decreasing profile (decreasing number of available machines) is NP-hard and we can cite also Dolev and Warmuth who have shown that the Hu algorithm is optimal in the case of intrees scheduling (for tasks with unitairy processing time) on machines subject to special cases of unavailability profiles.

In [12] the authors propose a new heuristic for the $P_2|intree, p = 1, c = 1, nr-a|C_{max}$ problem where both processors are subject to unitary non-availability constraints. It was shown that the heuristic gives good results for particularly intree graphs. For this same problem, where one processor is always available, we have proposed an optimal algorithm [13,14].

The paper is organised as follows. In Sect. 2, the considered problem is presented. The adopted solution is detailed in Sect. 3. In Sect. 4, the lower bounds is presented. The experimental results are illustrated in the Sect. 5. Finally, we give some concluding remarks.

2 Problem Description

In this paper, we consider the problem of scheduling a set of n tasks with intree precedence constraints on m identical parallel machines. Each machine is subject to at most one non-availability period, and can not perform more that one task at any given time.

Furthermore, the preemption is not allowed and we consider a semi-resumable scenario with a resumable factor ($0 \leq \alpha \leq 1$). That is to say, if p is the part of a task that has already been processed before the non-availability period, then αp needs to be repeated immediately after the machine becomes available again.

The goal is to schedule the n tasks on the m machines so that the precedence constraints are observed and the makespan is minimized. It is assumed that all the execution times of the tasks and unavailability periods are known in advance. The problem is denoted $Pm|intree, sr-a|C_{max}$.

We will use the following notation in the subsequent analysis.

m	Number of machines.
n	Number of tasks.
T_j	Task number j in the intree.
p_j	Processing time of T_j.
S_i	Start time of the non−availability period on machine i.
F_i	End time of the non−availability period on machine i.
α	Resumable factor.
$Path(T_j)$	The sum of the processing times of all the tasks that form the chain linking T_j to the root.
$leave$	A leave in an intree is a task with no predecessors.

In the following we propose a heuristic procedure to solve the described problem. The heuristic combines several selection and assignement rules. Then, we present an experimental study.

3 Adopted Solution

The construction of a complete schedule requires solving two questions. An assignement problem (i.e. to which machine each job will be assigned), and a sequencing problem (i.e. in what order the jobs should appear in each machine). Hereafter, we propose a constructive two-phase heuristic which each time proceeds to the selection of a job then assign it to a machine. We test and compare several selection and assignement rules (See Fig. 1):

HLF: stands for **H**igh **L**evel **F**irst.
LPT: stands for **L**ongest **P**rocessing **T**ime.
SPT: stands for **S**hortest **P**rocessing **T**ime.
LPF: stands for **L**ongest **P**ath **F**irst.
SPF: stands for **S**hortest **P**ath **F**irst.
Leaves: This rule gives the priority to tasks having no predecessors.

We also select two Assignement rules for task allocation. The first one consists in assigning the task (chosen by the selection rule) to the processor in which the execution of the task starts as early as possible. The second rule consists in allocating the chosen task to the processor which minimizes the completion time of the task.

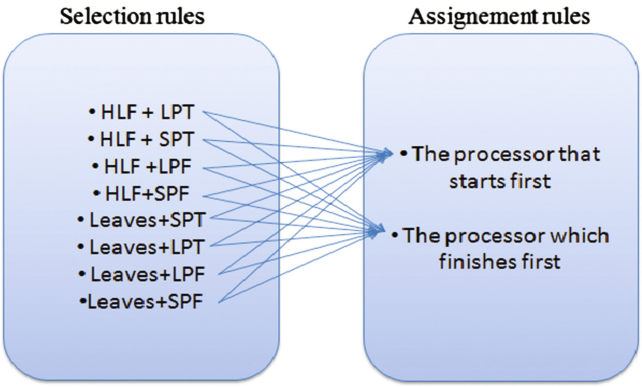


Fig. 1. Assignement and selection rules

3.1 First Selection Rule: HLF+LPT

This rule gives the priority to the tasks of the highest level. If that level contains many tasks we select the one with the longest processing time (see example in Fig. 2).

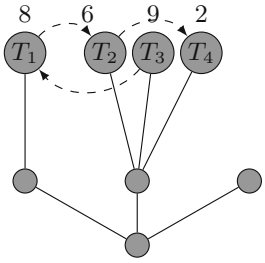


Fig. 2. Task graph

In this instance, there are four tasks in the highest (third) level. Based on the processing times the priority order is $T_3 \rightarrow T_1 \rightarrow T_2 \rightarrow T_4$.

3.2 Second Selection Rule: HLF+SPT

This rule gives the priority to the task of the highest level with the lowest processing time. If that level contains many tasks we select the one with the shortest processing time.

In considering Fig. 2, tasks of level 3 will be selected in the order $T_4 \rightarrow T_2 \rightarrow T_1 \rightarrow T_3$.

3.3 Third Selection Rule: HLF+LPF

This rule sorts the tasks of the highest level according to the decreasing order of the lengths of their paths from the root (see Fig. 3).

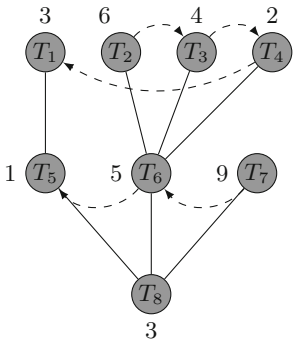


Fig. 3. Task graph

According to $HLF+LPF$ rule, the order of the tasks selection is $T_2 \rightarrow T_3 \rightarrow T_4 \rightarrow T_1 \rightarrow T_7 \rightarrow T_6 \rightarrow T_5 \rightarrow T_8$.

3.4 Fourth Selection Rule: HLF+SPF

This rule sorts the tasks of the highest level in the increasing order of the lengths of their paths from the root.

According to $HLF + SPF$ rule and considering Fig. 3, the order of the tasks selection is: $T_1 \rightarrow T_4 \rightarrow T_3 \rightarrow T_2 \rightarrow T_5 \rightarrow T_6 \rightarrow T_7 \rightarrow T_8$.

3.5 Fifth Selection Rule: Leaves+SPT

The principle of this rule, is to start with the leaves (the tasks that have no predecessors) sorted in the increasing of their processing times (see Fig. 4).

In the instance described by Fig. 4, the priority order by $leaves + SPT$ rule is: $T_7 \rightarrow T_1 \rightarrow T_3 \rightarrow T_5 \rightarrow T_4 \rightarrow T_2 \rightarrow T_6 \rightarrow T_8$.

3.6 Sixth Selection Rule: Leaves+LPT

This selection rule has same principle of the precedent rule but the tasks are sorted in the decreasing order of their processing times.

If we consider the task graph described by Fig. 4, the priority order by $leaves + LPT$ rule is: $T_2 \rightarrow T_4 \rightarrow T_3 \rightarrow T_1 \rightarrow T_5 \rightarrow T_7 \rightarrow T_6 \rightarrow T_8$.

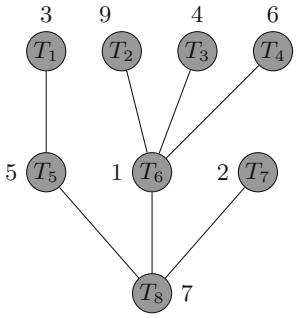


Fig. 4. Task graph

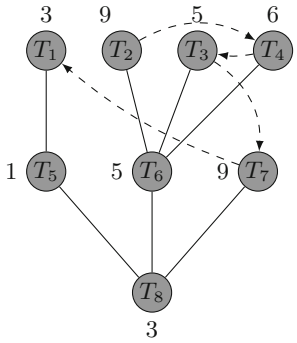


Fig. 5. Task graph

3.7 Seventh Selection Rule: Leaves+LPF

The principle of the seventh rule is to start by the leaves, then sort them in the decreasing order of the path.

According to the precedence graph described by Fig. 5 the priority order of tasks is: $T_2 \rightarrow T_4 \rightarrow T_3 \rightarrow T_7 \rightarrow T_6 \rightarrow T_1 \rightarrow T_5 \rightarrow T_8$.

3.8 Eighth Selection Rule: Leaves+SPF

According to the precedence graph described by Fig. 5 the priority order of tasks computed by "Leaves+SPF rule" is: $T_1 \rightarrow T_5 \rightarrow T_7 \rightarrow T_3 \rightarrow T_4 \rightarrow T_2 \rightarrow T_6 \rightarrow T_8$.

4 Lower Bounds

In order to study the behaviour of the different rules' combinations, we propose to compare them to a lower bound on the optimal makespan. In this section, three lower bounds will be described. The highest will be retained for the comparison with the heuristics' solutions.

For the first lower bound, we ignore both availability and precedence constraints, and we consider the sum of all processing times of all tasks divided by the number of machines (i.e. $\frac{\sum_{i=1}^{n} p_i}{m}$). For the second bound we consider the sum of the processing times of the longest chain in the intree. Finally, for the third bound, we consider the availability constraints and we suppose that preemption is allowed, we then distribute the sum of all processing times (i.e. $\sum_{i=1}^{n} p_i$) over the machines such that their completion time is the same (see Fig. 6). It is clear that the third bound always dominates the first one.

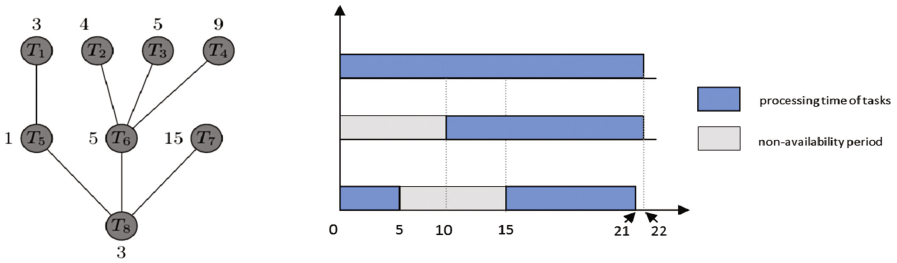


Fig. 6. Lower bound computing

- First bound: $\sum_{i=1}^{N} p_i/m = 45/3 = 15$.
- Second bound: $p_7 + p_8 = 18$.
- Third bound: According to the machine environment described by Fig. 6 the bound is 22.

We then retain the bound 22.

5 Experimental Results

The lower bounds and the heuristic scheduler were implemented using Java over the framework Netbeans IDE 8.0.2, and all tests were carried on an Intel Core i7 CPU running at 2.40 GHz. In order to test the influence of different data structures, we consider two sets of instances.

To compare the heuristic results with the lower bound, we determine the gap between each heuristic solution, and the lower bound by computing $\frac{C_{max}-LB}{LB} * 100$. Then, we derive the average gap over the instances.

5.1 First Test Set

For the first test set, the heuristic was tested on a set of 500 instances varying in size between 10 and 3900 tasks. The number of machines is fixed at 5 and each node (task) in the intree can have at maximum 5 predecessors. All the

processing times and the unavailability periods are randomly generated using a discrete uniform distribution over the interval $[0, 100]$.

The generation of the intree, is done step by step starting by the root. For each node, a randomly number of predecessors is generated between 0 and 5. The number of tasks in the intree is between a predetermined number of levels (lower bound) and an upper bound computed by this formula: $1 + \sum_{k=1}^{levels}(predecessors)^k$. This war we obtain larges intrees.

Table 1. .

Selection rules	Affectation rules					
	Starts first			Finish first		
	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
HLF+LPT	2.10%	2.87%	1.77%	2.10%	2.7558%	1.9466%
HLF+SPT	2.44%	3.07%	2.02%	2.43%	2.98%	1.96%
HLF+LPF	1.84%	2.33%	2.74%	1.84%	2.33%	1.55%
HLF+SPF	1.86%	2.21%	2.73%	2.44%	2.94%	1.95%
Leaves+SPT	17.04%	17.66%	16.85%	17.30%	17.79%	16.60%
Leaves+LPT	17.10%	17.66%	18.97%	16.94%	17.62%	18.53%
Leaves+LPF	16.86%	17.40%	18.82%	16.68%	17.26%	16.09%
Leaves+SPF	17.47%	18.03%	18.81%	17.29%	17.62%	19.59%

5.2 Results

The obtained results for the first test set are given in Table 1. We observe that the heuristics generated schedules that have an average difference of 2.5% from the lower bound for the four first selection rules, and an average difference of 18% for the rest of rules.

For rules that start with the leaves, the number of tasks from which we have to select will be much higher than when using HLF rule. For this reason, more than half of the tasks can be assigned in the first step, then the heuristic doesn't have much chance to provide the schedule. For example, a leave which is in the second level shouldn't be assigned early since it has only one successor which is the root. In this case, it is better to give priority to tasks having many successors, which is ensured by HLF rule. Therefore, the two combinations $HLF + SPF$ and $HLF + LPF$ are the best because they give the priority to the tasks having the longest path.

We also remark that whenever the resumable factor increase, the makespan tends to increase as the supplementary charge (αp) also increases. Naturally this is not always the case because the heuristic may assign the considered task to another machine that does not interrupt the task by an unavailability. For this type of graphs, in most of the cases, it is the third bound which leads to the best lower bound.

5.3 Second Test Set

For this test set, the heuristic was tested on a set of 500 instances varying in size between 40 and 60 tasks. The number of machines is fixed at 5 and each task in the intree can have at maximum 1 predecessor except for the root which can have more than one. Hence the graph is a set of chains with one root. Here too all the processing times and the unavailability periods are randomly generated using a discrete uniform distribution over the interval $[0, 100]$.

Table 2. .

Selection rules	Affectation rules					
	Starts first			Finish first		
	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
HLF+LPT	9.22%	11.16%	13.95%	1.19%	1.10%	1.35%
HLF+SPT	9.70%	11.11%	13.84%	1.36%	1.64%	1.92%
HLF+LPF	9.83%	10.86%	13.38%	0.83%	0.86%	1.03%
HLF+SPF	9.10%	10.10%	13.60%	1.38%	1.64%	2.07%
Leaves+SPT	9.16%	10.16%	14.01%	1.36%	1.64%	1.9232%
Leaves+LPT	9.22%	11.16%	13.95%	1.19%	1.10%	1.35%
Leaves+LPF	9.15%	11.21%	13.47%	0.83%	0.86%	1.03%
Leaves+SPF	9.44%	11.66%	14.02%	1.38%	1.64%	2.07%

5.4 Results

The obtained results for the second test set are given in Table 2. The heuristic generated schedules that have an average gap of 1.5% from the lower bound for all heuristics that use finish first affectation rules and an average difference of 12% for the rest of rules. It is clear that the computed makespan for rules that use finish first affectation is better than the makespan computed for the rest of rules.

For the selection rules, the results are almost the same because for the considered form of tree at each step of selection tasks leaves are in the highest level.

6 Conclusions

In this paper, a parallel machine scheduling problem in which machines are subject to non-availability constraints was considered. We study the semi-resumable situation with the objective of minimizing the makespan. We then gave a comparative study for the intree case. The considered problem can be extended to outtree case to study the behaviour of the heuristic for this form of trees.

References

1. Huang, K., Wang, Z.Y., Weng, X.X., Lin, W.: A scheduling algorithm on heterogeneous star and tree grid computing platform. In: International Conference on Convergence Information Technology, vol. 78, pp. 347–351 (2007)
2. Fan, S.: Session scheduling algorithm of grid computing. In: Third International Conference on Knowledge Discovery and Data Minning, vol. 78, pp. 3–6 (2010)
3. Patni, J.C., Aswal, M.S., Pal, O.P., Gupta, A.: Load balancing strategies for grid computing. In: 3rd International Conference on Electronics Computer Technology (ICECT), vol. 3, pp. 239–243 (2011)
4. Fan, B., Li, S., Zhou, L., Zhang, L.: Scheduling resumable deteriorating jobs on a single machine with non-availability constraints. Theoret. Comput. Sci. **412**, 275–280 (2011)
5. Graham, R.L., Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G.: Optimization and approximation in deterministic sequencing and scheduling. Ann. Discrete Math. **5**, 287–326 (1979)
6. Billaut, J.C., Moukrim, A., Sanlaville, E.: Flexibility and Robustness in Scheduling. Hermes Science/Lavoisier, Paris (2005)
7. Coffman, E.G., Graham, R.L.: Optimal scheduling for two-processor systems. Acta Informatica **5**, 200–213 (1972)
8. Yee Lee, C.: Parallel machines scheduling with non simultaneous machine available time. Discrete Appl. Math. **30**, 53–61 (1991)
9. Yee Lee, C.: Machine scheduling with an availability constraint. J. Global Optim. **9**, 395–416 (1996)
10. Safi, A., Trystram, D., Canon, L.: A proactive approach for coping with uncertain resource availabilities on Desktop grids. In: High Performance Computing, Goa, India, vol. 9, pp. 395–416 (2014)
11. Garey, M.R., Johnson, D.S., Tarjan, R.E., Yanakakis, M.: Scheduling opposite forests. SIAM **4**, 72–93 (1983)
12. Sanlaville, E., Mahjoub, A., Guinand, F.: Problemes d'ordonnancement sur machines paralleles avec tches communicantes et indisponibilites (In french), ROADEF, Marseille, France (2014)
13. Ben Abdellafou, K., Korbaa, O.: A novel algorithm for scheduling in trees on two parallel machines with unavailabilities. In: IEEE International Conference on Systems Man and Cybernetics, Budapest, Hungary (2016)
14. Ben Abdellafou, K., Sanlaville, E., Mahjoub, A., Korbaa, O.: Scheduling UECT Trees with communication delays on two processors with unavailabilities. In: INCOM 2015, Ottawa, Canada, vol. 48, pp. 1790–1795 (2015)

Network Intrusion Detection Using Danger Theory and Genetic Algorithms

João Lima Santanelli(✉) and Fernando Buarque de Lima Neto

Computing Engineering Programme, Polythecnic School of Pernambuco, University of Pernambuco (UPE), Recife, PE, Brazil
{jls2,fbln}@ecomp.poli.br

Abstract. One of the most concerning problems faced by practitioners, within the development and operation of IT communication networks, is the crescent number of network intrusion attempts. That kind of attacks compromise the integrity of several services provided through the Internet. This paper presents a technique capable of optimize Danger Theory-based Intrusion Detection Systems through the use of a Genetic Algorithms. To validate the approach, tests were performed on the KDD Cup 1999 database, provided by the University of California Irvine (UCI).

Keywords: Intrusion detection · Artificial immune system · Danger theory · Dendritic cells algorithm · Genetic algorithms · Optimization

1 Introduction

With the advancement of technology, society is becoming increasingly dependent on digital communications. Among these, Internet is doubtless the one gaining more ground over the time and thus, deserves also progressively more attention.

The exponentially growing of the amount of messaging throughout the computer networks is the perfect scenario for the rise of new threats. The flaws on the security systems due to Denial of Service Attacks, for example, is one of the problems that brings proportionally great economic loss to many countries around the world [1].

Computer network attacks are, mostly, not a goal, but a means for attaining harmful actions and crimes over the internet. The most evident prove of that are the websites' security, which are never so secure that likes cannot be broken. However, along the time, there were other new motivations for attacks such as political questions, extortion and even contests for the supremacy among the attackers. Therefore, it is necessary to seek constantly improvement on network incident detection techniques. A system capable to detect malicious traffic in a computer network is called an Intrusion Detection System (IDS).

Among several techniques to attain cyber-attacks, Computational Intelligence (CI) stands out because of its autonomy, adaptability and lightweight.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_39

Within CI, Artificial Immune Systems (AIS) [2] seem to be a serious contestant because of its speed and flexibility in light of the scarce number of labeled examples. To explain the operation of the Human Immune System (HIS), Polly Matzinger elaborated the Danger Theory [3–5], in 1994. The core idea is that organisms identify dangerous situations based on alarms generated by the affected cells or tissues. Drawing inspiration in the human defense mechanisms, it is possible to associate cyber threats to pathogenic agents and the network to the human body. Therefore, it is reasonable to create an IDS based on AIS.

Another technique that draws inspiration on Biology is the Genetic Algorithm (GA) [6–8]. They are, for example, able to optimize the solution search for a given complex combinatorial problem. The approach is comprised of associating the problem solutions to individuals of a population and, throughout genetic operations (such as selection, cross-over and mutation), allow them to evolve, originating more adapted individuals, which, then represent better candidate solutions for the problem at hand.

This paper intends to use Genetic Algorithms to optimize the parameters associated to the input patterns of a Danger Theory-based Intrusion Detection System (DT-IDS). To reach this goal, a system was implemented that consists of a Danger Theory-based detection algorithm, more precisely, the Dendritic Cell Algorithm (DCA) [9], and a Genetic Algorithm that optimizes the detection parameters of the first one. This system extracts the data from the KDD Cup 1999 database [10], provided by the University of California Irvine (UCI), and compares the results of the performed classification with the labels contained in database records.

This work is divided in six sections. The first one contextualizes the research, exposing the motivations and objectives of this work. The second one provides the theoretic background needed to understanding the problem and the solution and addresses basic concepts of Danger Theory, Genetic Algorithms and Cyber Attacks. The third one discusses related works, pointing out the most important aspects of each one and comparing their solutions with the one adopted in this paper. The fourth one describes the experimental sets assembled to validate the approach presented in this paper. The fifth one discusses the results obtained throughout the performed experiments. The sixth one presents the final considerations about this research and suggests some improvements that would be done in future works.

2 Background

2.1 Danger Theory

The Danger Theory was developed by Polly Matzinger [3], in 1994, in order to explain the operation of the Human Immune System. Previously, it was believed that the Human Immune System acts based on the “self-non-self” theory, recognizing elements that were part of the system or not.

The main idea of Matzinger's theory is that the immune responses to pathogenic agents are given not based on the "self-non-self" concept, but activated by danger signals - in a population dynamics manner.

Matzinger's model includes three type of signals [4]:

1. *Signal Zero*: is the danger signal emitted by a compromised cell or tissue;
2. *Signal One*: indicates that an antigen has been detected by an immune cell; and
3. *Signal Two*: co-stimulatory signal emitted by an Antigen-Presenting Cell (APC) requesting an immune response from T-cells.

The Dendritic Cell Algorithm (DCA), proposed by J. Greensmith, is inspired by Matzinger's model, that describes the Antigen Presentation Cell (APC) as a structure capable to activate virgin T cells, starting the immune response. The Dendritic Cell (DC) is a type of APC [3].

According to Greensmith [9], there are three types of Dendritic Cells:

1. *Immature DC*: collect parts of antigens and signals from other cells;
2. *Semi-mature DC*: identifies the local signals as "safe" and present the antigen to T cells, resulting in tolerance; and
3. *Mature DC*: identifies the local signals as "danger" and present the antigen to T cells, starting the immune reaction.

The DCA consists, basically, in generating a group of entities representing Dendritic Cells and exposing them to input patterns, that play the role of antigen parts. The pattern can be interpreted by each cell as "normal" or "anomaly". The classification is based on two values associated to the pattern: *p_safe* and *p_danger*, that, respectively, represent the probabilities of the pattern is associated with a safe or a dangerous situation [11]. These values are intrinsic to the input patterns, and, based on them, an immature DC can migrate to the other states: mature or semi-mature.

After the migration of the DCs, the result of the classification of the pattern as "safe" or "dangerous" can be performed using a voting mechanism based on the number of mature and semi-mature cells with the respective antigen associated to the pattern.

In a scenario with multidimensional input patterns, the challenge to be faced is how to calculate the *p_safe* and *p_danger* values associated with each input pattern. They could be represented as a weighted sum of the probabilities associated with each intrinsic attribute of the pattern. Certainly there will be an optimum set of weights that will make these values more accurate, resulting in a better classification process.

2.2 Genetic Algorithms

The first Genetic Algorithm was created by John H. Holland, in the early seventies [6]. They can be implemented in several manners, but the central idea

consists of associating the solutions of a problem to individuals of a population (or chromosomes) and, throughout genetic operators, make them evolve, becoming more adapted. The function that measures the fitness of an individual of the population is called Fitness Function, and it is modeled according to the problem that is intended to be solved.

The genetic operators/concepts applied in this present study are (1) selection, (2) crossover, (3) mutation and (4) elitism [7].

1. *Selection*: To select the individuals that will survive to the next generation/that will be selected for reproduction; it was chosen the "roulette" mechanism, where the probability for an individual to be chosen is directly proportional to the value of its fitness function;
2. *Crossover*: For each pair of individuals chosen to reproduce, random attributes are selected to be permuted, originating two new individuals that may be added to the population poll;
3. *Mutation*: For each individual of the population, there is a probability of some of its attributes be randomly modified to originate a new individual; and
4. *Elitism*: The fittest individuals of each generation will pass automatically to the next, ensuring that the better solutions will be preserved.

In the previous section one can find methods to calculate *p_safe* and *p_danger* values for DCA, and there is an optimum set of weights capable to provide more accuracy to the classification process. To reach this goal, a good approach is the use of Genetic Algorithms.

In this present study, the set of weights that defines the values obtained for *p_safe* and *p_danger* is modeled as an individual of GA's population. More details about the operation of the algorithm will be provided in the fourth section of this paper, where is discussed the experimental methodology. For now, it suffices to say that these sets of weights are modified by the GA to provide more accurate results.

2.3 Cyber Attacks

Some intrusion experts believe that the most recent attack types are variants of known ones. Therefore, the known signature attacks must be sufficient to detect new types [10].

This work will use four frequent types of attacks, provided by the KDD Cup 1999 database: (1) DoS, (2) R2L, (3) U2R and (4) Probe.

1. *DoS*: Denial of service attack, which intend to turn the target unavailable;
2. *R2L*: Non-authorized remote access to a server;
3. *U2R*: Non-authorized access to super-user (root); and
4. *Probe*: Vulnerability exploiting of the target.

In the scenario of Intrusion Detection (as a classification kind of problem), there are four types of possible results: (1) true positive, (2) true negative, (3) false positive and (4) false negative.

1. *True Positive*: when the system notifies an intrusion that is being performed;
2. *True Negative*: when the system classifies correctly the normal traffic;
3. *False Positive*: When the system notifies an intrusion that does not exist in fact; and
4. *False Negative*: When the system does not notify an intrusion that is being performed.

The definitions above, make clear that the worst case among these four, considering a computer network intrusion scenario, is the *false negative*. This is because a *false negative* makes the system assume that there is no threat and, so, it is not necessary to take any counter-action, leaving the network unprotected. On the other hand, the *false positive* generates an alarm that leaves the network on alert. The only problem associated to this situation is that the network administrator is bothered, but it does not offer any harm to the network. So, in an Intrusion Detection System it is necessary that the *false negative* rate is as low as possible, and is is desirable that a low *false positive* rate is also achieved.

3 Related Works

Many works adopted KDD Cup 1999 Database [10] to test their theories. The use of Computational Intelligence to solve Intrusion Detection problems is very widespread.

Srinoy and Kurutach [12] combined Artificial Ant Clustering and K-Mean Particle Swarm Optimization (K-PSO) using the KDD Cup 1999 Database achieving a reduction of the dimensionality of the attribute set from 41 to 10.

The Particle Swarm Optimization (PSO) allied to Support Vector Machines (SVM) was used by Srinoy [13] to detect malicious network activities, reaching an accuracy of 96,11%, while the pure SVM, used by Somwang and Lilakiatsakun [14], performed 97,44% of correct classifications.

Dantziger and Lima Neto [2] purposed a hybrid approach using concepts of Danger Theory, Nave Bayes and Multi Agent Systems to perform intrusion detection in an IEEE 802.11 network.

In 2014, Zekri et al. [15] presented a summary of some major works on Artificial Immune System for Intrusion Detection Systems since 2005. Positive Selection, Generation of Detectors, Innate Immunity, Cooperative AIS and Clonal Selection figure among the techniques used in these works [15]. But it is notorious that the most recent studies focus on Dendritic Cell Algorithm (DCA) and Negative Selection Algorithm (NSA).

Zekri compared DCA with NSA in her work [15] and found most significant results for the first one, strengthening the idea that DCA is the AIS's state of art for IDS.

4 Experimental Methodology

This paper intends to present the use of Genetic Algorithms to optimize Danger Theory-based Intrusion Detection Systems (DT-IDS). For that we implemented

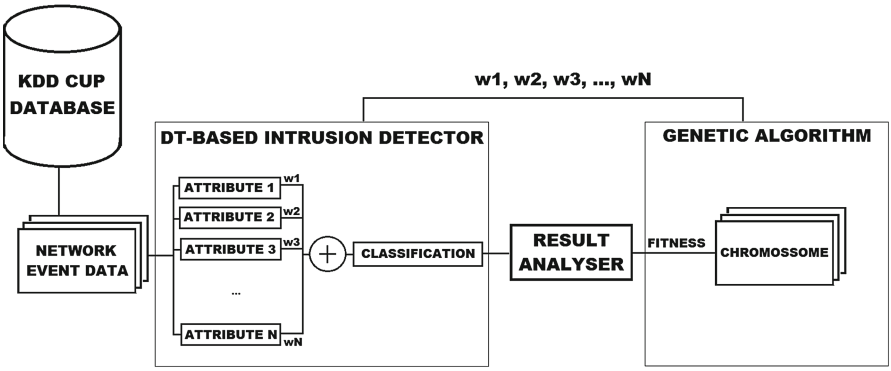


Fig. 1. Implemented model using danger theory and genetic algorithms

a DT-IDS optimized with a Genetic Algorithm. The overview of our approach is shown in Fig. 1, which components are better explained in the following subsections.

4.1 KDD Cup 1999 Database

The database for tests of our approach is provided by the KDD Cup 1999 [10] and is used for a contest where the competitors were challenged to create an Intrusion Detection System, classifying the network events as "good" or "bad" connections.

The amount of training and testing data, distributed by attack classes, are shown on the Table 1.

Table 1. KDD Cup 1999 database figures

Event type	Amount of data per type	
	Training	Testing
Normal	99.278	60.593
DoS	391.458	223.298
U2R	52	39
R2L	1.126	5.993
Probe	4.107	2.377
Unknown	-	18.729
TOTAL	**494.021**	**311.029**

The type of event defined as "unknown" represents the types of attacks that are not available in the training set, but as discussed in Subsect. 2.3, must be detected based on the known signature attacks.

4.2 Danger Theory-Based Intrusion Detector

The implemented system has a decision core based on the Dendritic Cell Algorithm (DCA) [9]. The Dendritic Cells (DC) receive signal zero from the sensors monitoring each attribute of a network event. These sensors could be associated to the cells and tissues compromised by the pathogenic agent. In each training iteration, the immature DCs are exposed to the antigen, represented by the set of attributes extracted from each network event registry obtained from the KDD Cup 1999 database, and so, they can evolve or not to the semi-mature or to the mature state. After some iterations, if the cell does not evolve, it is replaced by a new one. Each evolved cell migrates to a new group and is replaced by an immature cell, maintaining the immature DC's population size.

After training, the antigens represented by the testing registries are presented to the migrated cells and, throughout a voting mechanism, it is decided if the recognized pattern represents a threat or not.

Each network event registry extracted from the database represents one iteration of DCA. It means that DCA's training phase will have 494.021 iterations while the test phase will have 311.029 iterations, according to Table 1, totalizing 805.050 iterations per execution.

The challenge to be faced is how to model a set of attributes (many of them continuous) as an input pattern capable to be recognized by Dendritic Cells. To solve this problem, we scanned the training data set and, for each attribute, we set an interval that goes from the minimum to the maximum value, dividing them into a discrete number of sub intervals.

The hardest aspects to tackle are the number of sub intervals and the differences between two input patterns associated to different types of attacks. On the other hand, a large number of sub intervals represents a larger domain of input patterns to be recognized by DCs and, for example, two patterns associated to the same network event type would be classified differently. To determine the better number of sub intervals to be used, several experiments were performed running DCA for different configurations and, according to the results exposed on Table 2, we choose to divide each interval in three parts.

Another important aspect to be considered is that the *p_safe* value is associated to the probability of the antigen pattern represent a non-dangerous situation [11], but, in most examples using the canonical form of DCA found in literature, this value is not numerically equal to this probability. To model this problem, we calculate the *p_safe* value by multiplying that probability with a constant value. Several experiments were performed and we can note in Table 2 that the values for the *p_safe* multiplier between 1.49 and 1.53 present the best results showing an accuracy variation less than 0.07%. So, we decided to set the *p_safe* multiplier as 1.5.

To avoid any bias in the final results, all the tests performed to determine the parameters presented on Table 2 used, as training set, 10% of the training data provided by UCI in the KDD Cup Database [10]. The validation is performed upon the complete training database and each accuracy value was calculated five times.

Table 2. DCA results for each number of sub intervals

Sub intervals	Accuracy for p_safe multiplier							
	1.43	1.45	1.47	1.49	1.51	1.53	1.55	1.57
2	0.9933	0.9934	0.9937	0.9938	0.9935	0.9933	0.9930	0.9929
3	0.9298	0.9326	0.9409	0.9955	0.9950	0.9949	0.9947	0.9946
4	0.9202	0.9234	0.9363	0.9856	0.9857	0.9937	0.9936	0.9935
5	0.9114	0.9143	0.9321	0.9759	0.9769	0.9929	0.9927	0.9926

On this work, DCA's *p_danger* value was modeled as 1 *p_safe*. The cell's maturation depends on the antigen's CMS value, defined as *p_safe* + *p_danger* [11]. The CMS value represents the relevance of antigen's pattern. The higher is antigen's *p_safe*, the larger is the certainty about its pattern to be associated with a non-dangerous situation. Analogously, the higher is antigen's *p_danger*, the larger is the certainty about its pattern to be associated with a dangerous situation. But in this case, the sum *p_safe* + *p_danger* is always equal to 1. Therefore, to adapt our approach to the idea that CMS represents the relevance of antigen's pattern, we defined it as the absolute value of the difference between *p_safe* and *p_danger*.

To be modeled as an antigen, the input pattern represented by the array containing the number of the sub interval in which is located the value of each attribute must be associated to a value of *p_safe* and *p_danger*, as discussed in Subsect. 2.1. These values determine how the DCs will respond to the exposure to the antigen.

The implemented decision logic to calculate the *p_safe* of each antigen is regulated by a set of weights (w1, w2, w3, ..., wN), as shown previously in the Fig. 1. Each attribute has its relevance associated to the respective weight and, therefore, it is desired to find the optimum weight set resulting in an optimal performance regarding security.

Each evaluated parameter is associated to a network attribute. For each one, a value between 0 and 1 is calculated, which represents the odds that the attribute is related to malicious traffic. Each value is multiplied by the weight representing the parameter's relevance, obtaining a partial result. The antigen's *p_safe* is calculated as the sum of these partial results.

Assigning the same value to the weights associated with each attribute would mean that all attributes have the same significance regarding the detection of anomalies. However, this is not true and it is not what is necessary for attaining high levels of intrusion detection. Thus, we decided to associate all DT-IDS decision core weights to chromosomes of a Genetic Algorithm so it can reach an optimal solution based on detection rates obtained through the Result Analyzer, as in Fig. 1.

4.3 Genetic Algorithm-Based Parameter Optimization

To optimize the DT-IDS, we associated the parameters used to calculate antigen's *p_safe* values to individuals in a Genetic Algorithm's population.

Due to the fact that DCA demands a considerable amount of time to be executed (approximately three minutes in these experiments, with ten immature cells in the population) the Genetic Algorithm was modeled with a small population (eigth individuals per iteration). To assure that the best individuals will survive to the next generation, it was applied with 25% elitism.

The individual's selection for reproduction (crossover) is based on a "roulette" mechanism where the probability to be chosen is directly proportional to the individual's fitness. The fitness is set as the accuracy of DCA in classification of the training events using the associated parameters.

The reproduction is performed as a crossover mechanism that results in two new individuals. Each new individual's attribute is equal to the same attribute in one of the parents.

The selection of the individuals which will pass to the next generation is based in the "roulette" mechanism described previously in this section. The best 25% of the population is automatically selected.

To mitigate the stagnation of the algorithm in a local maximum, we set the mutation rate as 10

Stop conditions for the GA were established as the stagnation of the best fitness value for 10 rounds or the execution of 60 iterations. If at least one of the conditions is reached, the algorithm stops.

5 Simulation Results

To validate the approach, several experiments were performed. The first experimental set considers only known attack types, representing a less dynamical environment, where new types of attacks are not a real threat.

The second one represents a dynamical environment where new types of cyber-attacks emerge as new threats. The great challenge is to detect malicious activities that did not figure on the training data set, proving the adaptability of the approach.

5.1 Considering only Known Attacks

The results obtained throughout the first experimental set are shown in Table 3, where the labels "Total" and "FP" represent respectively the total accuracy and the false positive rate of the detection process.

5.2 Considering Unknown Attacks

The results obtained throughout the second experimental set are shown in Table 4, where the labels "Total" and "FP" represent respectively the total accuracy and the false positive rate of the detection process. The symbol "?" represents the accuracy in detecting unknown types of attacks.

Table 3. DT-IDS optimized by genetic algorithm (known attacks)

Detection accuracy						
Normal traffic	Attacks				Total	FP
	DoS	U2R	R2L	Probe		
89,21%	99,42%	28,21%	2,92%	89,44%	95,23%	2,24%

Table 4. DT-IDS optimized by genetic algorithm

Detection accuracy							
Normal traffic	Attacks					Total	FP
	DoS	U2R	R2L	Probe	?		
89,21%	99,42%	28,21%	2,92%	89,44%	29,14%	91,25%	2,10%

6 Conclusion and Future Works

6.1 Discussion

The use of Genetic Algorithms for the optimization of Danger Theory-based Intrusion Detection Systems, as discussed in this paper, produced very interesting practical results.

Even though Tables 3 and 4 show that the approach has not proved effective in detecting attacks of the type U2R, R2L (corresponding, respectively, to only 0.01% and to 0,22% of the training set) and the unknown attacks. This shows that the technique adds strong dependency on training set, not proving very able to adapt to a scenario with new threats. Fortunately, for known attacks our approach has been proved quite satisfactory (95,23% of accuracy). The results, for known attacks shows that the use of Genetic Algorithms as optimizer of Danger Theory-based classifiers is quite efficient, especially in DoS attack detection and false negative reduction.

An important achievement of this study is the reduction of false positive rates. A 2,1% false positive rate makes the system more independent of human intervention, considering that each false positive must be analyzed by the network administrator. The obtained false negative rate is very acceptable too.

Considering the aspects presented in this study, it is concluded that the use of the approach based on Danger Theory and optimized by Genetic Algorithms is recommended for less dynamic environments, where the emergence of new threats do not occur very often.

6.2 Future Work

The use of Genetic Algorithms can bring some inconvenience. One of the negative aspects is the need for large computational effort and, somehow acts against the benefits of AIS (DT), such as speed and flexibility.

With the advent of distributed computing, this problem can be easily alleviated, allowing the implementation of Genetic Algorithms with larger populations and more iterations, as parallelism decreases the time for completion of each step.

Thus, as future work, it is suggested to apply the concepts of Distributed Systems to enhance the performance of the Genetic Algorithm, allowing it possibly to reach better results.

Acknowledgements. Thanks to Computer Engineering Program of School of Engineering of University of Pernambuco and the Brazilian Army which made this research possible.

Thanks to Moisés Danziger, whose work [2,3] and inspiration were seminal for the present study.

References

1. Pau, L.-F.: Business and social evaluation of denial of service attacks in view of scaling economic counter-measures. In: 2010 IEEE/IFIP Network Operations and Management Symposium Workshops, pp. 126–133 (2010)
2. Danziger, M., Lima Neto, F.B.: A hybrid approach for IEEE 802.11 intrusion detection based on AIS, MAS and Nave Bayes. In: International Journal of Computer Information Systems and Industrial Management Applications, vol. 3, pp. 193–201, Pernambuco, Brazil, 2011 (2011)
3. Matzinger, P.: Tolerance, danger and the extended family. Ann. Rev. Immunol. **12**, 991–1045 (1994)
4. Matzinger, P.: The danger model: a renewed sense of self. Science **296**, 301–305 (2002)
5. Matzinger, P.: The danger model in its historical context. Scand. J. Immunol. **54**, 4–9 (2001)
6. Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Ardor (1975)
7. Razali, N.M.: Genetic algorithm performance with different selection strategies. In: Solving TSP, Proceedings of the World Congress on Engineering 2011, Vol II, London, UK (2011)
8. Meng, Q.C.: Genetic Algorithms and Their Application. Publishing Company of Shandong University, Jinan (1995)
9. Greensmith, J., Aickelin, U., Cayzer, S.: Introducing dendritic cells as a novel immune-inspired algorithm for anomaly detection. In: Jacob, C., Pilat, M.L., Bentley, P.J., Timmis, J.I. (eds.) ICARIS 2005. LNCS, vol. 3627, pp. 153–167. Springer, Heidelberg (2005). doi:10.1007/11536444_12
10. KDD Cup 1999 Database. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. Accessed Jun 2016
11. Brownlee, J.: Clever Algorithms: Nature-Inspired Programming Recipes. LuLu, Raleigh (2011)
12. Srinoy, S., Kurutach, W.: Combination artificial ant clustering and K-PSO clustering approach to network security model. In: International Conference on Hybrid Information Technology (2006)
13. Srinoy, S.: intrusion detection model based on particle swarm optimization and support vector machine. In: CISDA. IEEE Symposium on Computational Intelligence in Security and Defense Applications, pp. 186–192 (2007)

14. Somwang, P., Lilakiatsakun, W.: Computer network security based on support vector machine approach. In: 2011 11th International Conference on Control, Automation and Systems, 26–29 October 2011. KINTEX, Gyeonggi-do, Korea, pp. 155–160 (2011)
15. Zekri, M., Souici-Meslati, L.: Immunological approach for intrusion detection. revue africaine de la recherche en Informatique et Mathématiques Appliquées (ARIMA) **17**, 221–240 (2014)

A Gamification Model for Resource Sharing in Malaysian Schools Using Cloud Computing Platform

M. Nordin A. Rahman(✉), Abdullahi Nababa Saidu,
M. Fadzil A. Kadir, Syadiah Nor Shamsudin,
and Syarilla Iryani A. Saany

Faculty of Informatics and Computing, University of Sultan Zainal Abidin,
Besut Campus, 22200 Besut, Malaysia
mohdnabd@unisza.edu.my

Abstract. Cloud computing is a technology that delivers a pool of configurable computing resources to the users. The insufficient systems for sharing of resources, poor extended pattern of resource sharing system, inadequate network and computing facilities have become an issue in school management. This article discusses a model for resource sharing in Malaysian schools using cloud computing platform. The sharing of resources is modelled based on Multi-Person Prisoner's Dilemma and Peer-Approved incentive mechanism that gives justification to users in the resource utilization. The concept of software as a service is adopted to make the model can be managed in the cloud environment. Based on the proposed model, a prototype is developed and deployed into cloud using OpenStack architecture. The prototype is evaluated using questionnaire instrument that is based on four attributes of usability metrics. The result shows that the model usability achieved 79.28% of effectiveness, 81.35% of efficiency, 91.8% of learnability and 93.18% of satisfaction. The proposed model could save the cost of buying computing and storage infrastructure since all the resources are cloud based.

1 Introduction

The concept of an intelligence school information management has been identified as not only representing a technological innovation for realizing global integration but also as considering the current demand on global issues such as socio economy, application of green technology and smart governance. An intelligence school information management has been characterized as a tool for generating a particular form of spatial innovation that is based on web services, integrated module, involved large data sets and real time information feeding and response. It is also covers the level of service and security such as robust connectivity, remote network access, dashboard for virtualization management, priority resourcing by client, integrated email and telephony services etc.

The Malaysian government always put the first priority on budget allocation for the Malaysian education system. However, student outcomes have not always matched the

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_40

resources channeled into the system. The government will maintain current levels of this investment for the aspiration to further maximize student outcomes within the budget level. The situation requires funds to be distributed to agencies involved in accordance with the priority needs such as environment and socio-economic. As stated in Malaysian Education Blue Print 2013–2025, one of the aspirations in Malaysian Education System is efficiency. It requires the ministry and schools to maximize the student outcomes within the allocation budget given by the government. However, to improve effectiveness in the teaching, learning, management and evaluation process, the use of technological innovation that requires a certain investment cannot be avoided. Therefore, the integrated technology of software, hardware and platform that can be shared by ministries, State Education Department (SED), District Education Office (DEO) and schools must be developed.

Cloud computing is a new technology service that provides several concepts such as distributed application design, virtualization, enterprise IT management and grid application. The main aims of cloud computing is to allow more openness methods in deploying and scaling applications with low investment for hardware, software or platform [9]. This technology refers to flexible network accessible computing resource pools that can be distributed as per user request. Resources and processing power can be managed through service dashboard that its configuration settings are available for users without the need for technical personnel involvement.

1.1 Information Management in Malaysian School

In January 2015, the Ministry of Education in Malaysia has launched an online system called School Management System (SPS). This system is used by schools, DEO and SED to achieve a single database application for information related to schools, teachers and students. SPS is also aimed at preventing teachers from performing repetitive data entry in multiple systems developed by the ministry. SPS does not include teaching, learning and assessment modules. SPS can be accessed through the address http://apps4.moe.gov.my/sps and has three (3) modules that are *school management, teacher management* and *student management*.

Currently, Malaysian schools, DEO or SED do not have a specific platform for information sharing of teaching, learning and student assessment materials. The archiving materials in this particular platform help student to make reference more easily and could be a comprehensive inventory for students to have an alternative learning support. While the management of education such as school administrators, DEO and SED can use the archival material for the purpose of planning, quality improvement and increase the effectiveness of the process of teaching, learning and assessment in schools.

1.2 Cloud Technology

The US National Institute of Standards and Technology (NIST) defines cloud computing as “a model for enabling ubiquitous, convenient, on-demand network access to a

shared pool of configurable computing resources such as network, servers, storage, applications and services that can be rapidly provisioned and released with minimal management effort or service provider interaction" [5]. Cloud computing offers significant cost savings, efficiencies, flexibility and scalability as well as opportunities for the innovative development and delivery of new services [1, 2, 4, 6, 8]. NIST is also divided cloud computing into three (3) service layers:

- Software as a Service (SaaS) – the software or applications run on the cloud (in provider's servers) and the consumer can access via the Internet to work on this application normally.
- Platform as a Service (PaaS) – an affective environment and powerful tool for developers or organizations to create and deploy the applications.
- Infrastructure as a Service (IaaS) – the basic layer of computing resources that provides on demand and scalable full infrastructure resources such as servers, software, network equipment and storage.

Deploying cloud computing can be implemented in four (4) methods: private cloud, public cloud, community cloud and hybrid cloud.

- Private cloud – cloud infrastructure is operated solely for particular organization.
- Public cloud – cloud infrastructure is available to several consumers, individuals or organizations and the provider of service owns the infrastructures.
- Community cloud – cloud infrastructure is shared by several organizations that have same concern with particular issues such as mission, policy or security.
- Hybrid cloud – a mix cloud infrastructure between two or more previous types

2 Related Works

To date, Malaysia has 10,154 schools that consist of 7,760 primary schools and 2,394 secondary schools. The overall number of students is 5,120,802 with a total of 196,077 students are in preschool, 2,704,046 students are in primary school and 2,220,679 students are in secondary level. While the number of teachers is 419,820 of which 238,073 teachers are served in primary school and 181,747 teachers are allocated in secondary school. In the blueprint Malaysian Education Development Plan 2013–2025, the Ministry of Education will ensure that students not only learn using ICT but also can use it in effective ways to improve their learning. Therefore, the Ministry attempts for strengthening basic ICT school infrastructure while introducing ICT solutions well proven in the educational system. Steps to be taken include:

- Provide students with the skills and knowledge to learn effectively and live productively in global and digital world
- Equip all 10,000 schools nationwide with 4G Internet access and virtual learning platform that can be used by teachers, students and parents through 1BestariNet program
- Provide training for all teachers to apply ICT in teaching and learning to support student learning

- Increase the ratio of students to ICT equipment 10:1. This ratio may reduce further depending the impact assessment and the availability of funds
- Leading the way with the delivery of learning using ICT innovations such as distance learning and self-learning content (self-paced)

The education stake holders such as SED, DEO and school management are striving to deliver higher levels of teaching and learning services, but challenged with flat funding to support this effort. A solution that can be taken is to use the vast of databases available to gain a better opportunity in providing services more efficiently. This can be initialized with engaging teachers, subject matter experts and school management with open data initiatives to create innovative information sharing programs. The model that applies advanced analytics and collaboration to the teaching, learning and assessment requirements provides a roadmap to achieve these aims, but the education stake holder IT department is often limited by staff size, skill and budget and is unable to implement the model on their own.

Cloud computing and software as a service can provide the means to acquire data-driven teaching, learning and assessment capabilities that support intelligence school information management. Trends driving intelligence school management growth are as follows:

- The increasing of ICT literature among school management, teachers and students
- The growing number of teachers and students who have a connection to the Internet
- Smart sharing of teaching, learning and assessment materials become an important issue discussed in modern school management
- SED, DEO and schools must provide services that span the digital divide
- Proliferating technology especially in cloud computing equals exploding amounts of learning and assessment materials

[1] proposed a model for the provision of shared archiving services using cloud computing technology. This archiving as service model is developed based on the Open Archival Information Reference Model. The developed model allows the sharing of functionality and information objects by making them as services at the user application level. The model has been experimented to the Japanese Government and National Archives as an application for Archives and Record Management. The experiment shows that the model provides efficient archiving functionality and improve document life cycle as well as simplify record transfer activities. [6] has developed a cloud computing based platform for supporting teaching and learning using a Virtual Computing Lab (VCL) approach and the model is experimented at North Carolina State University State University. Based on the evaluation of the model, the university achieved the following benefits:

- Maintaining a VCL not exceed 21% of its traditional counterparts
- Increase student access to professional and research tools
- Increase the IT agility to support teaching and learning
- Overcome several issues related to lab administration
- Enhanced the capability of distance education

3 The Gamification Model for Resource Sharing

The gamification model for resource sharing is developed based on the integration of three techniques: *prisoners' dilemma*, *peer-approved incentive* and *application layering*. The prisoners' dilemma and peer-approved incentive techniques managed the resource sharing and its justification in the user's resource utilization process.

3.1 Resource Sharing Method

The consolidation of resources from various providers requires for introducing some negotiation technique. In CARS, the multi-person prisoner's dilemma technique is used in order to encourage the users (providers) to share their resources. The multi-person prisoner's dilemma is a game where two players are partners in a crime with each player having two options of either confessing or not confessing [10]. The players are to choose an action simultaneously and at the same time each player wants to maximize his utility.

In the context of resource sharing, users can either choose to share more resources (which is considered as not-preferred option) or to share fewer resources (which is considered as preferred option). The multi-person prisoner's dilemma technique is combined with an incentive mechanism in order to provide justification and encouragement for users to share their resources. According to [7], there are three incentive mechanism model can be considered in designing information system for sharing resources: *token-exchange, peer-approved* and *service-quality*. Based on the simulation result provide by [7], peer-approved model is the best solution for any information system that implemented without involving direct payments. Therefore, the peer-approved model is used to integrate with multi-person prisoners' mechanism for developing the model of cloud based resource sharing in Malaysian school environment.

In this resource sharing problem, several users contrary to prisoner's dilemma that involves two players are considered. Therefore, the following assumptions are made:

- There are N users in the system, and each user has at least one resource to share.
- User can either share one resource (preferred alternative) or can decide to share many resources (non-preferred alternative).
- Rating of user is proportional to the number of resources shared by that user.
- The users are categorized into three types: *main-contributor, partial-contributor*, and *weak-contributor*.

Let: main-contributor, M_c, be a user that shared at least P resources; partial-contributor, P_c, be a user that shared at most Q resources and weak-contributor, W_c, be a user that shared only R resource or no resource, where the variables P, Q, and R can be customized by administrator and $P > Q > R$.

Suppose that M_{c1} shared d_1 resources, M_{c2} shared d_2 resources, … M_{ci} shared d_i resources where $d \in Z^+$. Thus, $D = d_1 + d_2 + \ldots + d_i$ is the total resources that are shared by main contributors. This condition can be represented by:

$$D = \sum_{k=1}^{i} d_k \tag{1}$$

Suppose that P_{c1} shared e_1 resources, P_{c2} shared e_2 resources, … P_{cj} shared e_j resources where $e \in Z^+$. Thus, $E = e_1 + e_2 + \ldots + e_j$ is the total resources that are shared by partial contributors. This form can be represented by:

$$E = \sum_{m=1}^{j} e_m \tag{2}$$

Suppose that W_{c1} shared f_1 resources, W_{c2} shared f_2 resources, … W_{ck} shared f_k resources where $0 \leq f \leq 1$. Thus, $F = f_1 + f_2 + \ldots + f_k$ is the total resources that are shared by weak contributors. This situation can be represented as:

$$F = \sum_{n=1}^{k} f_n \tag{3}$$

The total resources that are shared in the system is the sum of all the resources shared by users and can be denoted as:

$$G = \sum_{k=1}^{i} d_k + \sum_{m=1}^{j} e_m + \sum_{n=1}^{k} f_n \tag{4}$$

In the proposed model, category of user is calculated based on their contributions. The boundary of main contributor is set to at least P resources in order to encourage user to contribute more into resource database. The variable of P can be customized to accommodate the system requirements. Despite all these restrictions, the system is open and accessible to all users. However, user can access resources available to their category. There are several existing enterprise applications that give benefits to user based on usage and contribution. For example, in 4Shared.com, *4shared application* allows uploading, downloading and sharing of multimedia files, documents, and other contents based on cloud services. Initial user is given a space of 10240 MB. After a successful registration, the storage space can be expanded to 15 GB and user will be given more space as they keep using the application. The utility of each user's category can be denoted as follows:

- Main-contributors are those that shared at least P resources to the system. This group of users can access all the resources. The utility of main contributors is given by:

$$\text{Utility } M_c = \sum_{k=1}^{i} d_k + \sum_{m=1}^{j} e_m + \sum_{n=1}^{k} f_n \tag{5}$$

- Partial-contributor is the user who share at least Q resources where $Q < P$. Thus, peer-approved mechanism is applied where users can access resources based on their ratings. This category of users can access all the resources except the last S

resources (where S is a variable that can be customized and S < Q. The utility of partial-contributor is given by:

$$\text{Utility } \mathrm{P_c} = \left(\sum_{k=1}^{i} d_k + \sum_{m=1}^{j} e_m + \sum_{n=1}^{k} f_n \right) - \mathrm{S} \tag{6}$$

- Weak-contributors are those who shared only R resource or no resource at all. This category of user could access only resources that are shared by other weak-contributors. The utility of this contributor can be signed as:

$$\text{Utility } \mathrm{W_c} = \sum_{n=1}^{k} f_n. \tag{7}$$

3.2 Application Layering

The implementation of the proposed model is developed based on SOM that emphasize on applying the open data concept. The concept of open data is emerging as a popular approach in knowledge sharing. The basic requirement for this concept is to open up publicly owned data and information. Those data and information are available in machine readable formats and easy to use by users or providers. It will provide the transparent platform and enable collaboration between users, providers and related stakeholders. The proposed model adopts the concept of open data in the design and the data stored must be complied with the following principles:

- Complete – the data can be definable and agreed among users
- Accessible – the data can be accessed by the widest range of users
- Machine pre-processable – the data must be organized in an agreed structure
- License free – copyright, patent, trademark etc. is not apply to the available data

The proposed model is implemented by using the concept of SaaS and deployed under OpenStack environment. The hierarchy structure of the application model consists of three layers. These layers are: *Resource Layer, Management Layer* and *Service Layer*. The descriptions of each layer are as follows:

- Resource Layer: The layer consists of all the back-end database and resource pool. It supports for computing power, resource library and storage capability of the resource sharing system.
- Management Layer: The layer responsible for coordination and interaction between the resource layer and service layer for resource sharing. It is responsible for the management of all the system components. The prisoners' dilemma and peer-approved incentive mechanism are incorporated in this layer to handle the resource sharing with its justification.

- Service Layer: The layer provides the description of services offered by the reception that include self-service portal. It allows for resources to be uploaded and downloaded by school management or teachers.

4 Result Analysis

Quality evaluation of cloud based application is important before its service can be deployed to user. In this study, the developed prototype is evaluated based on its usability. The usability of the services will be measured using quality of service (QoS) metrics suggested by [2, 11]. A questionnaire has been developed and distributed to 331 teachers for hands-on evaluation session. Four attributes of usability are used: effectiveness, efficiency, learnability and satisfaction. Each attribute is evaluated based on specific questions with five scales: *strongly disagree, disagree, neutral, agree, and strongly* agree. In the context of developed prototype, the definition of each attribute is as follows:

- Effectiveness is the capability of a prototype to enable user for completing specific task accurately and providing a convenient and interactive interface for user to access
- Efficiency is the capability of prototype to enable user for completing specific task in a reasonable time and having a set of features to facilitate user in making interaction effectively
- Learnability is the prototype characteristic that ease the user to accomplish task effectively without specific guidance
- Satisfaction is defined as data security aspect and reporting quality offered by prototype

For the effectiveness evaluation, five (5) questions are examined (Table 1). Based on the results, more than 68% of respondents agreed with the effectiveness of functional in implementing resource sharing. For all questions involved, the average of service effectiveness of quality metrics measured is 79.28%. Regarding to quality metrics number four (*Enhance ability experience in resource sharing*), percent of respondents

Table 1. Effectiveness result

No.	Questions	Respondent feedback (%)				
		SD	D	N	A	SA
1	Allow to use appropriate resources	9.1	10.0	5.4	44.4	31.1
2	Application is accurate and reliable	8.2	8.2	0.3	47.4	36.0
3	Deliver all services at all time	3.3	9.1	0.3	56.8	30.5
4	Enhance ability experience in resource sharing	14.8	13.3	3.3	40.8	27.8
5	Enable to completeness retrieve resources	8.5	10.0	0.0	53.2	28.4

* SD: strongly disagree D: Disagree N: Neutral
A: Agree SA: Strongly agree

agreed with the effectiveness of the prototype is relatively low compared to other quality metrics. This is caused by the respondents still cannot fully adapt to the process flow of the prototype. It is expected that this effectiveness percentage will increase after user having continuous experience and train.

To evaluate the efficiency of the proposed model, six (6) quality metric questions are used. Table 2 tabulates respondent's feedback from the questionnaire given. The result shows that more than 73% of respondents agreed with the efficiency of prototype functions in resource sharing. The average of efficiency quality metrics is 81.35%.

Table 2. Efficiency result

No.	Questions	Respondent feedback (%)				
		SD	D	N	A	SA
1	Convenient to exchange resources with others	5.4	6.9	0.3	46.8	40.5
2	The effect of resource sharing is convincing and commendable	10.6	2.7	0.0	52.3	34.4
3	The process of uploading and downloading resources is fast and efficient	10.9	13.6	0.0	26.0	49.5
4	Efficient enough to be used and guide users through the process	6.3	9.4	0.0	49.2	35.0
5	Guideline resources are appropriate	14.4	4.2	0.0	29.9	51.4
6	Response time of each process is appropriate	17.8	4.8	4.2	41.4	31.7

* SD: strongly disagree D: Disagree N: Neutral
A: Agree SA: Strongly agree

There is also six (6) quality metric questions are used in measuring learnability of the proposed model. Respondent's feedback data is tabulated in Table 3. The result shows that more than 78% of respondents agreed with the learnability of prototype. Besides the friendly interface, the prototype has also provided enough information to guide user in completing tasks. The average for learnability quality metrics is 91.8%.

Table 3. Learnability result

No.	Questions	Respondent feedback (%)				
		SD	D	N	A	SA
1	Interface is easy to learn	8.2	3.9	0.0	40.2	47.7
2	Information is properly presented with clear interpretation	8.8	10.6	1.8	48.3	30.5
3	External resources are relevant and informative	0.0	0.0	0.0	44.4	55.6
4	Provide useful contents related to the shared resources	0.0	0.0	0.0	65.3	34.7
5	Provided guidelines are appropriate and sufficient	0.3	2.7	0.3	61.6	35.0
6	All processes are easy to learn and understand	5.7	1.5	5.1	37.8	49.8

* SD: strongly disagree D: Disagree N: Neutral
A: Agree SA: Strongly agree

In a cloud environment, security and data privacy features are very important. Respondent's satisfaction feedback data is tabulated in Table 4. Seven (7) quality metric questions are used in measuring satisfaction of the proposed model. The result shows that more than 88% of respondents agreed that the prototype has achieved the level of satisfaction. The satisfaction quality metrics average is 93.18%. This indicates that the respondents are confident and satisfied with the prototype as a resource sharing platform.

Table 4. Satisfaction result

No.	Questions	Respondent feedback (%)				
		SD	D	N	A	SA
1	Confident to share the resources	4.2	2.7	0.0	24.8	68.3
2	Satisfied with the security and data privacy	6.6	1.5	0.0	47.1	44.7
3	Fonts and style used in interface design are attractive	0.0	2.7	0.0	36.0	61.3
4	The module integration interfaces are interactive and user friendly	5.7	1.8	4.2	45.9	42.3
5	The model of sharing incentive is satisfied	0.0	4.2	0.0	52.6	43.2
6	The model can be easily understand by inexperience users	3.9	0.0	0.0	59.2	36.9
7	The application is very convenience as it can be used at any time and place	7.9	0.0	2.1	44.1	45.9

* SD: strongly disagree D: Disagree N: Neutral
A: Agree SA: Strongly agree

5 Conclusion

The problem of resource management, inadequate network and computing facilities as well as difficulty in computing services acquisition has become a big concern for Malaysian schools. The article proposed a model for sharing resources among Malaysian schools using SaaS that implemented under cloud platform. The sharing of resources is modeled based on Multi-Person Prisoner's Dilemma with an incentive mechanism that gives justifications and encouragement for user to participate. The cloud based prototype has been developed as a part to complete the concept of software as a service. Furthermore, the service oriented model and open data policy is applied in the architecture of the prototype. The proposed model can save the cost and physical space, particularly for the Malaysian Ministry of Education in providing facilities and infrastructure to schools, teachers and students.

Acknowledgement. This study is funded by Ministry of Higher Education Malaysia (FRGS/1 2015). Special thanks to University of Sultan Zainal Abidin for providing facilities to develop and evaluate the proposed model.

References

1. Askhoj, J., Nagamori, and M., Sugimoto, S. Archiving as a service – a model for the provision of shared archiving services using cloud computing. In: Proceeding of the iConference, pp. 151–158 (2011)
2. Bardsiri, A.K., Hashemi, S.M.: QoS metrics for cloud computing services evaluation. Int. J. Intell. Syst. Appl. **12**, 27–33 (2014)
3. Bell, M.: SOA Modeling Patterns for Service Oriented Discovery and Analysis. Wiley, New Jersey (2010)
4. Clohessy, T., Acton, T., Morgan, L.: Smart City as a Service (SCaaS) – a furute roadmap for E-Government smart city cloud computing initiatives. In: Proceedings of the IEEE/ACM 7th International Conference on Utility and Cloud Computing, pp. 836–842 (2014)
5. Mell, P., Grance, T.: The NIST Definition of Cloud Computing. NIST Special Publication, pp. 800–145 (2011)
6. Moothoor, J., Bhatt, V.: A Cloud computing solution for Universities: Virtual Computing Lab. IBM White Paper, pp. 1–13 (2010)
7. Raganathan, K., Ripeanu, M., Sarin, A., Foster, I.: Incentive mechanism for large collaborative resource sharing. In: Proceeding of the IEEE International Symposium on Cluster Computing and the Grid, pp. 1–8 (2004)
8. Shamsul, A.M., Haryani, S.S.A., Abdulkarem, S., Abdulaziz, A.: Cloud computing in academic institutions. In: Proceeding of the International Conference on Ubiquitous Information Management and Communication (2013)
9. Stancic, H., Rajh, A., Milosevic, I. Archiving-as-a-service. influence of cloud computing on the archival theory and practice. In: Duranti, L., Shaffer, E. (eds.) Proceeding of the Memory of the World in the Digital Age: Digitization and Preservation, UNESCO, pp. 108–125 (2013)
10. Szilagyi, M.N.: An investigation of N-Person Prisoners' Dilemmas. Complex Systems **14**, 155–174 (2003)
11. Tan, C., Liu, K., Sun, L.: A design of evaluation method for SaaS in cloud computing. J. Ind. Eng. Manag. **6**(1), 50–72 (2013)

A Modified Naïve Possibilistic Classifier for Numerical Data

Karim Baati[1,2(✉)], Tarek M. Hamdani[1,3], Adel M. Alimi[1], and Ajith Abraham[4]

[1] REGIM-Lab.: REsearch Groups on Intelligent Machines, University of Sfax, National Engineering School of Sfax (ENIS), BP 1173, Sfax 3038, Tunisia
{karim.baati,tarek.hamdani,adel.alimi}@ieee.org
[2] Esprit School of Engineering, Tunis, Tunisia
[3] Taibah University, College of Science And Arts at Al-Ula, Al-madinah Al-munawwarah, Kingdom of Saudi Arabia
[4] Machines Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, P.O. Box 2259, Auburn, WA 98071, USA
ajith.abraham@ieee.org

Abstract. In this paper, we propose a modified version of the Naïve Possibilistic Classifier (NPC) which has been already suggested to make decision from numerical data. As the former NPC, the modified classifier makes use of the probability to possibility transformation of Dubois et al. in the continuous case in order to estimate possibilistic beliefs. However, unlike the former NPC which uses the product as a fusion operator, the proposed classifier fuses possibilistic beliefs using the generalized minimum-based algorithm which has been recently proposed as an improvement of the minimum operator for combining possibilistic estimates. Experimental evaluations are conducted on 15 numerical datasets taken from University of California Irvine (UCI) and show that the new version of NPC largely outperforms the former one in terms of accuracy.

Keywords: Naïve possibilistic classifier · Possibility theory · G-Min algorithm · Numerical data

1 Introduction

In recent years, possibilistic classifier begins to attract attention among Bayesian-like classification models. Indeed, this classifier presents the advantage to combine the structure of the Naïve Bayes Classifier (NBC) as a good classifier that can often outperform more sophisticated models [1] with the possibility theory as a powerful framework for handling imperfect data [2].

Different versions of naïve possibilistic classifier which are based on the probability to possibility transformation rule of Dubois et al. [3] have been proposed. Indeed, two versions based on this transformation in the continuous case and dealing with numerical features [4], namely, Naïve Possibilistic Classifier (NPC)

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_41

for numerical data and Flexible Naïve Possibilistic Classifier (FNPC) for numerical data, have been suggested in [5]. Further, a version of naïve possibilistic classifier based on the transformation rule of Dubois et al. in the discrete case [3] and devoted to categorical data has been successfully used in [6,7] to make decision from the medical data of the UCI lymphography dataset.

Experimental results in [5] have been conducted on 15 numerical datasets and have shown that the FNPC outperforms all Bayesian-like classifiers for numerical data. However, even based on the same principle as FNPC, the NPC was among the worst-ranked classifiers. Moreover, for both classifiers, the product has been used in the aggregation step since it provides better results by comparison to the other classical fusion rule, namely the minimum. Therefore, to enhance the quality of decision within NPC, it seems to be mandatory to use a more suitable approach for the fusion of possibilistic estimations.

Recently, a new fusion strategy based on a Generalized Minimum-based (G-Min) algorithm has been suggested as an enhancement of the minimum operator in order to fuse possibilistic beliefs stemming from heterogeneous medical data [8]. Experimental evaluations in [8] have demonstrated the efficiency of the G-Min algorithm with regard to classical fusion operators for possibilistic beliefs namely, the product and the minimum.

The main focus of the work is to enhance the classification accuracy of the NPC proposed in [5] when dealing with numerical data by using the G-Min algorithm suggested in [8] instead of the product rule.

The rest of the paper is structured as follows. Related works are discussed in Sect. 2. Then, in Sect. 3, we restate the theoretical background of the possibilistic framework. Next, Sect. 4 provides details of the proposed classifier. Section 5 is devoted to show and discuss the experimental results. Finally, Sect. 6 presents some concluding remarks.

2 Related Work

In this section, we present main previous studies that have combined the Bayesian structure with the possibilistic data representation.

The first version of Bayesian-like possibilistic classifier was proposed in [9] in order to deal with imprecise training data. However, no efficient method was proposed in this work for the elicitation of possibilistic distributions. Indeed, it was stated in [9] that the procedure proposed to estimate possibility distributions which is based on the computation of the maximum-based projection [10], may construct pathological cases [9].

Later, Haouari et al. [11] have proposed a naïve possibilistic network classifier which models imperfect attribute values when no prior knowledge is available. In this particular context, possibility distributions are built using the partial ignorance which is expressed by an expert.

In [12], authors propose an efficient algorithm which makes use of the Jeffrey's rule in order to reconsider possibilistic knowledge included by a naïve product-based possibilistic network classifier on the basis of uncertain inputs.

The proposed algorithm presents the main advantage of being able to process the classification task in polynomial time with respect to the number of features.

As mentioned earlier, in [5], the authors propose two kinds of possibilistic classifiers for numerical data. In the first one, authors extend the classical and flexible Bayesian classifiers by applying a probability-possibility transformation to Gaussian distributions. In the second one, authors directly express data in possibilistic formats using the idea of proximity between data values. Experiments in [5] have shown the good performance of the flexible possibilistic classifier when dealing with numerical data.

In a more recent work [13], the authors extend the probability-to-possibility transform based classifiers which were proposed in [5] in order to deal with uncertainty and imprecision in data modeling. Experimental results reported in [13] have confirmed the robust behavior of the probability-to-possibility transform-based classifiers when they treat imperfect data.

3 Possibilistic Framework

In order to recall some basics of the possibilistic framework, the following notations are considered in this section as well as in the remainder of this paper:

$C = \{c_1, c_2, ..., c_j, ..., c_n\}$: an exhaustive and exclusive universe of discourse of classes.

$A = \{a_1, a_2, ..., a_i, ..., a_m\}$: a set of attributes which may stand for either categorical or numerical values.

v_{it}: the value taken by a given attribute a_i during test.

3.1 Possibility Theory

Possibility theory, introduced by Zadeh [14] and then developed by Dubois and Prade [15] is a fusion theory based on fuzzy sets theory and devoted to represent and combine imperfect information in a qualitative or quantitative way. Information imperfections treated by possibility theory may represent the uncertainty due to variability of observations, the uncertainty due to poor information, the information ambiguity, the information imprecision, etc. [2].

At the semantic level, the basic function in possibility theory is a possibility distribution denoted as π which assigns to each possible class c_j from C a value in [0, 1]. The possibility value assigned to a class c_j stands for plausibility i.e. the belief degree that this class is the right one.

By convention, $\pi(c_j) = 1$ means that c_j is totally possible and if $\pi(c_j) = 0$, c_j is considered as impossible. Intermediary values distinguish values which are more possible than others. Finally, note that in the normalization version of a possibility distribution, we require that at least one class of C is totally possible.

From the possibility distribution, two dual measures, namely possibility Π and necessity N, may be assigned to each subset E of the power set 2^C as follows:

$$\begin{aligned} \Pi(E) &= max[\pi(c_j)], c_j \in E \\ N(E) &= min[1 - \pi(c_j)], c_j \notin E \end{aligned} \tag{1}$$

Between these two measures there is a duality relation defined by:

$$N(E) = 1 - \Pi(\bar{E}), E \subset 2^C \tag{2}$$

$\Pi(E) = 1$ corresponds to the situation where at least one singleton $c_j \in E$ has a possibility degree equal to 1 and it is likely to be the right class but without being necessary. $N(E) = 1$ corresponds to the situation where $\Pi(\bar{E}) = 0$, meaning that the right class certainly does not belong to subset $\bar{E}$, so it is certainly a member of the subset $E \subset 2^C$.

On the other hand, as being based on fuzzy sets theory, possibility theory may recall all the panoply of fuzzy combination rules in order to fuse possibility estimates of a given class.

3.2 Conditional Possibility

Possibilistic conditioning corresponds to revising an initial possibility distribution when a new information becomes available. In possibility theory, conditioning may be defined through a counterpart of the Bayes rule. Indeed, it can be stated for two subsets E and F of 2^C by:

$$\Pi(E \cap F) = \Pi(E|F) * \Pi(F) \tag{3}$$

where * is a combination operator which is commonly chosen as the minimum or the product [15].

3.3 Naïve Bayes Style Possibilistic Classification

Similarly to Bayesian classification, possibilistic classification is based on the aforementioned possibilistic version of the Bayes rule and can be defined by:

$$\pi(c_j|a_1 = v_{1t}, ..., a_m = v_{mt}) = \frac{\pi(a_1 = v_{1t}, ..., a_m = v_{mt}|c_j) * \pi(c_j)}{\pi(a_1 = v_{1t}, ..., a_m = v_{mt})} \tag{4}$$

By assuming that there is no a priori knowledge about classes and the input vector to be classified, we can take $\pi(c_j) = 1$ and $\pi(a_1 = v_{1t}, ..., a_m = v_{mt}) = 1$. Moreover, as in naïve Bayesian combination, naïve possibilistic classification assumes that attributes $\{a_i\}$ $(\forall \ \ 1 \leq i \leq m)$ are independent. The conditional joint possibility $\pi(a_1 = v_{1t}, ..., a_m = v_{mt}|c_j)$ is equal in that case to the fusion of conditional possibility estimations stemming from each single attribute a_i. Therefore, in a context characterized by no a priori knowledge about classes and input vector as well as independent attributes, Eq. 4 becomes [9]:

$$\pi(c_j|a_1 = v_{1t}, ..., a_m = v_{mt}) = \pi(a_1 = v_{1t}|c_j) * ... * \pi(a_m = v_{mt}|c_j) \tag{5}$$

As stated for the conditioning rule, * may be taken as either the product or the minimum [16,17].

In practice, given a new instantiation $\{a_1 = v_{1t}, ..., a_m = v_{mt}\}$, we must establish a matrix Π of possibilistic estimates in order to perform the product-based or the minimum-based classification. This matrix is defined as follows:

$$\Pi = \{\pi(i,j)\} = \{\pi(a_i = v_{it}|c_j)\} \tag{6}$$

$\forall \ \ 1 \leq i \leq m$ and $\forall \ \ 1 \leq j \leq n$

Lastly, the final decision stands for the class c^* for which Eq. 5 yields the highest degree of possibility belief:

$$c^* = arg \quad \max_j(\pi(1,j) * ... * \pi(m,j)) \tag{7}$$

4 Proposed Classifier

In order to perform the classification of numerical data, the proposed classifier rests on two main steps, namely the estimation of possibility beliefs and then the fusion of these beliefs using the G-Min algorithm. In the following, technical details of these two steps are detailed.

4.1 Estimation of Possibility Beliefs

To estimate possibilistic beliefs from numerical data, attribute values are first normalized as follows:

$$a_{in} = \frac{a_i - min(a_i)}{max(a_i) - min(a_i)} \tag{8}$$

where min and max represent respectively the minimum and the maximum value of the attribute a_i over the training set. For simplicity reasons, we consider that all attribute value a_i's are normalized and then refer to the corresponding values a_{i_n}.

As in NBC for numerical data [1], we can assume that each attribute a_i is a random variable which is normally distributed over each class c_j [5]. Thus, a Gaussian distribution $g_{ij} = g(a_i, \mu_{ij}, \sigma_{ij})$ is computed for each feature a_i with regard to each class c_j. For this Gaussian, parameters μ_{ij} and σ_{ij} represent, respectively, the mean and the standard deviation of the variable a_i for the class c_j and they are estimated from training samples.

Based on the obtained Gaussian, we can estimate the probability $P(I_{ai})$ of the confidence interval centered at μ_{ij} as [5]:

$$P(I_{ai}|c_j) = 2 * G(a_i, \mu_{ij}, \sigma_{ij}) - 1 \tag{9}$$

where G is a Gaussian cumulative distribution which may be assessed using the table of the standard normal distribution.

Afterward, we make use of the probability-possibility transformation of Dubois et al. in the continuous case [3] in order to estimate the possibility $\pi(a_i = v_{it}|c_j)$ of each test value as follows [5]:

$$\begin{aligned} \pi(a_i = v_{it}|c_j) &= 1 - (2 * G(v_{it}, \mu_{ij}, \sigma_{ij}) - 1) \\ &= 2 * (1 - G(v_{it}, \mu_{ij}, \sigma_{ij})) \end{aligned} \tag{10}$$

4.2 Fusion of Possibilistic Beliefs

To fuse possibilistic beliefs, the minimum rule allocates the final decision to the class which satisfies:

$$c^* = arg \quad \max_j(\min_i \pi(i,j)) \tag{11}$$

where $\Pi = \{\pi(i,j)\}$ $(1 \leq i \leq m$ and $1 \leq j \leq n)$ the matrix of possibilistic estimates which has been defined in Eq. 6.

In many cases, the minimum-based possibilistic combination is likely to lead to a final decision that may have very close possibility estimate to other alternatives. In such situation, the quality of decision may be seriously altered since the final class is likely to be inaccurate. In order to resolve this problem, the G-Min algorithm has been proposed in [8] in order to avoid the ambiguity between the final decision and the rest of classes and hence to find a decision with a possibility estimate widely away from other alternatives.

The G-Min algorithm requires the matrix Π of possibilistic estimates and it is based on two main steps. The first aims to build a set of possible decisions whereas the second aims to filter this set in order to find a final class with a high score of reliability [8].

In the first step, columns of the matrix of possibilistic estimations Π are sorted increasingly and links with attributes are left out. Then, rows of the resulting matrix are sorted decreasingly and the argument (the class) related to each value is retained. Afterward, a decision set *dset* is established using the first row of the output matrix Π after the two sorting steps [8].

In the second step, the next row of Π is treated by picking possibility values of respectively each of the selected classes on *dset*. Based on the new possibility values, we eliminate each class of the decision set *dset* that have not a possibility value within the threshold α with respect to the maximum of the updated possibility values. Each time we find *dset* different from a singleton, we deal with the next row of Π in order to try to further filter the decision set. The second step is repeated until the decision set *dset* is a singleton or we are in the last row of the matrix Π. The final class is then assigned to the one having the maximum value of possibility in the decision set *dset* [8].

More details about the proposed algorithm may be found in Algorithm 1 [8].

Further, in order to more explain the proposed algorithm, we exhibit a numerical example in Fig. 1 in which we have taken $\alpha = 0.1$. In this example, cardinality of the decision set has decreased along rows and a unique class has been obtained at the end. One could plainly see that in this example, the final decision differs from the one that would be chosen if the minimum operator has been employed.

5 Experimental Evaluations

In order to compare performances of respectively the product-based NPC and the G-Min-based NPC to classify numerical data, experiments are conducted on 15 datasets provided by UCI. These datasets stand for the same ones used in [5].

Algorithm 1. G-Min Algorithm

Require: Matrix of possibilistic estimations Π
Ensure: Final decision c^*
  $\Pi \leftarrow sort(\Pi,'rows','increase')$
  $\Pi \leftarrow sort(\Pi,'columns','decrease')$
  $q \leftarrow 1$
  $dset[q] \leftarrow arg(\pi(1,1))$
  $c^* \leftarrow arg(\pi(1,1))$
  for $j \leftarrow 2$ *to* n **do**
    if $\pi(1,1) - \pi(1,j) < \alpha$ **then**
      $q \leftarrow q + 1$
      $dset[q] \leftarrow arg(\pi(1,j))$
    end if
  end for
  $card \leftarrow cardinality(dset)$
  $i \leftarrow 2$
  while $card \neq 1$ *and* $i \neq m$ **do**
    for $q \leftarrow 1$ *to* $dset.size$ **do**
      $poss[q] \leftarrow arg^{-1}(dset[q], i)$
    end for
    $max \leftarrow maximum(poss)$
    for $q \leftarrow 1$ *to* $poss.size$ **do**
      if $max - poss[q] \geq \alpha$ **then**
        $dset \leftarrow dset.remove(arg(poss[q])$
        $c^* \leftarrow max$
      end if
    end for
    $card \leftarrow cardinality(dset)$
    $i \leftarrow i + 1$
  end while

Characteristics of these datasets are illustrated in Table 1. As shown in Table 1, the number of cases ranges from 150 to 3050, the number of attributes from 4 to 37 and the number of classes from 2 to 26. Therefore, a wide variety of problems is represented.

During experiments, we have carried a 10-cross validation and as in [5], we have used the standard Percent of Correct Classification (PCC) to assess performances of the two classifiers. PCC is defined as:

$$PCC = \frac{number \;\; of \;\; well \;\; classified \;\; instances}{total \;\; number \;\; of \;\; instances} * 100 \tag{12}$$

Moreover, we have taken $\alpha = 0.1$ in order to apply the G-Min fusion algorithm.

Table 2 shows that the G-Min-based classifier is the best-ranked on 11 among the 15 benchmarks namely, "Letter", "Segment", "Yeast", "Satellite", "Magic", "German", "Transfusion", "Iosophere", "Glass", "Wine" and "Iris". Furthermore, the new classifier is the best in terms of average rank (Table 2).

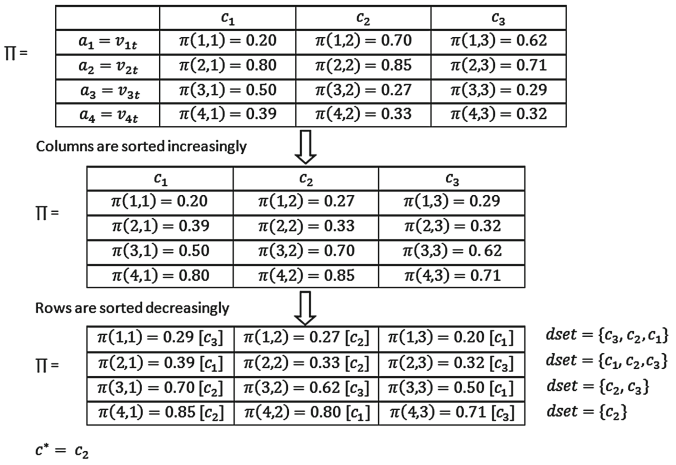


Fig. 1. A numerical example for the application of the G-Min algorithm

Table 1. Details of numerical datasets used in the experiments

Dataset	Instances	Attributes	Classes
Letter	3050	17	26
Segment	1500	20	7
Yeast	1484	9	10
Satellite Image	1090	37	6
Magic Gamma Telescope	1074	10	2
German	1000	25	2
Diabetes	768	7	2
Transfusion	748	4	2
W.B Cancer	699	8	2
Iosophere	351	35	2
Ecoli	336	8	8
Heart	270	14	2
Glass	214	10	7
Wine	178	13	3
Iris	150	4	3

Table 2. PCC obtained with the two versions of NPC given as the mean and the standard deviation

	NPC (Prod.)	NPC (G-Min)
Letter	60.42 ± 3.24 (2)	**73.46 ± 2.42(1)**
Segment	71.47 ± 4.15 (2)	**81.31 ± 1.77 (1)**
Yeast	49.67 ± 4.87 (2)	**69.97 ± 5.13(1)**
Satellite	88.26 ± 2.62 (2)	**96.27 ± 2.91(1)**
Magic	59.24 ± 7.09 (2)	**72.80 ± 3.29 (1)**
German	66.40 ± 3.97 (2)	**70.30 ± 3.19(1)**
Diabetes	**69.01 ± 3.99 (1)**	68.24 ± 3.72 (2)
Transfusion	61.67 ± 6.60(2)	**64.88 ± 6.80 (1)**
W.B Cancer	**95.03 ± 2.26 (1)**	92.85 ± 2.06 (2)
Iosophere	58.40 ± 10.95 (2)	**66.97 ± 7.44 (1**
Ecoli	**83.37 ± 4.46(1)**	80.06 ± 7.23 (2)
Heart	**84.04 ± 8.77 (1)**	81.85 ± 7.50 (2)
Glass	49.18 ± 11.80 (2)	**57.53 ± 8.99 (1)**
Wine	94.37 ± 5.56 (2)	**95.33 ± 3.89 (1)**
Iris	95.33 ± 6.00 (2)	**96.00 ± 6.21 (1)**
Average rank	1.73	**1.26**

6 Conclusion

This paper has proposed a new version of the NPC classifier which has been already suggested to deal with numerical data. The new version hybridizes the capability of the former NPC to estimate possibilistic beliefs from numerical data with the efficiency of the G-Min as a novel algorithm for the fusion of possibilistic beliefs. Experimental results have shown that the proposed classifier largely outperforms the former NPC in terms of accuracy.

The good performance of the proposed version may be encouraging for two major reasons. First, this version may be kept and strengthened by a feature selection stage in order to establish a more efficient classification technique for numerical data. On the other side, the good behavior of the proposed G-Min-based NPC may be promising in the sense that this classifier can be joined to one of the previously proposed possibilistic classifiers for categorical data in order to treat uncertainty stemming from mixed numerical and categorical data.

Acknowledgment. The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program.

References

1. Langley, P., Iba, W., Thompson, K.: An analysis of Bayesian classifiers. In: Proceedings of AAAI, pp. 223–228 (1992)
2. Khaleghi, B., Khamis, A., Karray, F.O., Razavi, S.N.: Multisensor data fusion: a review of the state-of-the-art. Inf. Fusion **14**(1), 28–44 (2013)
3. Dubois, D., Foulloy, L., Mauris, G., Prade, H.: Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. Reliable Comput. **10**(4), 273–297 (2004)
4. Baati, K., Kanoun, S., Benjlaiel, M.: Différenciation d'écritures arabe et latine de natures imprime et manuscrite par approche globale. In: Colloque International Francophone Sur l'Ecrit et le Document (CIFED) (2010)
5. Bounhas, M., Mellouli, K., Prade, H., Serrurier, M.: Possibilistic classifiers for numerical data. Soft Comput. **17**, 733–751 (2013)
6. Baati, K., Hamdani, T.M., Alimi, A.M.: Diagnosis of lymphatic diseases using a naïve bayes style possibilistic classifier. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 4539–4542. IEEE (2013)
7. Baati, K., Hamdani, T.M., Alimi, A.M., Abraham, A.: A Modified Naïve Bayes Style Possibilistic Classifier for the Diagnosis of Lymphatic Diseases. In: Proceedings of the 16th International Conference on Hybrid Intelligent Systems. Springer (2016)
8. Baati, K., Hamdani, T.M., Alimi, A.M., Abraham, A.: A new possibilistic classifier for heart disease detection from heterogeneous medical data. Int. J. Comput. Sci. Inf. Secur. **14**(7), 443–450 (2016)
9. Borgelt, C., Gebhardt, J.: A naïve bayes style possibilistic classifier. In: Proceedings of the 7th European Congress on Intelligent Techniques and Soft Computing (1999)
10. Borgelt, C., Kruse, R.: Efficient maximum projection of database induced multivariate possibility distributions. In: Proceedings of the 7th IEEE International Conference on Fuzzy Systems, pp. 663–668 (1988)
11. Haouari, B., Ben Amor, N., Elouedi, Z., Mellouli, K.: Naïve possibilistic network classifiers. Fuzzy Sets Syst. **160**(22), 3224–3238 (2009)
12. Benferhat, S., Tabia, K.: An efficient algorithm for naive possibilistic classifiers with uncertain inputs. In: Greco, S., Lukasiewicz, T. (eds.) SUM 2008. LNCS (LNAI), vol. 5291, pp. 63–77. Springer, Heidelberg (2008). doi:10.1007/978-3-540-87993-0_7
13. Bounhas, M., Hamed, M.G., Prade, H., Serrurier, M., Mellouli, K.: Naïve possibilistic classifiers for imprecise or uncertain numerical data. Fuzzy Sets Syst. **239**, 137–156 (2014)
14. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst. **1**(1), 3–28 (1978)
15. Dubois, D., Prade, H.M., Farreny, H., Martin-Clouaire, R., Testemale, C.: Possibility Theory: An Approach to Computerized Processing of Uncertainty 2. Plenum Press, New York (1988)
16. Baati, K., Hamdani, T.M., Alimi, A.M.: Hybrid naïve possibilistic classifier for heart disease detection from heterogeneous medical data. In: Proceedings of the 13th International Conference on Hybrid Intelligent Systems, pp. 235–240. IEEE (2013)
17. Baati, K., Hamdani, T.M., Alimi, A.M.: A modified hybrid naive possibilistic classifier for heart disease detection from heterogeneous medical data. In: Proceedings of the 6th International Conference on Soft Computing and Pattern Recognition, pp. 353–35. IEEE (2014)

Combinatorial Structural Clustering (CSC): A Novel Structural Clustering Approach for Large Scale Networks

Liang Chen, Hongbo Liu(✉), Weishi Zhang, and Bo Zhang

School of Information Science and Technology,
Dalian Maritime University, Dalian 116026, China
{lhb,teesiv}@dlmu.edu.cn

Abstract. With the development of the brain cognition science and big data technologies, effective graph clustering is a key technique to uncover the brain mechanism, especially in resting state functional connectivity analysis. In this paper, a combinatorial structural clustering (CSC) algorithm is proposed for large scale networks. A structural similarity feature from adjacency structures of outliers and hubs is introduced to brain functional connectivity networks. Experimental results illustrate that our approach has some advantages compared with SCAN.

Keywords: fMRI · Combinatorial structural cluster · Complex network · Big data

1 Introduction

With the development of the brain cognition science and big data technologies, the complex network theory has been widely used in functional connectivity (FC) which utilizes the self-spontaneous synchronism of brain to construct the topological graph.

Effective clustering approaches are important for the brain network detection, especially in resting state functional connectivity. The structural clustering algorithm for networks (SCAN) is one of the most successful clustering methods [1]. Comparing to the approaches based on degree or functional connectivity density, the SCAN emphasises the relationship between nodes rather than a "global score". This topological similarity method is more effective relatively [2,3].

To utilize the region of interest strategy the similarity of functional pattern between different brain regions could be transformed into the topological graph and the functional connectivity modular detection could be treated as a graph partition problem. The graph clustering is a kind of the solution of graph partition problem [4]. In previous studies, the brain network modular analysis are mainly based on k-means [5] or center-based cluster approaches [6,7] which may lead to be divided into several cluster with similar size. These algorithm are

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_42

usually supervised with the cluster number parameter. If the cluster number is unfit with the data then the nodes in a same cluster may be divided into different clusters while the nodes in different clusters into a same cluster. So if the number of clusters is unknown or the ROI/voxels number is imbalance for different subgraph then the center-based cluster approach might unsuit for these data. Unfortunately it usually occurs in fMRI data. It is primarily caused by the mechanism of k-means and other center-based approaches based on center of mass. On the other hand, structural clustering algorithm for networks (SCAN) utilize a topological similarity to cluster the network [1], which is a local complex network metric to measure the similarity between nodes with high correlation coefficient. Similar to density-based clustering, the main concept in the SCAN is that densely connected adjacent nodes should be in the same cluster. However unlike the traditional approaches, SCAN can successfully finds not only clusters but also hubs and outliers at not insignificant cost [8]. But in fMRI data analysis, if the adjacency graph is constructed according to the similarity between among the data sets, such as time series of ROIs, then the hubs turn out to be the buffer between different clusters. Therefore, a novel approach is proposed whose starting point is the sampling distribution where the vertices in graph may be divided into several part as the subgraph according to the density gap of edges between nodes.

In this paper, a combinatorial structural cluster (CSC) algorithm for fMRI functional connectivity is proposed. Although SCAN's effectiveness in detecting subnetworks, there still exists a weakness of feature, structural similarity δ. $\delta(u, v)$ between vertices u and v is described as (1)

$$\delta(u, v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u)|\,|\Gamma(v)|} \tag{1}$$

where $\Gamma(u)$ is the set of adjacent vertices for u. $|\Gamma(u)|$ is the number of vertices for $\Gamma(u)$. Then we assume that p and q are the probability of a nodes $u \in C_i$ connected with the inner nodes and inter nodes respectively, and C is the whole cluster result which include outlier, hubs and cluster nodes. Then the unbiased estimation of $|\Gamma(u)|$ could be determined as follow.

$$|\Gamma(u)| = p|C_i| + q|C \setminus C_i| \tag{2}$$

where $C \setminus C_i$ is the nodes set which contains nodes in C expect the nodes in C_i. While $u \in C_i \wedge v \in C_j \wedge C_i \neq C_j$ the unbiased estimation of intersection $\Gamma(u) \bigcap \Gamma(v)$ could be estimated as

$$\left|\Gamma(u) \bigcap \Gamma(v)\right| = \begin{cases} pq\,|C_i \bigcup C_j| + q^2\,|C/\{C_i \bigcup C_j\}| & C_i \neq C_j \\ p^2\,|C_i \bigcup C_j| + q^2\,|C/\{C_i \bigcup C_j\}| & C_i = C_j \end{cases} \tag{3}$$

where C is the whole clusters for graph G, $u \in C_i$ and $v \in C_j$ and $|C_i|$ is the nodes number of cluster C_i. Then $\delta(u, v)$ could be illustrate as (4).

$$\delta(u, v) = \begin{cases} \frac{pq|C_i \bigcup C_j| + q^2|C/\{C_i \bigcup C_j\}|}{\sqrt{(p|C_i| + q|C/C_i|)(p|C_j| + q|C/C_j|)}} & C_i \neq C_j \\ \frac{p^2|C_i| + q^2|C/C_i|}{(p|C_i| + q|C/C_i|)} & C_i = C_j \end{cases} \tag{4}$$

Therefore, the distribution of δ is influenced by the proportion between inner clustering edges and inter clustering edges. If the sizes of clusters in clustering results are imbalanced, then distribution of δ for the vertices of small clusters and distribution of δ for the vertices of large clusters are different. Even if the density of edges in small clusters are similar with that in large clusters. So, another type of feature is introduced in CSC in order to improve the accuracy of clustering result.

2 CSC Framework

In this section, the CSC framework includes 3 steps, i.e. the optimized structural cluster analysis step in Algorithm 1, the preliminary correction step in Algorithm 2 and the depth correction step included by Algorithm 3. The CSC framework is shown in Fig. 1.

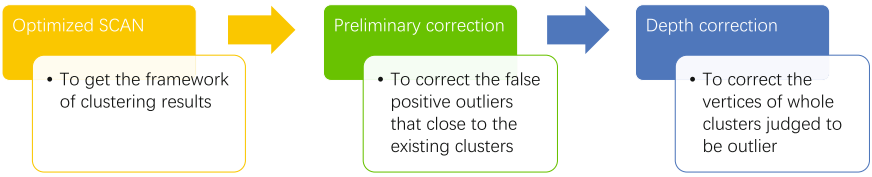


Fig. 1. CSC framework

In the optimized structural cluster analysis step, the clustering result are returned as SCAN and calculation approach is optimized via using math kernel library (MKL), and the Dijkstra approach in order to save the computational cost which is presented in Algorithm 1. The functional connectivity of resting state fMRI data are input as the adjacency matrix W and the clustering result $Cluster_ID$ is result as a logical matrix which illustrate the cluster index respect to.

In the preliminary step, the false positive outliers that close to the existing clusters in the clustering result of the last step are updated in this step which is presented in Algorithm 2. If there are some vertices closed to the existence cluster but still judged to be outlier for the mechanism of δ in the cluster result via Algorithm 1. In order to improve the accuracy of clustering result, we introduce a new feature $outlier_feature$ to preliminary correct the clustering result of SCAN. Assume that u is an outlier in clustering result. Then $outlier_feature$ is (5)

$$outlier_feature(u, c) = \frac{|\Gamma(u) \bigcap (c)|}{ln^{|c|}} \tag{5}$$

where c is one of the existent clusters set in clustering results. In this step, the adjacent matrix W and the clustering result $Cluster_ID$ via Algorithm 1 are input in this algorithm.

ALGORITHM 1. Optimized structural cluster analysis

Input: adjacency matrix W, δ threshold ϵ;
Output: $Cluster_ID$;
Calculate the δ matrix for each vertices pair with MKL;
$W = W + eye(size(W))$;$Degree = sum(W, 1)$;
$\delta = (diag(Degree)^{-0.5} * (WW^T) * diag(Degree)^{-0.5} - eye(size(W))) . * (W - eye(size(W)))$;
Find the vertex which could be acted as the seed vertex and construct the direct graph for $Dijastra$ approach;
$type = (sum(delta > \epsilon, 2) >= 2)$;
$\delta_{direct} = zeros(size(W))$; $\delta_{direct}(type, :) = (delta(type, :) > \epsilon)$;
for *each vertex* $u \in V$ **do**
  if *u is seed vertex and hasn't been concluded in any previous clusters* **then**
    Find the vertex which could reach u in the direct weighted graph δ and return the vertex list in the logical vector;
    $Cluster_num++$;
    $Dist = Dijastra(\delta_{direct}, u, 1 : length(type))$;
    $Cluster_ID(:, Cluster_num) = (Dist < \infty)$;
  end
end
if $Cluster_num > 0$ **then**
  $Cluster_ID = Cluster_ID(:, 1 : Cluster_num)$;
else
  $Cluster_ID = zeros(size(W, 1), 1)$;
end

In the deep correction step, the false positive outliers could be clustered into new clusters are corrected in this step which is illustrated as Algorithm 3, while the vertices of whole clusters judged to be outlier for the mechanism of δ in the cluster result via algorithm. Then we could further correct the clustering result with necessity of dense structural among vertices illustrated in Algorithm 3.

3 Experimental Setting, Result and Discussion

3.1 Experimental Setting

In order to verify the efficiency of CSC, both series of synthetic topological graph data and resting state functional connectivity datasets are used in this section.

In the first part of simulation, we analyze the synthetic data, 50 independent repeated trials are adapted for each parameter traverse. In the graph with balance cluster size, there are 200 vertices and 6 cluster with the same size of 30. And in the graph with imbalance cluster size, there are 640 vertices and 5 cluster with the size of 20, 40, 80, 160 and 320 are included. For the graph with the mechanism of ER random graph, the inner cluster property p and inter cluster property q are set 0.25 and 0.01 respectively. For the graph with the

ALGORITHM 2. Preliminary correction

Input: adjacency matrix W, $Cluster_ID$, parameter ε;
Output: the preliminary corrected $Cluster_ID$;
Call Algorithm 1 to calculate the clustering result logical matrix $Cluster_ID$;
Find the outlier from $Cluster_ID$ and record in $outlier_list$;
$outlier_feature = zeros(size(outlier_list, 1), size(Cluster_ID, 2))$;
for *each vertex* $i \in outlier_list$ **do**
    $tmp = zeros(size(outlier_list, 1), size(Cluster_ID, 2))$;
    $tmp(Adjacency(outlier_list(i), :), :) = Cluster_ID(Adjacency(outlier_list(i), :), :)$;
    for *each cluster* j **do**
        if $sum(any(Adjacency(tmp(:, j), tmp(:, j))))$ **then**
            $outlier_feature(i, j) = sum(tmp(:, j))$;
        end
    end
    $outlier_feature(i, :) = outlier_feature(i, :)./log(sum(Cluster_ID))$;
end
Correct the outlier closed to the vertices in clusters **for** *each vertex* $i \in outlier_list$ **do**
    if *any* $outlier_feature$ *for vertex* i *greater than* ε **then**
        Add outlier i to the cluster of biggest $outlier_feature$;
        $Cluster_ID(outlier_list(i), argmax) = true$;
    end
end

mechanism of scale free network, the inter cluster part is the sparse ER random graph whose property has been set 0.01 and the inner cluster part is generated via the BA approach whose parameter $|u_0|$ and core size are set $\{5, 7, 9, 13, 18\}$ and $\{5, 7, 9, 13, 18\}$ respectively [9]. The clustering performance is examined via adjusted rand index (ARI) [10] which is also adapted in other work [1]. The calculation cost will aslo be examined in order to test the effect of algorithm optimization. In this experiment, the vertices number are set $\{100, 200, \cdots, 1000\}$ while the cluster size of each clusters is increased along with the whole vertices number linearly. The graph with imbalance and balance cluster size are repeated 25 times respectively. The mechanism of topological graph is ER random graph.

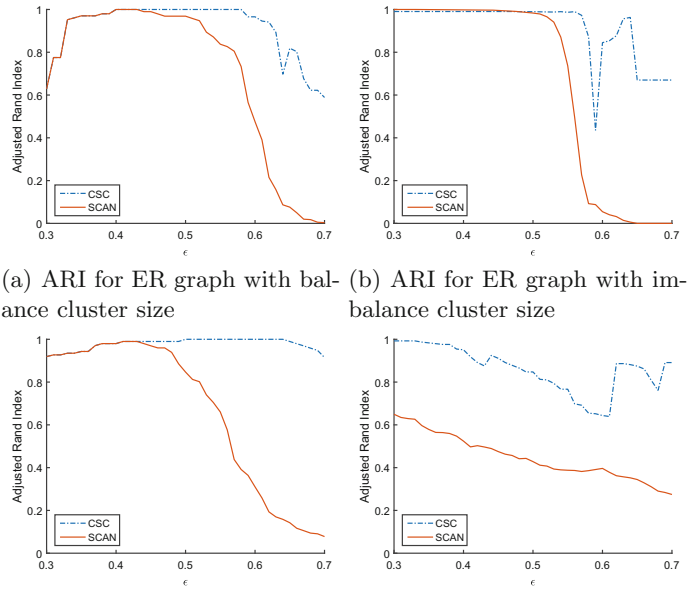
In the second part of resting state, datasets are collected with the assistance of Dalian Maritime University and Affiliated Zhongshan Hospital of Dalian University. During resting state scans, black screen with a white dot was shown for the subjects who were instructed to keep eyes open while blinking normally, keep head relax, still and stay awake throughout the data acquisition [11]. The scanning time is $360s$ [12]. Functional image preprocessing was performed using the statistical parameter mapping (SPM) [13] and Data Processing Assistant for Resting-State fMRI (Dparsf) [14] whose version are SPM12 and Dparsf V3.2 respectively.

ALGORITHM 3. Deep correction

Input: $Cluster_ID$, Adjacency matrix W, constraint φ
Output: the depth corrected $Cluster_ID$
Call Algorithm 2 to calculate the corrected clustering result $Cluster_ID$;
Find the remain outlier from $Cluster_ID$ and record in $outlier_list$;
$W = W(outlier_list, outlier_list)$;
Find all the edge in the subgraph of $outlier_list$;
for *each edge* k **do**
    if *The edge* k *hasn't been processed* **then**
        Extract the cluster C_{single} included edge k:
        To find one of their share vertex which is close to the others;
        Create a new set $Edge$ for C_{single} which includes the 3 edges of these vertices;
        Initialize the traversal edge $t_{edge} = 0$ and the length len_{edge} of queue $Edge$;
        while $t_{edge} < len_{edge}$ **do**
            $t_{edge} = t_{edge} + 1$;
            Find the share vertices set idx_k between two vertices of $t_{edge} \in Edge$; **if** idx_k *is not empty* **then**
                Update $idx_{potential}$: Find the vertices closed to each other in idx_k but not in current cluster;
            end
            if φ **then**
                $C_{new} = C_{single}$;
                Extract all the vertices in $idx_{potential}$ which has neighbours in C_{new} except the two points for the $t_{edge}-$th edge in queue $Edge$ into C_{new}, until there is no more vertex to be extracted from $idx_{potential}$;
                Put the vertices in C_{new} but not in C_{single} to $idx_potential$;
            end
            for *each vertex in* $idx_{potential}$ **do**
                Add the edge that $idx_{potential}(i)$ links to the vertices in C_{single} to the end of queue $Edge$;
                Update $C_{single}(idx_{potential}(i)) = true$;
            end
            Update $len_{edge} = length(Edge)$;
        end
        Record the edge whose two endpoints are both in C_{single} to be true. $proc_{Edge}(C_{single}, C_{single}) = true$;
        if *The number of vertices in cluster* C_{single} *is more than* 3 **then**
            Create a new cluster ID for the vertices contained in C_{single};
            $Cluster_ID(outlier_list, ++clusterID) = C_{single}$;
        end
    end
end

3.2 Result and Discussion

Simulation Result. In this section, the ER graph and the scale free network are applied in the simulation data generation which are both the major structure of complex network. Erdos-Renyi Random Graph (ER random graph) was first introduced in 1959 which is a particular case for complex network applied in amount of studies. Scale free network is another kind of complex network whose degree distribution follows a power law, or at least asymptotically. Many networks have been reported to be scale-free. Preferential attachment and the fitness model in graph theory have been proposed as mechanisms to explain conjectured power law degree distributions in real networks. It has been applied not only in social networks [15] but also in brain connectivity analysis [16].



(a) ARI for ER graph with balance cluster size

(b) ARI for ER graph with imbalance cluster size

(c) ARI for SF graph with balance cluster size

(d) ARI for SF graph with imbalance cluster size

Fig. 2. Adjusted rand index comparison between CSC and SCAN

The relationship of the mean adjusted rand index trends between CSC and SCAN are shown in Fig. 2 where the $x-$axis is the parameter ϵ is from 0.3 to 0.7 with the step of 0.01, $y-$axis is the value of adjusted rand index. Synthesizes each kind of situation, the approach, CSC, proposed in this paper have better clustering accuracy than SCAN. Then, the time cost between the CSC and SCAN are shown as Fig. 3. The vertices number are set $\{100, 200, \cdots, 1000\}$ while the cluster size of each clusters are increased along with the whole vertices number linearly. It is obvious that, the performance of CSC is less expensive than SCAN while the number of vertex is increased. This may suggest the optimization of SCAN is worked and the costs of two correction step is small.

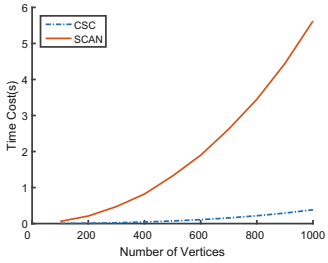
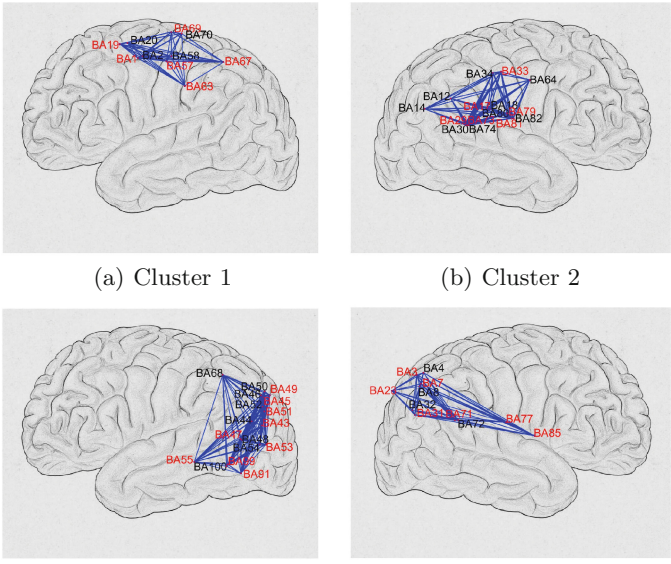


Fig. 3. The time cost for the vertices from 100 to 1000

Resting State Results. The brain subnetworks of resting state functional connectivity with CSC are shown in Fig. 4. Parameters of CSC are set as $\epsilon = 0.7$, $MinPts = 2$ and $\varepsilon = \varphi = 2$. According to previous research, the clustering result could be concentrated into these 4 clusters. Most of the region of interest in the same clusters illustrate number of potential functional relations. The ROIs in brain's left hemisphere are noted as blue label, otherwise are noted as black label. The regions such as superior frontal gyrus, media (SFGmed), the opercular part, triangular part and orbital part in inferior frontal gyrus (IFGoperc, IFGtriang, ORBinf respectively), anterior cingulate and paracingulate gyri (ACG), median cingulate and paracingulate gyri (DCG), posterior cingulate gyrus (PCG), PCUN, precentral gyrus (PreCG), rolandic operculum



(a) Cluster 1 (b) Cluster 2

(c) Cluster 3 (d) Cluster 4

Fig. 4. The brain subnetworks of resting state functional connectivity with CSC

(ROL), postcentral gyrus (PoCG), Inferior parietal (IPL), STG and inferior temporal gyrus (ITG) are contained in this cluster. Previous research suggest that SFGmed, ACG, DCG, PCG, PCUN, IPL, STG and ITG are contained in default mode network [17,18]. Furthermore, there is evidence that older adults exhibited a significant relation among SFGmed, IFGoperc, IFGtriang, ORBinf, PoCG and IPL, whereas no relation between behavior and functional connectivity was evident for younger adults which is suggesting that intrinsic functional connectivity is a potential mechanism of age-related cognitive performance [19].

4 Conclusion

This paper presented a novel clustering approach named combinational structural clustering (CSC). In our approach, the false positive outliers are reduced from clustering results while very few false negative outliers are created. Therefore, the accuracy of clustering results is increased compared with previous research.

Our CSC approach is suitable for the complex network with the mechanism of ER random graph and scale free network. And these two type of complex network has always been utilize to explain the mechanism of brain functional connectivity. The simulation results suggest that the parameter ϵ of CSC is insensitive to false positive outliers which makes the designer needn't to choose the parameter carefully.

The experimental results of resting state functional connectivity indicate the previous researches [17–19]. So we could consider the clustering results make sense to biological significance.

In general, CSC is such an effective approach with low computational cost for the brain functional connectivity in large scale networks.

Acknowledgments. This work is partly supported by the National Natural Science Foundation of China (Grant No. 61472058), and the Fundamental Research Funds for the Central Universities (Grant No. 3132016027).

References

1. Xu, X., Yuruk, N., Feng, Z., Schweiger, T.A.J.: Scan: a structural clustering algorithm for networks. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 824–833 (2007)
2. Polan, R., Nitsche, M.A., Paulus, W.: Modulating functional connectivity patterns and topological functional organization of the human brain with transcranial direct current stimulation. Hum. Brain Mapp. **32**(8), 1236–1249 (2011). S318
3. Lynall, M.E., Bassett, D.S., Kerwin, R., Mckenna, P.J., Kitzbichler, M., Mller, U., Bullmore, E.: Functional connectivity and brain networks in schizophrenia. J. Neurosci. **30**(28), 9477–9487 (2010)
4. Nascimento, M.C.V., de Carvalho, A.C.: Spectral methods for graph clustering - a survey. Euro. J. Oper. Res. **211**(2), 221–231 (2011)

5. Pulvermller, F., Kherif, F., Hauk, O., Mohr, B., Nimmo-Smith, I.: Distributed cell assemblies for general lexical and category-specific semantic processing as revealed by fMRI cluster analysis. Hum. Brain Mapp. **30**(12), 3837–3850 (2009)
6. Wang, J., Qiu, S., Xu, Y., Liu, Z., Wen, X., Hu, X., Zhang, R., Li, M., Wang, W., Huang, R.: Graph theoretical analysis reveals disrupted topological properties of whole brain functional networks in temporal lobe epilepsy. Clin. Neurophysiol. **125**(9), 1744–1756 (2014). Official Journal of the International Federation of Clinical Neurophysiology
7. Greicius, M.D., Ben, K., Reiss, A.L., Vinod, M.: Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proc. Natl. Acad. Sci. USA **100**(1), 253–258 (2003)
8. Shiokawa, H., Fujiwara, Y., Onizuka, M.: Scan++. Proc. Vldb Endow. **8**(11), 1178–1189 (2015)
9. Albert, R., Barabasi, A.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**(1), 47–98 (2001)
10. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971)
11. Horovitz, S.G., Fukunaga, M., de Zwart, J.A., van Gelderen, P., Fulton, S.C., Balkin, T.J., Duyn, J.H.: Low frequency BOLD fluctuations during resting wakefulness and light sleep: a simultaneous EEG-fMRI study. Hum. Brain Mapp. **29**(6), 671–682 (2008)
12. Van Dijk, K.R., Tvenkataraman, H.: Intrinsic functional connectivity as a tool for human connectomics: theory, properties and optimization. J. Neurophysiol. **103**(1), 297–321 (2009)
13. Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J.: Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. **2**(4), 189–210 (1994)
14. Yan, C.G., Zang, Y.F.: DPARSF: a matlab toolbox for pipeline data analysis of resting-state fMRI. Front. Syst. Neurosci. **4**, 13 (2010)
15. Granovetter, M.S.: The strength of weak ties. Soc. Sci. Electron. Publ. **78**(2), 1360–1380 (2015)
16. Rubinov, M., Sporns, O.: Complex network measures of brain connectivity: uses and interpretations. Neuroimage **52**(3), 1059–1069 (2010)
17. Whitfield-Gabrieli, S., Ford, J.M.: Default mode network activity and connectivity in psychopathology. Ann. Rev. Clin. Psychol. **8**(1), 49–76 (2011)
18. Greicius, M.D., Supekar, K., Menon, V., Dougherty, R.F.: Resting-state functional connectivity reflects structural connectivity in the default mode network. Cereb. Cortex **19**(1), 72–78 (2009)
19. Chen, N.K., Chou, Y.H., Song, A.W., Madden, D.J.: Measurement of spontaneous signal fluctuations in fMRI: adult age differences in intrinsic functional connectivity. Brain Struct. Funct. **213**(6), 571–585 (2009)

Memetic Algorithms for the Automatic Discovery of Software Architectures

Aurora Ramírez(✉), Rafael Barbudo, José Raúl Romero, and Sebastián Ventura

Department of Computer Science and Numerical Analysis, University of Córdoba, Campus of Rabanales, 14071 Córdoba, Spain
{aramirez,rbarbudo,jrromero,sventura}@uco.es

Abstract. Designing complex software systems involves making choices. Particularly in the early stages, software architects need to consider a wide range of design alternatives in order to satisfactorily meet all the requirements. Search techniques like evolutionary algorithms can support them to better explore these choices, requiring the resolution of the formulated search problem under the guidance of software metrics. Aspects like the system complexity and the diversity of metrics have a significant impact on the algorithm performance. Therefore, trying to improve its effectiveness by combining it with local search techniques represents an interesting option to fine tune the returned results. However, designing this kind of hybrid approaches, a.k.a. memetic algorithms, pose new challenges such as determining where these improvements should be located within the evolutionary process. Two memetic approaches are proposed with the aim of analysing and comparing their performance when dealing with a complex problem like the discovery of software architectures.

Keywords: Search-based software engineering · Memetic algorithm · Software architecture · Local search

1 Introduction

The design of complex software systems involves a continuous decision-making process, in which engineers should reach a solution satisfying both the functional and non-functional requirements. Within the software development cycle, the architectural analysis constitutes an early but critical phase, as its aim consists in specifying the highest level structure that will guide the subsequent system development and deployment. At this point in the project, software engineers still rely on their own judgement and past experiences, potentially causing errors that may dramatically increase the final costing for the project. Consequently,

Work supported by the Spanish Ministry of Economy and Competitiveness, project TIN2014-55252-P, and the Ministry of Education, FPU program (FPU13/01466).
© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_43

providing them with intelligent techniques to support their design process represents an appealing opportunity for the artificial intelligence field [1].

Since the appearance of search-based software engineering (SBSE) [2], metaheuristic techniques like evolutionary algorithms (EAs) and local search (LS) are being applied to the resolution of software engineering (SE) tasks like service composition [3] and software modularisation [4]. Focusing on software architecture optimisation [5], a subfield of SBSE aimed at improving software artefacts at the architectural level, the automatic identification of the underlying component-based architecture of a software system represents a complex optimisation task that has been recently addressed by means of EAs [6].

Given the richness and complexity of SE as an application domain, and particularly of those problems related to the earliest analysis, the adaptation of search methods demands great effort and knowledge in order to fully exploit their potential. That is why the use of more sophisticated alternatives like hybrid approaches still constitutes an area to be explored in SBSE, even though several studies reveal their suitability when compared to *pure* techniques [7]. For instance, memetic algorithms (MAs) propose the combination of EAs and LS techniques to better balance the exploration and exploitation of the search space. For the optimisation of software architectures, performing local improvements might help to bring those candidate solutions found by the EA closer to the outcome expected by the expert, specially when dealing with the most complex systems. However, the design of hybrid methods raises new questions regarding how they should be combined or how to correctly configure an even greater number of parameters than those required by each technique separately.

As discussed in [7], all these design issues greatly influence the performance of MAs and their optimal values are however quite often problem-dependent. It does not only mean that there are no guidelines for the design of general and competitive MAs, but also that an in-depth understanding of the problem-specific characteristics is required. Accordingly, this paper explores the design of two different types of MAs to deal with the automatic discovery of component-based software architectures, where both hill climbing (HC) and simulated annealing (SA) are considered as the specific LS technique. An experimental study is conducted to analyse the role of LS and its set-up, and consequently the capabilities of the resulting MAs when compared to the baseline EA.

The rest of the paper is organised as follows. Section 2 describes the related work, focused on the elements surrounding the design of MAs. Section 3 explains the baseline evolutionary approach to discover software architectures, as well as the two memetic approaches proposed on top of it. Section 4 presents the experimental results, while Sect. 5 outlines some conclusions and future work.

2 Background

A memetic algorithm is a special type of hybrid approach in which two well-known optimisation techniques, i.e. evolutionary algorithms and local search, work together with the aim of improving the exploitation capability of the EA [8].

Starting from an initial solution, a local search method iteratively explores its neighbourhood in order to find better solutions. HC represents one of the simplest procedures as it would only accept a movement to a better neighbour before going to the next iteration. Since HC can get stuck in local optima, other techniques were proposed to overcome such a limitation. For instance, SA allows replacing the current solution by another one with a lower quality, defining an acceptance probability that gradually decreases as the search progresses.

Frequently, MAs are created by injecting the local search procedure into the evolutionary search, acting as a new genetic operator. However, as already discussed in [7], the way in which hybridisation happens should be considered as a design decision that can even be motivated by the specific optimisation problem at hand. Furthermore, there are no general insights about the performance of memetic algorithms, specially in real-world applications. For these reasons, the following aspects need to be studied when implementing this kind of approaches:

- How to find the best trade-off between local and global search, e.g. the number of iterations performed by each method.
- How to select the most appropriate LS algorithm.
- How often LS is applied, and when.
- How to select the set of individuals that will represent the initial solutions for the local search.

Even though memetic computing is a popular technique within the artificial intelligence field, it rarely appears in the context of SBSE. The knowledge required to design and tune these algorithms clearly hampers their general adoption, so only a few works in the SBSE literature can be found. Focusing on software design, a comparison between EAs and ant colony optimisation applied to the design of object-oriented systems is presented in [9]. In this case, a greedy algorithm implements the LS procedure. The authors discussed that even when the hybrid approach does not always provide better results, it allows speeding-up the generation of good enough solutions. Similarly, the influence of SA in the specification of low-level software architectures is analysed in [10]. The experimental outcomes demonstrate that invoking the local search in different steps of the evolution have a significant impact on the results.

3 Memetic Algorithms

3.1 Baseline Evolutionary Algorithm

The discovery of the software architecture consists in identifying the constituent high-level elements of a system structure, in terms of its components and interfaces. This can be formulated as a combinatorial problem [6], where each component is derived from an group of strongly related classes, extracted from an early analysis model. Further, interfaces are defined considering how classes belonging to different components link together, and each connector is formed by a pair of provided and required interfaces.

Table 1. Design metrics used to evaluate the candidate solutions.

Metric	Formula
ICD: Intra-modular Coupling Density	$ICD_i = ((\#cl_t - \#cl_i)/\#cl_t) \cdot \left(CI_i^{in}/(CI_i^{in} + CI_i^{out})\right)$ $ICD = \sum_{i=1}^{n} ICD_i/n$
ERP: External Relations Penalty	$ERP = \sum_{i=1}^{n} \sum_{j=i+1}^{n} (w_{as} \cdot n_{as_{ij}} + w_{ag} \cdot n_{ag_{ij}} + w_{co} \cdot n_{co_{ij}} + w_{ge} \cdot n_{ge_{ij}})$
CS: Critical size	$CS = \sum_{i=1}^{n} CC_i$, $CC_i = 1\ if\ \#cl_i > threshold,\ 0\ otherwise$
CB: Component Balance	$SB(n) = \frac{n-\gamma}{\mu-\gamma}\ if\ n < \mu, = 1 - \frac{n-\mu}{\omega-\mu}\ if\ \mu < n < \omega, = 0\ if\ n \geq \omega$ $CSU(n) = 1 - Gini(\{\#cl_i \forall i \in [1,n]\}),\ CB = SB(n) \cdot CSU$

A first proposal for the evolutionary discovery of software architectures was presented in [6], where classes from an input analysis model are randomly distributed among a random number of components, properly delimited by a minimum and a maximum value. Then, candidate interfaces are identified accordingly. Each solution encodes a candidate component-based software architecture, represented as a component diagram.

The evaluation of the individuals is based on four design metrics (see Table 1), which are related to aspects like low coupling, high cohesion and the balance among the component sizes. ICD, which should be maximised, calculates the mean ratio between the number of internal (CI^{in}) and external relationships (CI^{out}) in each component i, also considering the number of classes inside that component (cl_i) with respect to the total number of classes (cl_t). Next, ERP penalises the presence of links outside the interface between a couple of components, i and j, which clearly increases coupling. Thus, this measure has to be minimised. ERP differentiates between different sorts of relationships like associations (n_{as}), aggregations (n_{ag}), compositions (n_{co}) and generalisations (n_{ge}), each one having its own weight (w). CS counts the number of components exceeding a given size ($threshold$), defined in terms of a percentage of the total number of classes of the input model. CS has to be minimised too. Finally, CB looks for the maximum balance between the number of components (n) and their size. γ, ω, and μ represent the minimum, maximum and average number of components, respectively. These four metrics are then aggregated using their ranking values to return a single fitness value for each solution, s, as follows:

$$f(s) = \begin{cases} r(ICD(s)) + r(ERP(s)) + r(CS(s)) + r(CB(s)) & \text{if } s \textit{ is feasible} \\ populationSize \cdot \#metrics + 1 & \text{if } s \textit{ is infeasible}. \end{cases} \quad (1)$$

The rest of elements of the EA are: a binary tournament to select the parents; a customised mutation operator to generate new solutions, which can perform up to five different transformations (create or remove a component, merge or divide components and move a class); and a replacement strategy preserving 10% of the best current individuals and 90% of the best offspring.

3.2 Algorithm EA(LS)

The first MA presented here, referred as EA(LS), follows the most commonly used approach, which includes the local search as a genetic operator to be executed right after the mutation. For this problem domain (see Sect. 3.1), the neighbourhood of a solution is determined by the set of solutions presenting the same distribution of classes into components as that solution, with the exception of only one class. Thus, neighbours are obtained by executing the mutator that moves a class from one component to another. One only neighbour is generated per iteration to control the number of evaluations assigned to the local search.

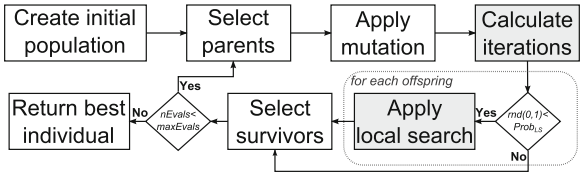


Fig. 1. Flowchart of the memetic algorithm EA(LS).

Another important element is the number of iterations performed by the LS operator. This value is dynamically determined based on the number of evaluations currently available in each generation of the evolution. The proposed method allows increasing the number of iterations as the search progresses, trying to progressively give more opportunities to the best solutions found by the evolutionary algorithm. Given a minimum (min) and a maximum (max) number of iterations, and a maximum number of evaluations ($maxEvals$), the logistic function that determines the number of iterations, which depends on the number of evaluations already used ($nEvals$), is defined as follows:

$$iterations = \frac{max - min}{1 + e^{-k \cdot (nEvals - maxEvals/2)}}. \tag{2}$$

The complete algorithm is shown in Fig. 1, shadow boxes representing those additional steps required by the MA. After the creation of the population, the algorithm iterates until a maximum number of evaluations is reached. Once the mutated individuals have been generated, the LS procedure is invoked with the number of iterations calculated according to the current point of the search. LS is executed over each offspring with a specific probability, which can be configured. Then, the current population and their descendants compete for survival.

The proposed scheme allows including any LS procedure, and only special attention has to be paid when comparing the current solution against the generated neighbour. Notice that the fitness value of a given solution is based on its ranking functions, so it is relative to the population to which it belongs. Therefore, as the fitness of the neighbour is computed by replacing the current solution, the movement is only accepted if the ranking position is equal to or better than the position of the previous solution.

3.3 Algorithm EA+LS

The second approach, named EA+LS, looks for the improvement of the solutions found after the evolution is over (see Fig. 2). Solutions will then compete against the final population, the best individual being returned. The internal behaviour of the LS technique is the same as that defined for EA(LS) in Sect. 3.2.

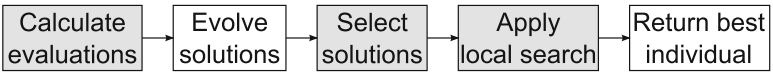


Fig. 2. Flowchart of the memetic algorithm EA+LS.

This method defines two complementary parameters to decide the amount of resources employed by the LS procedure. More specifically, only a percentage of the best final individuals will constitute the input for the local search, the number of iterations remaining the same for all of them, since their quality is assumed to be similar. Both parameters are known before running the algorithm, so the resulting number of iterations is subtracted from the total number of evaluations.

Table 2. Parameter set-up

Parameter	Value
Population size	150
Max. Evaluations	24000
t_0 in SA-m, SA-l	650, 100000
α in SA	0.95
Probability in EA(LS)	0.01, 0.05, 0.10 0.20, 0.80, 0.90 0.95, 0.99, 1.00
Iterations in EA(LS)	$min = 10$ $max = 50$ k =0.0005
% Sol. in EA+LS	10, 15, 20
Iterations in EA+LS	50, 100, 200

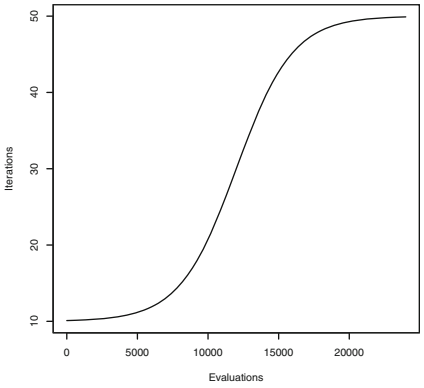


Fig. 3. Logistic function computing the number of iterations in EA(LS).

4 Experimental Framework

All the algorithms have been coded using JCLEC[1] and Datapro4j[2] Java libraries. Experiments were run on an Ubuntu 14.10 computer with 8 cores Intel Core i7

[1] http://jclec.sourceforge.net/.
[2] http://www.uco.es/grupos/kdis/datapro4j.

2.67 GHz and 15.6 GB RAM. Table 2 details the parameter set-up. Firstly, common parameters, e.g. population size, stopping criterion and those related to the evaluation metrics (see Table 1), have been determined according to the previous work [6]. More specifically, $w_{as} = 2$, $w_{ag} = 3$, $w_{co} = 3$, $w_{ge} = 5$, $threshold = 30\%$, $\gamma = 2$, $\mu = 5$ and $\omega = 8$. Moreover, four real-world software systems are used as problem instances: *Dataprо4j* (59 classes and 61 relationships), *JSapar* (46 and 89), *Marvin* (32 and 51) and *NekoHTML* (47 and 73).

Three local search methods have been considered: hill climbing and simulated annealing, the latter with two different acceptance functions: Metropolis (SA-m) and logistic (SA-l). Both methods apply a geometric cooling factor (α), whereas the initial temperature (t_0) was determined after preliminary experiments. Focusing on the algorithm EA(LS), the probability of applying LS over offspring might represent a key factor for its success, so a wide range of values has been considered. Similarly, the algorithm EA+LS has been tested with different configurations, adapting the percentage of solutions and the number of iterations. It has provided 9 variants, each one with a distinct distribution of evaluations for the EA and the LS. Figure 3 shows the function determining the number of iterations in EA(LS) according to Eq. 2 and the values detailed in Table 2. The *pure* evolutionary algorithm (referred as EA) acts as the reference algorithm.

4.1 Results Obtained by EA(LS)

Table 3 compiles the results obtained by EA(LS). The first column represents the percentage of satisfactory movements during the LS procedure, whereas the second column shows the number of evaluations required to find the best individual, i.e. the solution to be returned. The mean and standard deviation considering all the problem instances is shown. Then, the average percentage of improvement for each software metric used to compute the fitness function has been obtained. These values have been calculated for each problem instance, given that the algorithms are more variable here. Due to space limitations, the third column in Table 3 depicts only the best improvement achieved by the corresponding algorithm, including the metric and problem instance referred. Notice that the percentage of satisfactory movements varies between 16.70% and 36.70% depending on the LS operator and its probability. On the one hand, the performance of SA with the metropolis function (SA-m) is higher to that obtained by using the logistic function (SA-l). On the other hand, HC only performs relatively well when a low probability is applied.

In general, increasing the probability of applying local search cannot be always translated into a greater percentage of satisfactory movements, probably due to the fact that the number of evaluations assigned to the EA is not enough to properly locate the most promising areas of the search space. Nevertheless, this memetic model can provide good enough solutions by using a smaller number of evaluations than the original algorithm (EA). The speeding-up is particularly relevant when any type of SA is applied with a high probability. However, the behaviour of the resulting algorithm tends to vary among different executions according to the obtained standard deviation.

Table 3. Results for EA(LS)

Algorithm	% Satisfactory movements	Number of evaluations	Metric improvement		
			%	*Metric*	*Instance*
EA	-	23615.90 ± 732.20	-	-	-
EA(HC)_0.001	35.60 ± 3.20	23589.80 ± 663.80	1.17	CS	NekoHTML
EA(HC)_0.05	21.60 ± 2.90	23246.60 ± 1137.20	2.30	CB	Marvin
EA(HC)_0.10	21.30 ± 2.50	23067.20 ± 1473.70	1.75	CB	Datapro4j
EA(HC)_0.20	20.80 ± 2.90	22462.40 ± 2493.20	2.52	CB	Datapro4j
EA(HC)_0.50	20.10 ± 2.80	22421.80 ± 2431.00	2.59	CB	Datapro4j
EA(HC)_0.80	19.50 ± 3.00	22284.00 ± 2813.70	2.33	CS	JSapar
EA(HC)_0.90	19.47 ± 3.00	22535.70 ± 2718.10	2.67	CS	NekoHTML
EA(HC)_0.95	19.50 ± 3.10	21960.60 ± 3038.00	2.45	CB	Datapro4j
EA(HC)_1.00	19.40 ± 3.00	22460.60 ± 2801.20	2.83	CS	NekoHTML
EA(SA-m)_0.001	35.20 ± 3.00	23616.00 ± 677.30	1.53	CB	Datapro4j
EA(SA-m)_0.05	21.90 ± 2.70	23272.10 ± 1112.40	3.31	CB	Datapro4j
EA(SA-m)_0.10	21.80 ± 2.80	23111.30 ± 1340.10	1.30	CB	Datapro4j
EA(SA-m)_0.20	22.00 ± 2.50	22699.50 ± 1897.40	3.24	CB	Datapro4j
EA(SA-m)_0.50	26.50 ± 2.50	22617.00 ± 2070.60	2.68	ERP	NekoHTML
EA(SA-m)_0.80	32.90 ± 1.90	22545.50 ± 3958.00	2.25	CB	JSapar
EA(SA-m)_0.90	34.80 ± 2.00	19079.10 ± 7452.80	5.51	CB	JSapar
EA(SA-m)_0.95	35.76 ± 2.00	19490.70 ± 6899.00	1.22	CS	JSapar
EA(SA-m)_1.00	36.70 ± 2.00	19876.60 ± 6845.10	6.02	CB	JSapar
EA(SA-l)_0.001	31.90 ± 3.00	23707.30 ± 587.50	2.34	CB	JSapar
EA(SA-l)_0.05	17.70 ± 2.00	23187.30 ± 1394.80	2.82	CB	Datapro4j
EA(SA-l)_0.10	16.90 ± 2.00	22877.00 ± 1557.80	2.21	CB	Datapro4j
EA(SA-l)_0.20	16.70 ± 1.90	22803.50 ± 1663.90	1.33	CS	Marvin
EA(SA-l)_0.50	18.70 ± 1.50	19013.50 ± 6477.00	2.20	CB	JSapar
EA(SA-l)_0.80	19.30 ± 1.70	14709.00 ± 7251.50	2.08	CB	JSapar
EA(SA-l)_0.90	19.67 ± 1.70	14986.50 ± 7308.20	1.17	CB	Datapro4j
EA(SA-l)_0.95	19.60 ± 1.60	13019.30 ± 6787.90	3.00	CS	NekoHTML
EA(SA-l)_1.00	19.70 ± 1.60	15455.00 ± 7173.00	1.67	CS	NekoHTML

Focusing on the design metrics, it should be noted that CB and CS are frequently improved, achieving up to 6.02% improvement for one of the most complex problem instances, *JSapar*. These outcomes reveal that the MA is able to find room for improvement in contrast with the size of the architectural solutions found by the original EA. In this sense, the application of more sophisticated techniques has proven to be useful to software architects, as the algorithm provides them with more balanced components that help improving the understandability of the architectural model. On the contrary, it should be noted the original algorithm maintains a better trade-off among all the design metrics.

4.2 Results Obtained by EA+LS

Table 4 shows how the behaviour of the approach EA+LS changes drastically. On the one hand, EA+LS seems to perform a greater number of satisfactory movements. On the other hand, the number of evaluations required to find the best solutions is the same for the three LS techniques under the same percentage of solutions and number of iterations. It reveals that the best individual within the population was always obtained by the EA before invoking the LS procedure, independently of the problem instance. The high effectiveness achieved in the number of movements is therefore explained by the fact that a neighbour solution with the same fitness value is accepted as a better solution, since it allows the exploration of a new search direction. In fact, the final population often improves

Table 4. Results for EA+LS

Algorithm	% Satisfactory movements	Number of evaluations	Metric improvement		
			%	*Metric*	*Instance*
EA	-	23615.90 ± 732.20	-	-	-
EA+HC_10%_50	100.00 ± 0.00	23347.50 ± 55.90	1.91	CB	Marvin
EA+HC_10%_100	100.00 ± 0.00	22598.40 ± 55.90	1.62	CB	Marvin
EA+HC_10%_200	100.00 ± 0.00	21096.80 ± 56.90	1.62	CB	Marvin
EA+HC_15%_50	100.00 ± 0.00	22945.50 ± 29.60	1.49	CB	Marvin
EA+HC_15%_100	100.00 ± 0.00	21883.30 ± 23.30	1.78	CB	Marvin
EA+HC_15%_200	100.00 ± 0.00	19647.30 ± 30.20	2.00	CB	Datapro4j
EA+HC_20%_50	100.00 ± 0.00	22598.40 ± 55.90	1.62	CB	Marvin
EA+HC_20%_100	100.00 ± 0.00	21096.80 ± 56.90	1.62	CB	Marvin
EA+HC_20%_200	100.00 ± 0.00	18099.00 ± 57.20	1.53	CB	Marvin
EA+SA-m_10%_50	100.00 ± 0.00	23347.50 ± 55.90	1.91	CB	Marvin
EA+SA-m_10%_100	100.00 ± 0.00	22598.40 ± 55.90	1.62	CB	Marvin
EA+SA-m_10%_200	100.00 ± 0.00	21096.80 ± 56.90	1.62	CB	Marvin
EA+SA-m_15%_50	100.00 ± 0.00	22945.50 ± 29.60	1.49	CB	Marvin
EA+SA-m_15%_100	100.00 ± 0.00	21883.30 ± 23.30	1.78	CB	Marvin
EA+SA-m_15%_200	100.00 ± 0.00	19647.30 ± 30.20	2.00	CB	Datapro4j
EA+SA-m_20%_50	100.00 ± 0.00	22598.00 ± 55.90	1.62	CB	Marvin
EA+SA-m_20%_100	100.00 ± 0.00	21096.80 ± 56.90	1.62	CB	Marvin
EA+SA-m_20%_200	100.00 ± 0.00	18099.00 ± 57.20	1.53	CB	Marvin
EA+SA-l_10%_50	50.00 ± 1.80	23347.50 ± 55.90	1.91	CB	Marvin
EA+SA-l_10%_100	49.80 ± 1.20	22598, 40 ± 55.90	1.62	CB	Marvin
EA+SA-l_10%_200	49.90 ± 0.90	21096.80 ± 56.90	1.62	CB	Marvin
EA+SA-l_15%_50	50.03 ± 1.40	22945.50 ± 29.60	1.70	CB	Marvin
EA+SA-l_15%_100	49.82 ± 1.10	21883.30 ± 23.30	1.78	CB	Marvin
EA+SA-l_15%_200	49.93 ± 0.90	19647.30 ± 30.20	2.00	CB	Datapro4j
EA+SA-l_20%_50	49.90 ± 1.30	22598.40 ± 55.90	1.62	CB	Marvin
EA+SA-l_20%_100	49.90 ± 0.90	21096.80 ± 56.90	1.62	CB	Marvin
EA+SA-l_20%_200	50.00 ± 0.70	18099.00 ± 57.20	1.53	CB	Marvin

in global terms with the respect to that obtained by the EA, even when it does not modify the quality of the best individual finally returned.

Regarding the metric values, results also present similarities among the three LS techniques. This MA can perform small enhancements with respect to CB, *Marvin* frequently being the problem instance that obtains the highest percentage of improvement. Therefore, it can be concluded that the application of LS after the evolution (EA+LS) does not cause a significant difference to the software architect, since the EA has already converged to good solutions.

5 Concluding Remarks

Memetic approaches represent powerful techniques to address complex optimisation problems, as they apply complementary search methods to overcome the limitations of *pure* resolution methods. However, its application to a new domain often requires an in-depth study of the most appropriate search technique, hybridisation point and parameter set-up. A comparative study has been proposed to analyse how local search could improve the performance of an EA aimed at solving the automatic identification of component-based software architectures.

Experimental outcomes indicate that there exist several differences between the two methods proposed. Although less effective in the number of neighbours found, injecting the local search as a genetic operator allows the algorithm to reduce the number of evaluations required to find good enough solutions, while obtaining a greater improvement in terms of two software metrics, CB and CS. On the contrary, executing the local search after the evolution does not present a real benefit with respect to the baseline algorithm, since the evaluations devoted to the final improvement do not result in a best solution than that obtained by the EA. Future research will be directed towards refining the local search procedure, e.g. using domain knowledge to guide the exploration of the neighbourhood, as well as to address the design of MAs from a multi-objective perspective.

References

1. Harman, M.: The role of artificial intelligence in software engineering. In: 1st International Workshop on Realizing AI Synergies in Software Engineering, pp. 1–6. IEEE Press, Piscataway (2012)
2. Harman, M., Afshin Mansouri, S., Zhang, Y.: Search based software engineering: trends, techniques and applications. ACM Comput. Surv. **45**, 1–64 (2012)
3. Mucientes, M., Lama, M., Couto, M.I.: A genetic programming-based algorithm for composing web services. In: 9th International Conference on Intelligent Systems Design and Applications, pp. 379–384. IEEE (2009)
4. Mahdavi, K., Harman, M., Hierons, R.M.: A multiple hill climbing approach to software module clustering. In: International Conference on Software Maintenance, pp. 315–324. IEEE Computer Society, Washington (2003)
5. Aleti, A., Buhnova, B., Grunske, L., Koziolek, A., Meedeniya, I.: Software architecture optimization methods: a systematic literature review. IEEE Trans. Soft. Eng. **39**, 658–683 (2013)

6. Ramírez, A., Romero, J.R., Ventura, S.: An approach for the evolutionary discovery of software architectures. Inf. Sci. **305**, 234–255 (2015)
7. Krasnogor, N., Smith, J.: A tutorial for competent memetic algorithms: model, taxonomy, and design issues. IEEE Trans. Evol. Comp. **9**, 474–488 (2005)
8. Neri, F., Cotta, C.: Memetic algorithms and memetic computing optimization: A literature review. Swarm Evol. Comput. **2**, 1–14 (2012)
9. Smith, J., Simons, C.L.: A comparison of two memetic algorithms for software class modelling. In: 15th Annual Conference on Genetic and Evolutionary Computation, pp. 1485–1492. ACM, New York (2013)
10. Sievi-Korte, O., Erkki, M., Poranen, T.: Simulated annealing for aiding genetic algorithm in software architecture synthesis. Acta Cybernetica **21**, 235–265 (2013)

Multi-objective Dynamic Analysis Using Fractional Entropy

E.J. Solteiro Pires[1(✉)], J.A. Tenreiro Machado[2], and P.B. de Moura Oliveira[1]

[1] INESC TEC – INESC Technology and Science (UTAD Pole),
Escola de Ciências e Tecnologia, Universidade de Trás-os-Montes e Alto Douro,
5000-811 Vila Real, Portugal
{epires,oliveira}@utad.pt

[2] Department of Electrical Engineering, ISEP–Institute of Engineering,
Polytechnic of Porto, Rua Dr. António Bernardino de Almeida,
4249-015 Porto, Portugal
jtm@isep.ipp.pt

Abstract. Multi-objective optimization evolutionary techniques provide solutions for a specific problem using optimally concepts taking into consideration all the design criteria. In the last years, several multi-objective algorithms were proposed but usually the performance is measured at the end neglecting, therefore, the solution diversity along the interactions. In order to understand the evolution of the solutions this work studies the dynamic of the successive iterations. The analysis adopts the fractional entropy for measuring the statistical behavior of the population. The results show that the entropy is a good tool to monitor and capture phenomena such as the diversity and convergence during the algorithm execution.

Keywords: Multi-objetive genetic algorithm · Fractional entropy · Diversity · Convergence · Dynamic

1 Introduction

Multi-objective optimization, inspired in nature and biologically phenomena, is based on a population of solutions that evolves during several iterations. In genetic algorithms (GA), the evolution is supported by operators, namely selection, crossover and mutation. The performance of multi-objective algorithms is usually analyzed at the end of their execution and its success is measured using several metrics proposed in the literature. Additionally, in ambiguous situations, the use of nonparametic tests can be adopted, to clarify the algorithm supremacy [6,17]. Some of the proposed indices are based on Shannon entropy. One method presented by Wang *et al.* [20] revealed interesting results, because (i) the computational effort increases linearly with the number of solutions, (ii) the metric

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_44

qualifies the combination of uniformity and coverage of Pareto set and (iii) it determines when the evolution has reached its maturity. Another method, proposed by LinLin and Yunfang [9] to measure the performance of multi-objective problems, not only indicates when the algorithm should be stopped, but can also compare the performance of multi-objective algorithms. The entropy is evaluated regarding the solution density in a grid space.

Other studies tried to understand the population dynamics during time evolution [12,14,18,19]. Myers and Hancock [12] suggest the use of Shannon entropy to evaluate the run-time performance of a GA to solve labeling problems. The entropy measured in the parameter space is used to provide useful information about the algorithm state. Myers and Hancock concluded that populations with entropy smaller than a given threshold become saturated and the respective diversity disappears. Pires *et al.* [18] studied the signal propagation during the evolution of a GA. The mutation operator signal suffers a perturbation during some generations and the corresponding fitness variation is analyzed. Pires *et al.* also used the Shannon entropy to study the dynamic of multi-objective particle swarm algorithm [14] and non-sorting genetic algorithm II [19] during their execution.

So far, most of reported studies were carried out using Shannon entropy. However, there are different entropy concepts which can be used. Recently, Machado [10] proposed a fractional entropy, sensitive to small signals and useful in describing the dynamics of complex systems. Taking these ideas in consideration, this paper studies the dynamic and self-organization of solutions during the execution of multi-objective genetic algorithm (MOGA). The study considers three optimization functions and different population sizes, using the fractional entropy for evaluating the MOGA performance.

2 Multi-objective Genetic Algorithm

GA is a population based algorithm used in search and optimization problems. The most usual application is for a single objective however, the algorithm can be easily modified to deal with more that one objective. Indeed, the GA can take advantage of its population to store one non-dominated solution by chromosome in order to form a representative set of the Pareto front [3]. During the MOGA execution the solutions are guided in the search space, by means of objective functions, using techniques to promote dispersion, towards the Pareto front. The algorithm tries to achieve a compromise between all the criteria, promoting the arisen of different non-dominated solutions. Based on these concepts the MOGA has been successfully used in many engineering problems [2].

3 Entropy

Entropy is a quantitative measure that is fundamentally well defined via a mathematical equation. Although, entropy was proposed in thermodynamic as a state

variable, currently it is applied in a plethora of fields. Many entropy interpretations have been suggested over the years, such as *disorder*, *mixing*, *chaos*, *spreading*, *freedom* and *information* [1].

The first description of entropy was proposed by Boltzmann to describe systems that evolve from ordered to disordered states. *Spreading* was used by Guggenheim to indicate the diffusion of a energy system from a smaller to a larger volume. Lewis stated that, in a spontaneous expansion gas in an isolated system, *information* regarding particles locations decreases while, the missing information or, *uncertainty* increases.

Shannon [16] developed the information theory to quantify the information loss in the transmission of a given message. The study was carried out in a communication channel and Shannon focused in physical and statistical constraints that limit the message transmission.

Machado [10] proposed the fractional order, $\alpha \in \mathbb{R}$, entropy:

$$S_\alpha = \sum_{x \in X} \left(-\frac{p_i^{-\alpha}(x)}{\Gamma(\alpha+1)} [\log p_i(x) + \Psi(1) - \Psi(1-\alpha)] \right) p_i(x) \tag{1}$$

where $\Gamma(.)$ and $\Psi(.)$ represent the gamma and diagamma functions, respectively. The entropy S_α tries to incorporate the behavior verified in fractional systems models.

4 Entropy Index

This section describes the measurement of the diversity during the process evolution. The index uses the entropy proposed by Machado [10] inspired in the fractional calculus theory [7,11,13,15].

The index expresses the spreading of the population elements, from a randomly dispersed elements to a well dispersed population elements. Moreover, another propriety to be analyzed is the convergence rate of the algorithm during the time evolution.

For this purpose, the position of all solutions along the iterations are considered. Moreover, the minimum distances between the N_p solutions, *i.e.* the minimum spanning tree, are taken into account to evaluate a kind of isolation solution probability. Let d_i be one of the minimum spanning tree distances, where $i \in \{1, 2, \ldots, N_p - 1\}$, and p_i, its isolation probability for one iteration given by the following expression:

$$p_i = \frac{d_i}{\sum_{j=1}^{N_p - 1} d_j} \tag{2}$$

After evaluating the probabilities, their values are sorted and stored in vector $\mathbf{P}_t^e$ relative to the iteration t of run e.

Since the algorithm is stochastic, each time it is executed a different vector $\mathbf{P}_t^e$ (representing the best front, a different convergence evolution and a different

vector $\mathbf{P}_t^e$) is obtained. In order to have a numeric result, 21 runs are executed, $e = \{1, 2, \ldots, 21\}$, and the median obtained as representative of the $\mathbf{P}_t$ vector in opposition to the average usually used in the Monte Carlo algorithm.

Thereafter, the entropy index, at iteration t, is given by:

$$S_{\alpha,t} = \sum_{i=1}^{N_p-1} \left(-\frac{p_i^{-\alpha}(x)}{\Gamma(\alpha+1)} [\log p_i(x) + \Psi(1) - \Psi(1-\alpha)] \right) p_i(x), \tag{3}$$

where p_i is the element i of vector $\mathbf{P}_t$ and $\alpha = 0.5$.

5 Experimental Results

To investigate the MOGA dynamics, three benchmark functions, each one with three objectives, are optimized using the NSGA-II algorithm [4]. During the optimization the entropy index, for every iteration, is evaluated. The functions are DTLZ2, DTLZ4 [5] and UF8 [21]:

$$F_1 = \begin{cases} f_1(X) = [1 + g(X)] \cos(x_1\pi/2) \cos(x_2\pi/2) \\ f_2(X) = [1 + g(X)] \cos(x_1\pi/2) \sin(x_2\pi/2) \\ f_3(X) = [1 + g(X)] \sin(x_1\pi/2) \\ g(X) \ = 1 + 9 \sum_{i=3}^{m} (x_0.5)^2 \end{cases} \tag{4}$$

$$F_2 = \begin{cases} f_1(X) = [1 + g(X)] \cos(x_1^\beta\pi/2) \cos(x_2^\beta\pi/2) \\ f_2(X) = [1 + g(X)] \cos(x_1^\beta\pi/2) \sin(x_2^\beta\pi/2) \\ f_3(X) = [1 + g(X)] \sin(x_1^\beta\pi/2) \\ g(X) \ = 1 + 9 \sum_{i=3}^{m} (x_i^\beta - 0.5)^2 \end{cases} \tag{5}$$

$$F_3 = \begin{cases} f_1(X) = \cos(0.5x_1\pi) \cos(0.5x_2\pi) + \frac{2}{|J_1|} \sum_{j \in J_1} g_j(X) \\ f_2(X) = \cos(0.5x_1\pi) \sin(0.5x_2\pi) + \frac{2}{|J_2|} \sum_{j \in J_2} g_j(X) \\ f_3(X) = \sin(0.5x_1\pi) + \frac{2}{|J_3|} \sum_{j \in J_3} g_(X) \\ g_j(X) = \left(x_j - 2x_2 \sin(2\pi x_1 + \frac{\mathrm{j}\pi}{m}) \right)^2 \\ J_1 \quad = \{j | 3 \le j \le m, \text{ and } j-1 \text{ is a multiplication of 3}\} \\ J_2 \quad = \{j | 3 \le j \le m, \text{ and } j-2 \text{ is a multiplication of 3}\} \\ J_3 \quad = \{j | 3 \le j \le m, \text{ and } j \text{ is a multiplication of 3}\} \end{cases} \tag{6}$$

The parameters intervals are $x_i \in [0, 1]$ for $\{F_1, F_2\}$, and $x_i \in [-2, 2]$ for $\{F_3\}$. The number of parameters is $m = 12$ for all functions and $\beta = 12$ for function F_2.

To solve the problem, real codification, simulated binary crossover (SBX) with $n_c = 2$, uniform mutation with $p_\mathrm{m} = 0.05$ and $\Delta = 0.1(x_\mathrm{max} - x_\mathrm{min})$ are considered. The experiments adopt $t = 1000$ iterations. The population and

archive have the same size, predefined at the beginning of each execution, and maintained constant during each experiment

The MOGA algorithm considers the ϵ-dominance technique, *i.e.* for two solutions a and b, a dominates b if $(1-\epsilon)f_i(a) \succeq f_i(b)$, $i = \{1,2,3\}$ [8]. In the experiments the value of $\epsilon = 0.01$ is adopted.

This section presents the entropy evolution for six cases test set, *i.e.* for population sizes of $N_p = \{150, 200, \ldots, 400\}$ chromosomes, using the functions F_1, F_2 and F_3. In order to obtain a numerical result, 21 runs were executed, per experiment, and the medians of the distances were taken to obtain the probabilities. Thereafter, the entropy was evaluated for each experiment and the dynamic behavior is measured.

5.1 F_1 Optimization

Figure 1 illustrates six experiments for different number of chromosomes $N_p = \{150, 200, 250, 300, 350, 400\}$. It is observed that entropy indices increase with the population size. This phenomena results from the increasing of the number of terms in the sum of Eq. 3.

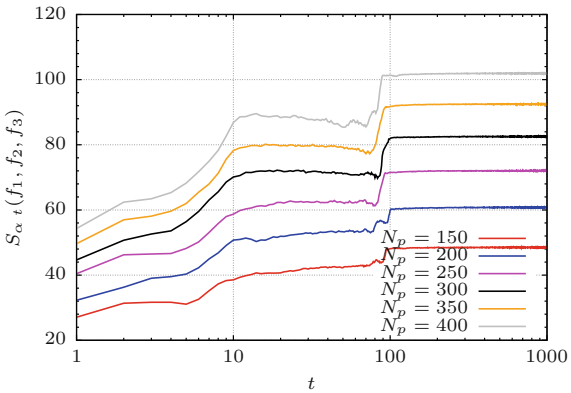
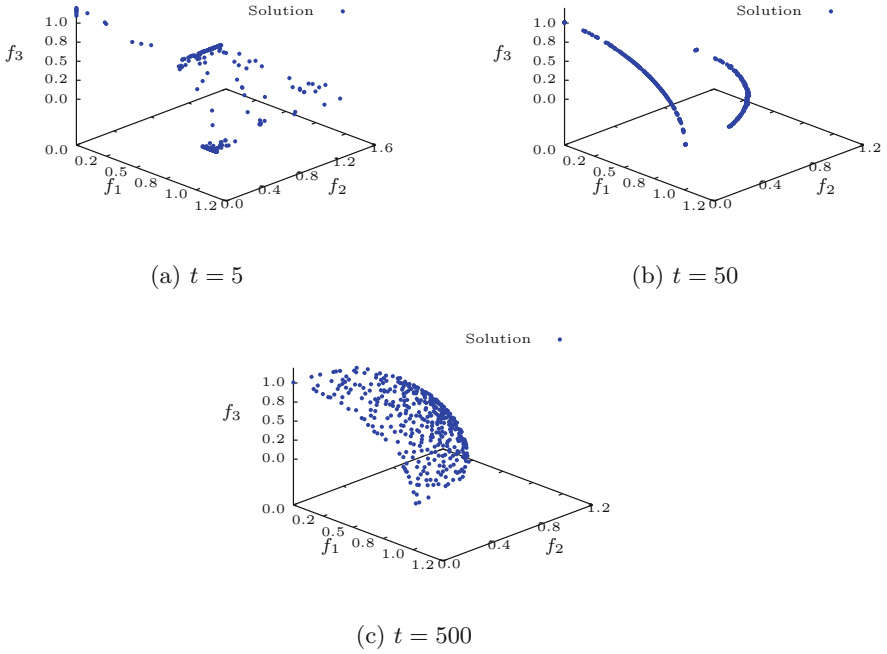


Fig. 1. Entropy $S\alpha, t(f_1, f_2, f_3)$ during the NSGA-II evolution for F_1

In Fig. 1, three evolution stages can be observed. The first stage concerns to an initial transitory phase where solutions are spread all over the objective space ($t \in [0, 10]$). The algorithm begins with a low entropy, and as solutions are converging, the entropy index increases progressively (see Fig. 2a). The second stage corresponds to the interval iterations $t \in [10, 100]$. In this phase the solutions are spread over two curve lines of the front (see Fig. 2b) and the entropy increases slowly. At the end of this stage, the entropy increases sharply in parallel with the discovery of the Pareto front. Finally, in the last stage (see Fig. 2c), entropy index becomes stagnated or in a steady-state value, and therefore, indicates that the algorithm has converged and the solutions are well spread over the Pareto



(a) $t = 5$

(b) $t = 50$

(c) $t = 500$

Fig. 2. Archive solutions at iteration t for F_1 and $N_p = 400$

front, therefore, can be concluded that the maximum entropy corresponds to the maximum spread.

5.2 F_2 Optimization

When function F_2 is optimized, the pattern convergence of the entropy is identical to the one observed with function F_1 as illustrated in Fig. 3.

5.3 F_3 Optimization

Figure 4 presents the entropy for the F_3 function. The entropy for population sizes of $N_P = \{150, 200\}$ does not present any abrupt increase during the execution. Indeed, at the end, the entropy signals reveal some decreasing behavior corresponding to loosing population diversity. Additionally, these two experiments do not converge to the Pareto font, demonstrating that both results are in agreement. On the other hand, for all the other cases, the experiments show a slight increase about generations 300–600 and converged to the Pareto front. Figure 5 shows the entropy between the archive solutions distance probability in the objective space.

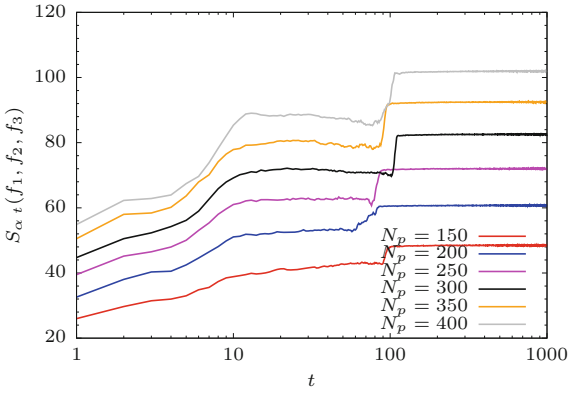


Fig. 3. Entropy $S_{\alpha,t}(f_1, f_2, f_3)$ during the NSGA-II evolution for F_2

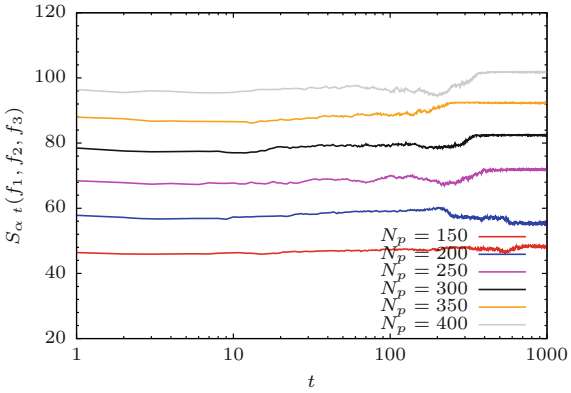
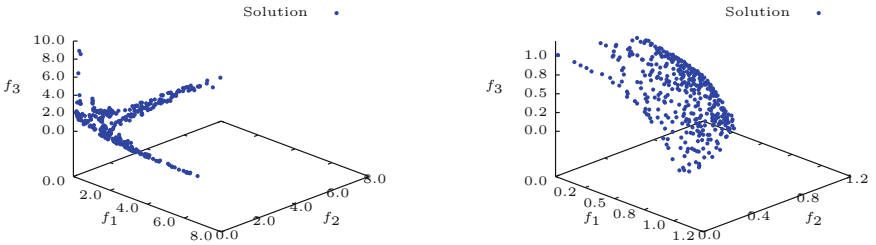


Fig. 4. Entropy $S_{\alpha,t}(f_1, f_2, f_3)$ during the NSGA-II evolution for F_3



(a) $t = 95$ (b) $t = 950$

Fig. 5. Archive solutions at iteration t for F_3 and $N_p = 400$

6 Conclusion

The entropy for describing the MOGA dynamic behavior was proposed. This work adopted the diversity measure notion expressed by one interpretation of entropy to measure the population diversity over the experiment of a MOGA. The fractional entropy was adopted in order to capture variations during the evolution. Additionally, the index stagnation can be used to indicate the algorithm convergence.

The optimization of three functions were tested. It was observed that, the optimization of the first two functions, the population size has a lower impact on the algorithm convergence. However, in the third function the population size has an important role during the algorithm iterations.

The entropy index can be used to compare the solution diversity during algorithm run with different parameters. For most MOGA cases reported in the literature, the performance evaluation is performed at the end of the algorithm, by comparing the final front extension, spreading and diversity. The algorithm proposed in this paper was used in MOGA to analyze the convergence rate during the time evolution.

The fractional entropy is sensitive to small signal variation leading to better results than the Shannon entropy [19].

References

1. Ben-Naim, A.: Entropy and the Second Law: Interpretation and Misss-Interpretations. World Scientific Publishing Company (2012)
2. Coello Coello, C.: Evolutionary multi-objective optimization: a historical view of the field. IEEE Comput. Intell. Mag. **1**(1), 28–36 (2006)
3. Deb, K.: Multi-objective Optimization Using Evolutionary Algorithms. Wiley, New York (2001)
4. Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In: Schoenauer, M., Deb, K., Rudolph, G., Yao, X., Lutton, E., Merelo, J.J., Schwefel, H.-P. (eds.) PPSN 2000. LNCS, vol. 1917, pp. 849–858. Springer, Heidelberg (2000). doi:10.1007/3-540-45356-3_83
5. Deb, K., Thiele, L., Laumanns, M., Zitzler, E.: Scalable multi-objective optimization test problems. In: Fogel, D.B., El-Sharkawi, M.A., Yao, X., Greenwood, G., Iba, H., Marrow, P., Shackleton, M. (eds.) Proceedings of the 2002 Congress on Evolutionary Computation, CEC2002, pp. 825–830. IEEE Press (2002)
6. Derrac, J., Garca, S., Molina, D., Herrera, F.: A practical tutorial on theuse of nonparametric statistical tests as a methodology for comparing evolu-tionary and swarm intelligence algorithms. Swarm Evol. Comput. **1**(1), 3–18 (2011). http://www.sciencedirect.com/science/article/pii/S2210650211000034
7. Gement, A.: On fractional differentials. Proc. Philos. Mag. **25**, 540–549 (1938)
8. Laumanns, M., Thiele, L., Deb, K., Zitzler, E.: Combining convergence and diversity in evolutionary multiobjective optimization. Evol. Comput. **10**(3), 263–282 (2002). http://dx.doi.org/10.1162/106365602760234108

9. LinLin, W., Yunfang, C.: Diversity based on entropy: a novel evaluation criterion in multi-objective optimization algorithm. Int. J. Intell. Syst. Appl. **4**(10), 113–124 (2012)
10. Machado, J.T.: Fractional order generalized information. Entropy **16**(4), 2350–2361 (2014). http://www.mdpi.com/1099-4300/16/4/2350
11. Méhauté, A.L.: Fractal Geometries: Theory and Applications. Penton Press, London (1991)
12. Myers, R., Hancock, E.R.: Genetic algorithms for ambiguous labelling problems. Pattern Recogn. **33**(4), 685–704 (2000). http://www.sciencedirect.com/science/article/pii/S0031320399000801
13. Oustaloup, A.: La Commande CRONE: Commande Robuste d'Ordre Non Intier. Hermes, Paris (1991)
14. Pires, E.J.S., Machado, J.A.T., de Moura Oliveira, P.B.: Entropy diversity in multi-objective particle swarm optimization. Entropy **15**(12), 5475–5491 (2013). http://www.mdpi.com/1099-4300/15/12/5475
15. Podlubny, I.: Fractional Differential Equations. Academic Press, San Diego (1999)
16. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal 27, 379–423, 623–656 (July, October 1948), http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf
17. Solteiro Pires, E.J., de Moura Oliveira, P.B., Tenreiro Machado, J.A.: Multi-objective MaxiMin sorting scheme. In: Coello Coello, C.A., Hernández Aguirre, A., Zitzler, E. (eds.) EMO 2005. LNCS, vol. 3410, pp. 165–175. Springer, Heidelberg (2005)
18. Pires, E.J.S., Machado, J.A.T., de Moura Oliveira, P.B.: Dynamical modelling of a genetic algorithm. Signal Process **86**(10), 2760–2770 (2006)
19. Pires, E.J.S., Machado, J.A.T., de Moura Oliveira, P.B.: Diversity study of multi-objective genetic algorithm based on shannon entropy. In: 6th World Congress on Nature and Biologically Inspired Computing (NaBIC), Porto, Portugal (30 July - 1 August) (2014)
20. Wang, L., Chen, Y., Tang, Y., Sun, F.: The entropy metric in diversity of multi-objective evolutionary algorithms. In: 2011 International Conference on Soft Computing and Pattern Recognition (SoCPaR), pp. 217–221 (2011)
21. Zhang, Q., Zhou, A., Zhao, S., Suganthan, P.N., Liu, W., Tiwari, S.: Multiobjective optimization test instances for the cec 2009 special session and competition. Technical report CES-487, University of Essex and Nanyang Technological University (2008)

Multi-agent Based Truck Scheduling Using Ant Colony Intelligence in a Cross-Docking Platform

Houda Zouhaier(✉) and Lamjed Ben Said

Laboratory SOIE, Higher Institute of Management of Tunis,
University of Tunis, Le Bardo, Tunis, Tunisia
Houda.Zouhaier@gmail.com, lamjed.bensaid@isg.rnu.tn

Abstract. The management of trucks in a cross-docking platform is a process under five steps: the arrival, the control, the unloading, the transfer and finally the loading. In each of these steps, a sequence of decisions arise. To achieve an optimal and robust solutions, the inter-dependencies between the different planning functions should be taken into account, and scheduling decisions must be made simultaneously. The truck scheduling should incorporate a real-time information regarding the resource availability and truck arrival and departure times which are crucial in a cross-docking platform. In this work, we present how the autonomous, distributed, and dynamic nature of the multi-agent paradigm by introducing ant colony intelligence (ACI) can provide a framework for the cooperation of various functions of the cross-dock to develop a robust schedule. The goal of this paper is to find an optimal dynamic scheduling system related to the parking lot and dock operations at the cross-dock facility. The proposed approach represents ACI integrated with both truck agents and resource agents to solve the truck scheduling problem in a dynamic environment.

Keywords: Resolution · Distributed approach · Real-time scheduling · Ant colony intelligence · Disruption · Agent based modeling

1 Introduction

The evolution of trade globalization, with the variability of demand and logistic costs, has complicated the task of managing the supply chain and distribution. The cross-docking is a logistic technique which seeks to reduce the logistic and transportation costs [2]. A cross-docking platform (also called a cross-dock) consolidates shipments of the same destination with different sizes for full loads. To implement the success factors of the cross-dock, the coordination of operations (the physical flow of goods and informations) must be effectively managed and the handling tools must be used properly [2]. Schaffer [16] considers that the communication among all members of a supply chain has become a key factor for the successful of cross-docking. Therefore, the data inaccuracy is a crucial issue in a cross-docking operation [5].

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_45

The information about the truck arrival times are subject to uncertainties due to traffic congestion, weather conditions, or engine failures [6]. In fact, if trucks have a deadline, it may be unrealizable to complete their unloading or their loading on time. This may lead to an earliness or a tardiness in the departure time: this is called a *"truck processing time deviation*[1]". Thus, it is important to ensure that the punctuality of the transportation provider is respected. The question is *how* a truck should be processed not only in the dock but also in the parking lot.

In this paper, a truck operations management taking into account the real-time circumstances are modeled and optimized using ant colony algorithm-based multiagent method. The method takes advantage of autonomous and dynamic nature of the Multi-Agent System (MAS) [8] to realize distributed computation and to react flexibly to dynamic events. Ant Colony Optimization (ACO) [3] is proposed to achieve global optimization and to reduce processing time.

The remainder of the paper is organized as follows. Section 2 provides a scope of literature review related to studies about the truck scheduling problem under dynamic environment. Section 3 explains the definition of the problem at hand. Section 4 is especially devoted to describe the ACI integrated in our MAS. In Sect. 5, some test runs are demonstrated to show the capabilities of the proposed model. Section 6 concludes the paper with final thoughts.

2 Literature Review

There are few comprehensive reviews on real-world aspect related to the daily internal operations management of the cross-docking platform. The cross-dock system is decomposed into a set of sub-planning problems that may be solved in sequential fashion: truck loading/unloading, parking lot handling and truck arrival management. Implementing a real-world cross-docks seems unexpected events that happen in these sub-problems. In instance, the arrivals number, the arrival time of trucks, the resources availability and the processing time are external factors prone to uncertainty. Furthermore, the parking lot layout and the flows uncertainty of trucks are among the factors that influence on the covered distance moved by trucks as well as on the docking time [15]. The output of one sub-problem is taken as the input of the subsequent sub-problem. Most of the uncertainty in cross-docks is caused by un-known truck arrival times as confirmed [10].

Most of works are devoted to the congestion problem at the gate of cross-dock facility due to unexpected arrival time of trucks. Boysen [18] define the entrance of a cross-docking area as a gate where inbound trucks are registered and assigned to a parking position on a parking lot. As mentionned by Gue [9], the cross-dock must schedule trucks so as to avoid excessive congestion due to short term storage. Ladier et al. [13] explore the concept of appointment system to register trucks to their preferred arrival times with the aim to being as close as possible

[1] *truck processing time deviation* is considered as a performance measure for a truck scheduling problem as mentioned [12].

to the wishes of the transportation providers. Another goal to maximize the transportation providers' satisfaction, linked to the management of unexpected arrival time of trucks is to minimize the earliness and the tardiness of arrivals. This objective appearing with the consideration of resource capacities in the work of Ladier in which the problem is modeled with an integer program and three heuristics [13,14]. Arabai [4] also considers a multi-criteria scheduling in which primary objective is to minimize earliness and tardiness simultaneously via a unified objective function. Three meta-heuristics are applied for this matter. The case of full uncertainty of arrival times is addressed by a set of papers. In instance, Konur et al. [11] study a cross-dock operator's truck scheduling problem at inbound doors in case of unknown truck arrival times by proposing a biobjective GA. Amini [1] adresses another sort of uncertainty in a truck scheduling problem in a cross-dock facility, in which trucks may confront breakdowns during their service times. The author proposes three multi-objective meta-heuristics to solve the problem. Shakeri et al. [17] address a truck scheduling problem when a lack of the availability of dock doors and material handling systems cause a need to capable plans, and a two-phase heuristic is proposed to solve large-scale problems. Heidari et al. [10] address the problem of incoming and outgoing trucks scheduling at a cross-dock facility, when vehicle arrival times are unknown, through a cost-stable scheduling strategy using two meta-heuristics for solving the designed sample problem.

3 Problem Description

The problem studied is this paper is to schedule inbound trucks using a real-time information about trucks transferring within the cross-dock. Our model chooses on-line the critical truck to be treated at each available parking position and at each dock door.

This problem is modeled as a network as shown in Fig. 1-(a) where the nodes are of three types of physical resources (gate G, parking position P and dock doors D) and where the links are the flows ϕ done by trucks. Each node represents the truck location l_i where $l_i \in \{G_i, P_i, D_i\}$. Every pair of nodes (l, l'), where $l \neq l'$, is associated with a travel time $t_{ll'}$ and a distance traveled $d_{ll'}$ that are symmetrical ($t_{ll'} = t_{l'l}$ *and* $d_{ll'} = d_{l'l}$). At each location, a set of resources assigned to a truck i, indexed by j, where $j \in \{1, 2, ..., n\}$ and n the number of resources of two main types $J = (H, M)$. Each resource j is designated to serve and to handle a set of m inbound trucks I indexed by i where $i \in \{1, 2, ...m\}$. H represents a set of human resources $e_j^i \in \{e_1^1, e_2^1, e_1^2, e_2^2, .., e_n^m\}$ with diversity of skills s_j (specialist, versatile, interim, junior, etc.) based on their training. M is a set of handling equipment. Each resource j has a limited time period at which it is available during the week. We assume that there are a limited number of available resources N^t at each time window t.

3.1 Time Evolution of Truck Process

We imitate a realistic truck scheduling process at the cross-dock that is composed of working gaps. We will introduce a continuous variable $t \in [0, \infty[$ to control time evolution of truck i as shown in Fig. 1-(b). The process of a particular truck i is defined as a time period between its effective arrival time a_i and its departure time d_i. In the case that trucks arrive with or without reservation, we suppose that each truck has a preferred arrival time p_i. We introduce the estimated beginning time b_i for handling a truck i to define the time period of waiting time w_i where $w_i = b_i - a_i$. The handling time h_i of truck i, between its current location l and next location l', is the sum of the travel time $t_{ll'}$ from l to l' and the total estimated process time pt_{ij} spent by a set of resources $j = \{1, .., k\}$ in location $l'$to perform the operation (mainly the unloading/loading operation or the parking operation). k is the number of assigned resources to truck i. We define the completion time $C_i(t)$ as a sum of arrival time a_i, waiting time w_i and handling time h_i of truck i in the cross-dock such as $C_i(t) = a_i(t) + w_i + h_i(t)$. To deal with disturbances, we propose to introduce the slack time s_i in the departure time d_i such as $s_i = [d_i^-, d_i^+]$ where $d_i(t) = C_i(t) + s_i$. A slack time is a source of flexibility that can provide a certain freedom to postpone the start of handling trucks, because trucks should not be ordered too near to each other.

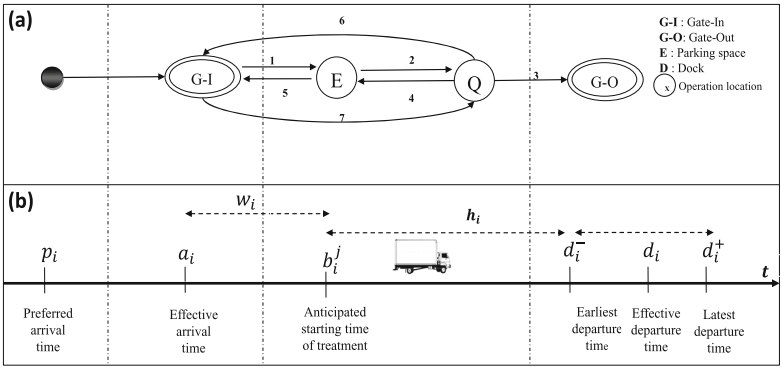


Fig. 1. (a) Representation of trailers treatment process within the crossdock; (b) Time parameters of the truck

4 Multi-agent Based Dynamic Truck Scheduling Model

In this section, the agent coordination mechanism inspired by both foraging and division of labor of ant colony in a MAS is proposed for an efficient dynamic truck scheduling problem. We solve two integrated problems; optimizing resource allocation and truck scheduling. The ACO is a metaheuristic for constructing incrementally the shortest path between agents. The idea of ACO is, at each iteration, the ant moves from one point to another according to a probability depending on traces of "pheromone" and a local heuristic information [3].

The pheromone is important to treat the priority of tasks to be accomplished. We consider in this paper two types of agents: truck agent and resource agent. Thus, each agent deposits pheromones according to the constraints described in the following sub-sections, and make decisions based on the detected pheromone.

4.1 ACI for the Truck Agent

In practice, it is important to define the truck requirements in order to determine its estimated processing time PT_i as well as the maximum number of resource agents Max_i that will handle it. We note that the requirements of TA_i depend on its type (its dimension and its content) and on the set of its operations to be handled in the cross-dock. Furthermore, we note that if a truck agent TA_i has a wished departure time, there will be a deadline for its departure. We denote each truck by its index i to facilitate the notation. But if the completion time $C_i(t)$ of truck agent i ends early or late compared to the due date $Due_i(t)$, we are getting a *"truck processing time deviation"* denoted as D_i^p. So, a penalty d_{CD} will be incurred and it is measured by a time difference between $Due_i(t)$ and $C_i(t)$. This condition must be verified for each truck agent. That's why a slack time s_i is integrated to give more flexibility for the truck handling. Based on PT_i, we can calculate the slack time $s_i(t)$ of each truck agent TA_i.

The pheromone of the truck agent depends on the most critical due date. In another way, the truck agent; that has the least slack s_i before the current due date Due_i; has the highest initial pheromone. The minimum slack time s_i considers the remaining work on a truck i as well as the time remaining until $Due_i(t)$. The minimum slack time is calculated as follows:

$$s_i = Due_i(t) - C_i(t) \tag{1}$$

$$\tau_{RA_i}(t) = s_i \tag{2}$$

The probability is calculated to determine the truck priority. We note that the resource agent has to evaluate the truck priority before accepting to handle it. Otherwise, the question is on which probability the truck will be; selected by resource agent RA_j; to be processed. It is defined as follows:

$$p_{TA_i}(t) = \frac{[\tau_{TA_i}(t)]^{\alpha}.[\eta_{TA_i}(t)]^{\beta}}{\sum\limits_{i \in I} [\tau_{TA_i}(t)]^{\alpha}.[\eta_{TA_i}(t)]^{\beta}} \tag{3}$$

The parameter α is used to control the influence of $\tau_{RA_i}(t)$ where $\alpha \geq 0$. The heuristic function for TA_i depends on the estimated processing time PT_i. We consider that smallest the PT_i is, more the TA_i is critical and has a high-priority. We note that PT_i represents a legal time that a truck agent i must be treated. The heuristic function is denoted as $\eta_{TA_i}(t)$ representing the desirability of a truck agent. It is given as $\eta_{TA_i}(t) = \frac{1}{PT_i}$. The parameter $\beta \geq 1$ is used to control the influence of $\eta_{TA_j}(t)$.

Schedule Construction. To build the truck schedule Sc_i, we need to determine the order of trucks that form the workload of cross-dock per time window. The solution is obtained by the construction of: (1) a list of trucks ordered by priority using Eq. 3 and (2) a list of most proficient resources that serve each truck agent using a steps described in 4.2 and (3) a daily schedule Sc_i per truck agent i. We represent the steps for building a daily truck schedule Sc_i using an activity diagram as shown in Fig. 2.

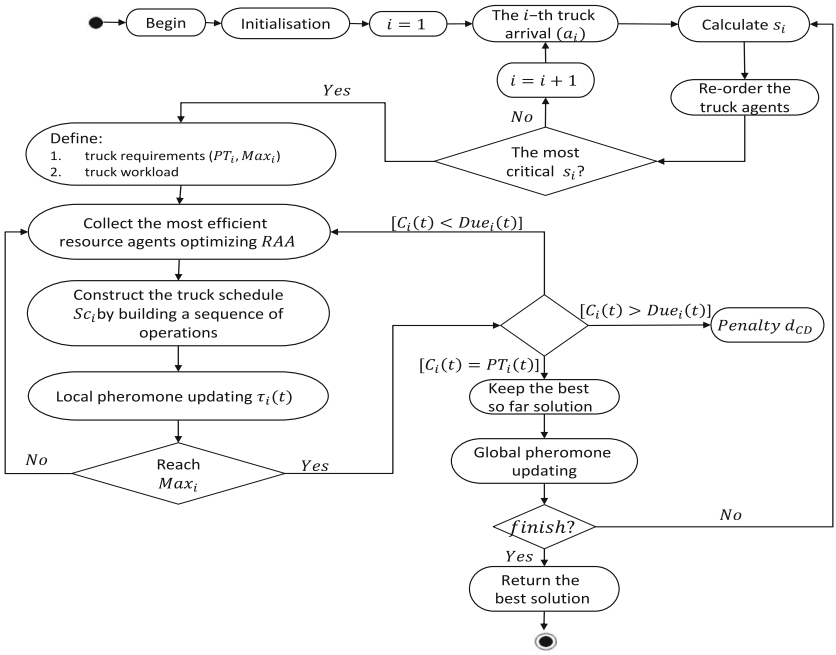


Fig. 2. Flow diagram of the proposed model

4.2 ACI for the resource agent

The assignment of a resource agent RA_j to a truck agent TA_i depends on four properties. We denote each resource agent by its index j to facilitate the notation. The goal is to choose the appropriate resource agents of different types J needed for performing each requested operation that belongs to TA_i. The quantity of pheromone affected to each RA_j depends on the 4-tuple properties of a resource agent $RA_j = x_j(t), Qu_j(t), prof_j, PT_{kl}^j$. After choosing the highest priority truck agent, we evaluate the pheromone value of all resource agents. The resource agent RA_j with the highest value of pheromone $\tau_{RA_j}(t)$ is selected for TA_i depending on the requested operation. The construction of the Resource Agent Allocation (RAA) is based on the following steps:

Step(a). We select a resource agent RA_j from the set of needed resource agents which have not yet been chosen for TA_idepending on the value of pheromone calculated from the following sub-steps:

Step(a-1). The selection depends first on the current status of RA_j inspired from [20]. If $x_j(t) = 1$, implies that RA_j is "$free$". If $x_j(t) = 0$, implies that RA_j is "*unavailable*" or "$buffered$" meaning that RA_j is in maintenance or it reaches the legal normal working hours per day so it will not accept any other operation. Accordingly, if $x_j(t) = 0$, the pheromone quantity of RA_j will be equal to 0 and has no attraction for truck agents.

Step(a-2). The selection of RA_j depends on a qualification property $Qu_j(t)$ inspired from [19]. The RA_j, that has the adequate qualification to perform the operation o requested by a TA_i, is chosen. In instance, a TA_i of *container truck* type can't be stationed in a parking of *light trucks*. The qualification of RA_j is given as $Qu_j = w_k c_k$ and it is represented by one-dimensional matrix. One RA_j may have more than one capability. The set of capabilities of RA_j is given as $w_k c_k = \{w_1 c_1, w_2 c_2, .., w_r c_r\}$; where $k \in \{1, .., r\}$ and the w_k is the weighting coefficient of each capacility. if $w_k = 1$, RA_j has the capability c_k. The number of capabilities r is the same for all resource agents.

Step(a-3). The selection of RA_j depends on a proficiency property inspired from [7]. The number of skills z of RA_j is given as $s_k^j = \{s_1^j, s_2^j, .., s_z^j\}$ where $k \in \{1, .., z\}$. We denote that a list of skills are the same for all resource agents of the same type. We denote that ech truck requires a set of skills $SK_l^i = \{sk_1^i, sk_2^i, .., sk_m^i\}$ where $l \in \{1, .., m\}$. The idea is to evaluate each RA_j by affecting a score $C_k$to each skill s_k^j of RA_j such as $sk_{lj}^i = \sum_{k=1}^{z} s_k^j C_k$ to obtain a level of proficiency in the required skill sk_{lj}^i. We note that $sk_{lj}^i \in [0, 5]$. When $sk_{lj}^i = 0$, RA_j does not have the skill and when $sk_{lj}^i = 5$, RA_j is masterly on that skill. So, the proficiency of $RA_j$is calculated as $prof_{ij} = \frac{\prod_{l=1}^{m} sk_{lj}^i}{5}$.

Step(a-4). Finally, we select the resource RA_j that serves the truck agent TA_i in the shortest processing time PT_i^j.

The resource agent RA_j; that responds more to a four properties; will has the highest initial pheromone. The value of the pheromone $\tau_{RA_j}(t)$ must be updated when the resource has completed a truck or a new truck is added to its workload.

$$\tau_{RA_j}(t) = \frac{prof_j * Qu_j(t) * x_j(t)}{PT_j^i(t)} \tag{4}$$

The probability with which the TA_i selects the RA_j, is a function of the pheromone quantity of the RA_j and the heuristic information associated with the operation $p_{RA_j}(t)$. α and β are the setup items to balance the effects of pheromones and heuristics. The agent resource with the highest probability will has a greater chance of being captured by the truck agent.

$$p_{RA_j}(t) = \frac{[\tau_{RA_j}(t)]^{\alpha}.[\eta_{ij}(t)]^{\beta}}{\sum_{j}[\tau_{RA_j}(t)]^{\alpha}.[\eta_{ij}(t)]^{\beta}} \tag{5}$$

The location of each resource agent is measured by its coordinates (x, y) where$x$, y denote the agent's abscissa and ordinate, respectively. The heuristic function is calculated based on the distance d_{ij} between the location of resource agent j and the location of truck agent i. It is modeled as $\eta_{ij}(t) = \frac{1}{d_{ij}}$ such as the nearest resource agent to truck agent will be chosen.

Step(b). It is necessary to limit the number of resource agents for a truck agent TA_i. The maximum number Max_i for each TA_i must be estimated to avoid a higher communication overhead between resource agents that results to a low efficiency. If the collection of different kinds of resource agents selected to work on TA_i reaches Max_i, the workload assignment process for TA_i is finished. Otherwise, we go to Step(a) to select other resource agents to work on TA_i. After the workloads for all truck agents have been assigned, the procedure for building the RAA is complete.

5 Numerical Results on the Performance of the Model

In this section, we will present the results of some numerical experiments carried out to evaluate the performance of implementing ACI in SMA in order to observe the impact of the input parameters on the solution obtained. In the following figures, each point on the curves are calculated as an average of 5 test runs. The program tests the dock door D as resource agent to serve the m arrival trucks. We have two stop conditions; either because of the unavailability of the docks or because the trucks do not find the adequate qualification in the list of resource agents.

5.1 Tests on the Number of Receiving Docks

The results given in Fig. 3-(a) is obtained for inbound trucks $m = 20$ and the number of receiving dock doors D varying between 1 and 9. The program shows the incremental number on inbound docks decrease the curve of total completion time after $D = 6$. We notice that greater we increase the number of inbound docks, greater the number of treated trucks $rate$ is and greater the total completion time decrease. We observe that increasing a number of resources will increase the performance of the solution and will minimize, at the same time, the total completion time. This is natural because imposing a limited number of availability resources will increase a total completion time.

5.2 Tests on the Number of Inbound Trucks

The test runs are done for $D = 9$ and for $5 \leq m \leq 45$. The results obtained are presented in Fig. 3-(b). The curves give the total completion time and the *rate* versus the number of incoming trucks. From the shapes of the curves, we observe that the curve of the total completion time surpasses the curve of *rate* after $m > 15$, but it is inferior for $m < 10$. Therefore, we interpret that greater the number of incoming trucks is, less the number of treated trucks is and greater the total completion time is considerably higher.

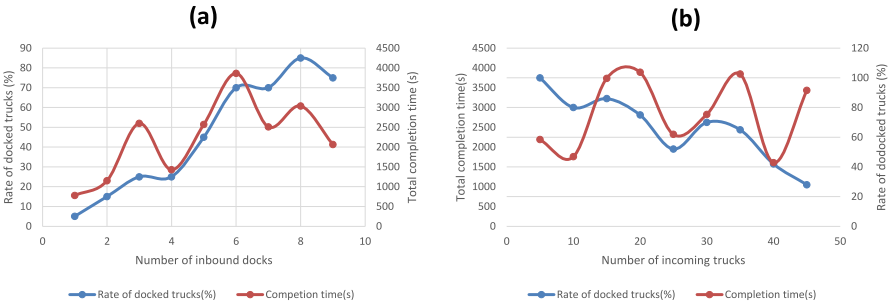


Fig. 3. (a) Execution time as a function of number of inbound dock doors; (b) Execution time as a function of number of inbound trucks.

6 Conclusion and Research Opportunities

In this paper, a new flexible and effective model for dynamic truck scheduling problem has been developed to improve coordination efficiency. The model specifies how the sequence of trucks per priority are processed and how the workload assignment process is optimized. The main contribution of our model is the combination of intelligent agents and ant-inspired coordination that can effectively adapt to the dynamic circumstances. The objective is to solve a truck scheduling problem both in the parking lot and in the dock of a cross-docking platform where truck agents bid for a processing sequence.

The future work will also focus on exploring agent coordination for dynamic re-scheduling in a cross-dock which can provide a schedule immediately and efficiently.

Acknowledgments. These research and innovation are carried out as part of a MOBIDOC Phd thesis as part of PASRI program funded by the EU and administered by the ANPR.

References

1. Amini, A., Tavakkoli-Moghaddam, R.: A bi-objective truck scheduling problem in a cross-docking center with probability of breakdown for trucks. Comput. Ind. Eng. **96**, 180–191 (2016)
2. Apte, U.M., Viswanathan, S.: Effective cross docking for improving distribution efficiencies. Int. J. Logistics Res. Appl. **3**(3), 291–302 (2000)
3. Blum, C.: Ant colony optimization: introduction and recent trends. Phys. Life Rev. **2**(4), 353–373 (2005)
4. Boloori Arabani, A., Fatemi Ghomi, S., Zandieh, M.: A multi-criteria cross-docking scheduling with just-in-time approach. Int. J. Adv. Manuf. Technol. **49**(5–8), 741–756 (2010)
5. Bonney, J.: Non-stop Logistics **36**(5), 63–64 (1994)
6. Boysen, N., Fliedner, M.: Cross dock scheduling: classification, literature review and research agenda. Omega **38**(6), 413–422 (2010)
7. Chen, W.N., Zhang, J.: Ant colony optimization for software project scheduling and staffing with an event-based scheduler. IEEE Trans. Softw. Eng. **39**(1), 1–17 (2013)
8. Ferber, J.: Les Systàmes Multi Agents: Vers une intelligence collective. France, intereditions edn, Paris (1995)
9. Gue, K., Kang, K.: Staging queues in material handling and transportation systems. In: Proceedings of the Winter Simulation Conference, vol. 2, pp. 1104–1108 (2001)
10. Heidari, F., Zegordi, S.H., Tavakkoli-Moghaddam, R.: Modeling truck scheduling problem at a cross-dock facility through a bi-objective bi-level optimization approach. J. Intell. Manuf. pp. 1–16 (2015)
11. Konur, D., Golias, M.M.: Analysis of different approaches to cross-dock truck scheduling with truck arrival time uncertainty. Comput. Industr. Eng. **65**(4), 663–672 (2013)
12. Ladier, A.L., Alpan, G.: Optimisation des operations dans une plateforme logistique: prise en compte des flux d arrivee et de la capacite des ressources internes (2013)
13. Ladier, A.L., Alpan, G.: Crossdock truck scheduling with time windows: earliness, tardiness and storage policies. J. Intell. Manuf. pp. 1–15 (2014)
14. Ladier, A.L., Alpan, G., Greenwood, A.G.: Robustness evaluation of an IP-based cross-docking schedule using discrete-event simulation. In: Industrial and Systems Engineering Research Conference. p. I211. Canada (2014)
15. Marcotte, S., Durand, M., Crainic, T.G.: Amenagement et gestion des flux de la cour d'un centre de distribution: une etude de cas (2013)
16. Schaffer, B.: Cross docking can increase efficiency. Automatic I.D. News **14**(8), 34–37 (1998)
17. Shakeri, M., Low, M.Y.H., Turner, S.J., Lee, E.W.: A robust two-phase heuristic algorithm for the truck scheduling problem in a resource-constrained crossdock. Comput. Oper. Res. **39**(11), 2564–2577 (2012)
18. Stephan, K., Boysen, N.: Cross-docking. J. Manag. Control **22**(1), 129–137 (2011)
19. Wang, L., Wang, Z., Hu, S., Liu, L.: Ant colony optimization for task allocation in multi-agent systems. China Commun. **10**(3), 125–132 (2013)
20. Xiang, W., Lee, H.: Ant colony intelligence in multi-agent dynamic manufacturing scheduling. Eng. Appl. Artif. Intell. **21**(1), 73–85 (2008)

From Traditional Data Warehouse To Real Time Data Warehouse

Senda Bouaziz[1(✉)], Ahlem Nabli[1], and Faiez Gargouri[2]

[1] MIRACL Laboratory, Faculty of Sciences,
Sfax University, 1171 Sfax, BP, Tunisia
bouaziz.senda@hotmail.fr, ahlem.nabli@fsegs.rnu.tn
[2] Tunisia MIRACL Laboratory, Institute of Computer Science and Multimedia,
Sfax University, 1030 Sfax, BP, Tunisia
faiez.gargouri@isimsf.rnu.tn

Abstract. The Traditional data warehouse did not contain data as today. Hence, it is difficult to retrieve these data and treat them. Furthermore, its content is not updated, which may lead to bad decisions. Data are typically loaded from conventional operational systems. Given that today's decisions, in the business world, are becoming more and more in real time. Thus, the system supporting these decisions is to be held. In this paper, we are interested in giving a survey on data warehousing starting from a traditional data warehouse to a real time data warehouse. This survey, focus firstly, on data warehouse architecture. Secondly, it details the changes in the Extract-Transform-Load process to deal with real time data warehousing. Thirdly, it sketches the integration data in the real time data warehouse. Finally, a comparative study concerning the real data warehouse approaching is also presented in this paper.

Keywords: Traditional data warehouse · Real time data warehouse · Near real time extract-transform-load

1 Introduction

A Data Warehouse (DW) is an integrated and historised collection of data generally used to make strategic decisions by means of Online Analytical Processing (OLAP) techniques. It integrates data that come from independent heterogeneous operational data-sources and create a single view about the organization. Once in the DW, these data are turned into structured information that can easily be handled by decision making processes.

Because of the new business requirements for ever up-to-second updated information and since timely data ensure better-informed decisions, real-time data warehousing is one of the savior trends that provide an access to accurate, integrated, consolidated view of the organizations information in real-time [8].

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_46

For a real time data warehouse, no official definition currently exists in the literature. A Real Time Data Warehouse (RTDW) can be defined as a system that represents the characteristics and the actual situation of the organization. For instance, if we have a request to analyze a particular facet of the organization embarked on an RTDW, the answer will be represented in the real state of the organization at the time of the request sending [6]. Unlike most traditional data warehouses, an RTDW is seen to contain current data (real time) of the organization. Thus, the refreshing frequency plays a predominant role in RTDW. Generally, based on the literature [1,3,6] a RTDW undergoes refreshments several times a day, which enables the decision makers to have access to the current data of the organization.

To provide data freshness, performance, availability, prediction capabilities and to offer an integrated information repository to drive and tactical decisions, many approaches are proposed to treat the architectures RTDW and the data integration in the RTDW.

The remainder of this paper is organized as follows. In Sect. 2, we introduce why RTDW is needed and explain the major comparison between the RTDW and the traditional DW. Section 3 presents some research studies that enable us to offer solutions to change the traditional Extract Transform Load (ETL) process to get closer to a RTDW. Section 4 describes some architecture proposed in some works. Section 5 presents some of the proposed approaches that address the problems of the data integration. Section 6 summarizes and discusses all the approaches studied in this paper. Finally, Sect. 7 gives a conclusion and suggests some future research directions.

2 Need for the Real Time Data Warehouse

The data warehouse can be updated either conventionally or in real time. The drawback of the traditional method is that the data content is not updated, therefore, bad decisions may be made. In fact, there is a need for near real-time data warehouse where some data are updated in real time and the remaining data are traditionally refreshed. Thus, there will be an overload on the source system because only critical data will be frequently extracted from the source system or strategic decisions are made using old data. Storing data in near-real time will reduce the latency between business transactions to operational sources and their appearance in the data warehouse. This facilitates the analysis of more recent data and makes decision making faster.

However, data warehouses and business intelligence real time applications have been proposed to answer exactly the types of questions that users would like to ask for real-time data. A Real-Time Data Warehouse enables data to store the data at a time when they are produced and immediately captured, cleaned and stored within the structure of the data warehouse. Then, traditional refreshment cycles are no longer valid. The data warehouse must be able to read the same data that move around the operating systems at the same time of its generation.

The purpose of RTDW is to enable enterprises to rapidly access information and notify the user or decision-making system to react almost immediately to the information. There are two arguments that justify the use of real-time data warehouse. (1) The traditional data warehouse which provides the state of the company at a specific time in the past (every day, every week or every month). However, some applications require a smaller temporal resolution (every hour). The real-time warehouse enables the organization to find variants of its state even within a day. (2) The second argument focuses on the fact that when a large volume of data is entered into the transaction systems in a single day (financial sector); processing ETL sometimes causes problems [1]. The operation of the real time warehouse suggests rather frequent and regular additions in a single day. Table 1 presents a detailed comparison between the traditional data warehouse environment and the real time data warehouse environment.

Table 1. Major differences between a Traditional DW and an RTDW

Traditional data warehousing environment	Real time data warehousing environment
For strategic decisions only	For strategic and tactical decisions
Historical data Updating data periodically	Real-time data
Highly restrictive reporting used to check the existing processes and patterns	Flexible ad hoc reporting and machine modeling to discover new relationships and hypothesis
Results hard to measure	Results measured with operations
Modetare user currency	High number of users accessing
Daily, weekly, monthly data concurrency is acceptable	Only data available within minutes is acceptable

In the following sections, we will present some approaches that can move closer to a real-time data warehouse and other approaches which represent an RTDW architecture and discuss the integration of real-time data.

3 Closer to a Real Time Data Warehouse

In most cases, updating a DW also means putting it out of use (shutdown) during the updating or at least this one makes its use much more difficult and with poor performances. Using a warehouse for an update may also cause some inconsistencies in the results which are returned by the query execution. They will be rarely logical and correct if interviewed data are updated at the same time. The criteria required for continuous updates without involving a shutdown are generally inconsistent with traditional ETL tools. To address this problem,

new solutions specialize in ETL real-time and data loading. There are also solutions modifying conventional ETL systems in order to load a warehouse on a frequency approximating the real-time.

Existing ETL system can be modified to perform real-time or near real-time data warehouse loading. Some of these solutions are described in [3]:

Near Real-Time ETL: This first solution consists in eliminating the real-time reality of the choice if the need is absent. We could simply increase the frequency of loading new data into the warehouse. For example, a load that is usually done on a weekly basis could be executed once a day. This approach may, however, involve shutdown problems. This change would enable the users of the DW to have access to more recent data without having to make major modifications to the loading process or data model. Not being the real-time reality, the near-real time can be a good first low cost solution.

Direct Trickle: If an application requires real-time, the simplest approach would be to continuously load the DW with the new data. In this way, we would eliminate any intermediate storage step. However, this solution can cause a loss of performance when the exploitation of the DW by one or more user(s) since the update of a warehouse itself can require a lot of work from the machine that hosts it.

Trickle and Flip: This approach involves making the second partition of the fact table of the data warehouse on which we make the load. On a periodic basis, we replace the fact table of the warehouse with the second partition, which is responsible for new data. In this way, the searchable fact table is updated at a low frequency to limit the performance loss during its operation. It also contains all the new data since the last update. This solution can be used according to a cycle which can vary from hours to minutes [3].

External Real-Time Data Cache: This approach consists in storing the data in real time in a real-time memory (Real Time Data Cache) external to the warehouse by completely avoiding any potential problem of performance issue. The Real Time Data Cache can be simply another dedicated warehouse loading of the storage and data processing. The applications which handle a big volume of data or which ask for a very short processing time could benefit from this solution. The major disadvantage of this solution is that it involves an additional database that must be installed and maintained. The work done to realize this approach is higher, but the costs that would be spent to buy more efficient equipment or additional memory are justified in many cases by the use of this method.

In [10], the authors proposed some challenges and possible solutions for near real time ETL. They identified two problems for each extraction, transformation and loading phase. In the extraction phase, there are problems, such as the integration of multiple heterogeneous data sources solved using Change Data Capture with the stream processor and data integration tools. The second problem of this phase is the overload data source using a service of updating. In the transformation phase, the authors identified two problems, the first is master data

overhead and the second focuses on the need immediate server to aggregate data. The solution for the first is to maintain a master cache of data and database queue. However, instead of transforming and loading the data, the solution to the second problem consists in switching both tasks, in other words, loading data first, and then transforming them, which could possibly reduce the time consumed for aggregation. During the loading phase, problems are: performance degradation and OLAP internal inconsistencies. Among the solutions who allows to solve these problems are: the staging table (Trickle and Flip), a multiple stage trickle and flip, organizing outside the DW update period, snapshot data, Real Time Data Cache and layer based view have been proposed. A part from this work not much work has been done on the investigation ETL challenges in near real time.

4 Real Time Data Warehouse Architecture

A lot of work has been done on the real time data warehouse architecture. As an example of a proposed architecture, we detail in the following, the approach of Vassiliadis et al., 2009 and Obali et al., 2013.

4.1 Architecture of Vassiliadis et al., 2009

The general proposed architecture called RTDWg is detected in Fig. 1. It deals with a continuous loading as opposed to the periodic basic in traditional

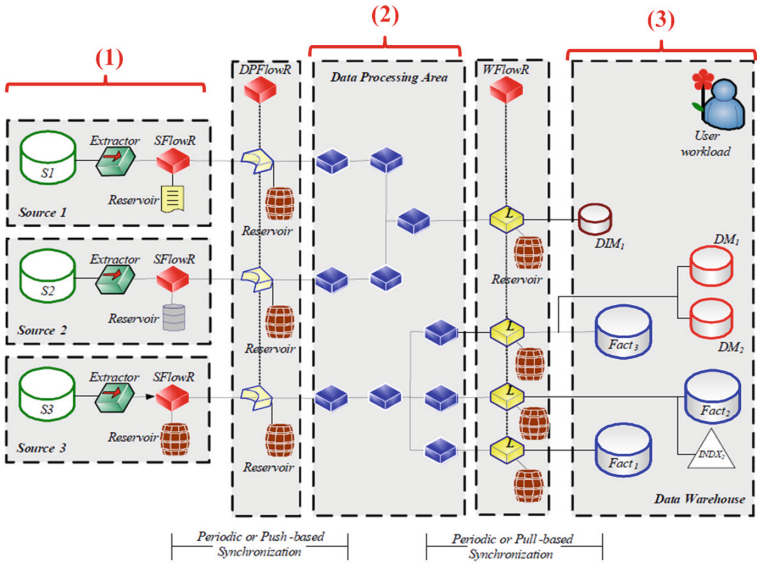


Fig. 1. The Proposed Architecture of near real time data warehouse [9]

approaches. It is composed of three main parts: (1) Data source hosting the data production systems that populate the data warehouse, (2) An intermediate Data Processing Area (DPA) where the cleaning, extraction and transformation of the data takes place and (3) the data warehouse.

Each data source hosts a Source Flow Regulator (SFlowR) module that is responsible for the identification of relevant changes and propagates them towards the DW. Then, the Data Processing Flow Regulator (DPFlowR) module is responsible for deciding which source is ready to transmit the data. A Warehouse Flow Regulator (WFlowR) orchestrates the propagation of data from the DPA to the warehouse. This propagation is based on the current work load from end users posing queries and the requirements for data freshness, ETL throughput and query response time [9].

4.2 Architecture of Obali et al., 2013

The proposed RTDW assures the access to acquired data from different data sources in near real time [7]. The components of this architecture are: Web Service Client, Web Service Provider, Metadata, ETL, Real Time Partition, Data Warehouse and Real Time Data Integration (Fig. 2).

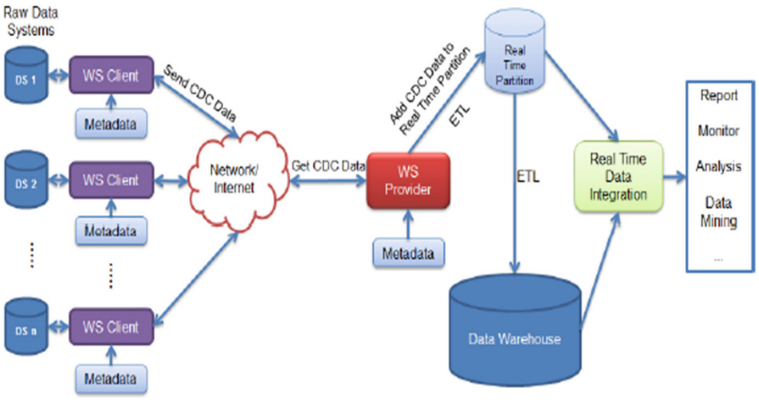


Fig. 2. Web services based real time data warehouse architecture [7]

Web Service (WS) Client is used to get the data changes (known as Change Data Capture - CDC) from raw data systems and send them to the Web Service Provider by calling the related web service. The data that is required to DW are collected from the operational systems and other external sources. Data capturing (CDC) techniques in use include: source data extraction, log capture, triggered capture, application-assisted capture, timestamp-based capture, and file comparison capture [7].

A Web Service Provider is basically a web service which gets data sent by RTDW Web Service Client and adds them to Real Time Partition. This component gets Data Transfer Object which is sent by RTDW Web Service Client, and decomposed them into two parts: data and metadata. The RTDW Web Service uses metadata to generate the Structured Query Language (SQL) via SQL-Generator to insert the data into RTDW log tables then execute this generated SQL on RTDW database and insert the data. In the Real Time Partition, the authors used three stages: the first consists in putting the CDC data into a related warehouse log table. The second step consists in cleaning the CDC log data on demand and the last step is the aggregation of CDC data cleaned on demand.

The component real time data integration is used to integrate the data both in Real Time Partition and Data Warehouse. When a user sends a query to this component; if query only wants historical data, then this component sends the query to Data Warehouse; if query wants both historical and instant data, then this component rewrites the query to get and integrate data [7].

5 Approaches that Treat Real Time Data Integration

To make data integration near real-time, different approaches [2,4,5] have treated the integration of data in real-time data warehouse.

5.1 Approach Lebdaoui et al., 2013

In [4], the authors handled the problems of data integrity issues towards the need for accessing to the RTDW and introduced an IA-RTDWg model that preserves, at the same time, the data integrity and availability. In this work, the authors discussed the integrity of data regarding the violation of RTDWg time constraints. They also presented their model which is based on a temporary duplication of fact tables that are concerned by the change of data, while keeping their integrity constraints. To manage the change in real-time data, RTDW uses specific criteria or measures as: real-time tables, real time partition and intelligent search tools to prioritize the integration of data according to their importance.

The RTDW systems must ensure data integrity. If someone has access to the data just captured by the CDC mechanism or transformed by ETL tools and modifies illegally them, therefore integrity is stripped. If a program or a tool in a DW solution has not processed the data according to predefined rules, and if DW step is ignored or skipped, integrity can also be corrupted.

5.2 Approach Ferreira et al., 2013

On the basis of the industry and theoretical ideas, [2] defined a data warehouse architecture for the constant data integration without compromising the performances of the requests, and assessed its ability to provide real-time. In this

real-time data warehouse, the authors defined a dynamic storage component and a static storage component to represent the recently data integrated and the rest of the data, respectively, with appropriate choices regarding how the components merge. They also proposed design choices concerning the calculation of the query mechanisms and estimated the alternatives to conclude what is the most effective implementation of these mechanisms.

The authors proposed the RTDW architecture from a basic DW architecture. This approach enables to updates the continuous data while keeping the most recent information in DW and enabling any desired degree of freshness to queries. The architecture is made of three main components. These are Dynamic Data Warehouse (DW-D), the Static Data Warehouse (S-DW), and fusion. In [2], the authors also proposed a solution that allows the integration of new data in real time. The solution involves effective loading data into a component that holds the most recent information and offers fast integration mechanisms, D-DW. This component was created to allow a very fast rate of integration. Queries run quickly and introduce the low-efficiency charging, due to small amounts of data. This D-DW component can run in memory. However, a data warehouse is also an extensive set of historical data, and the processing of such data is a task with a high cost. These historical data are stored in the S-DW component and, given that the D-DW component operates independently of S-DW. With this architecture (Fig. 3.), it is possible to provide guarantees of constant integration, new data and, at the same time, access to both the most and the least, without significant additional performance cost.

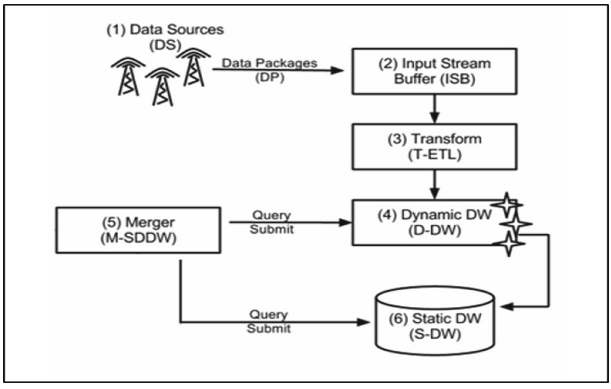


Fig. 3. Sketch of proposed RTDW [2]

5.3 Approach Lebdaoui et al., 2014

For quick and effective decisions for the future, it is necessary that the data warehouse reflects the real changes in operational data and provides the freshest

data for analysis systems. In [5], the authors presented a new method of improving data integration, while preserving the data integrity respecting the existing safety rules. They treated the problem of the integration of large data in the data warehouse in a short time and proposed a new model called Divide-Join Data Integration (DJ-DI). This model, which is built on dividing the volume of data changes and the duplication of the fact table, shows a real improvement in data integration rate compared with the normal configuration. The authors conducted various simulations DJ-DI model as an experimental platform. The results show that the DJ-DI model offers a remarkable improvement of data integration rate.

6 Discussion

It should be noted that, the majority of the presented studies of a RTDW did not specify the ETL real time architecture. Only the work of Wibowo A., [10] is specialized in the identification of the problems and their solutions at the level of every phase in the ETL process. The works [2,4,5] treat the integration data in

Table 2. Summary of the literature review

Paper	Proposed approach	C1	C2	C3	C4	C5	C6
Obali, N., et al., 2013	Construction a RTDW based on the Web services	Yes	Not treated	Yes	Change Data Capture Log Table	No	No
Lebdaoui, I., et al., 2013	Presentation a new model IA-RTDWg to protect at the integrity and the accessibility of the data at the same time	Yes	Not treated	No	Change Data Capture Temporary Table Real Time partition	Yes	No
Lebdaoui, I., et al., 2014	Proposal a new model called DJ-DI to handle the problem of the integration big data in the data warehouse in a short time	Yes	Not treated	Yes	–	Yes	No
Wibowo, A., 2015	Identification of the problems and their solutions at the level of every phase of the ETL process	No	Near Real Time	No	–	No	No
Ferreira, N., et al., 2013	Proposal a solution for the real time storing, with the capacity of integration of constant data and/or at high speed	Yes	Not treated	Yes	Change Data Capture	Yes	No

the RTDW. Furthermore, all the presented approaches are concentrated on the architecture of the RTDW and not detailed the design and the implementation of these architectures. Table 2 gives a summary of the literature review which is based on some criteria: C1 (RTDW Architecture), C2 (ETL Real Time), C3 (Prototype), C4 (The techniques used in the approach), C5 (Data Integration), C6 (Modeling approach).

There is no work that models the various stage of ETL real-time and real-time data warehouse with the platform Talent Real-Time Big Data[1]. This platform simplifies the integration in real time for big data and advanced analytics and cases of real-time use of that guide innovation in the company. Talend Real-Time Big Data Platform generates native code; so we can use it immediately with Apache Spark and Spark streaming today. As part of the fabric data solution, Talend Real-Time Big Data Platform provides high-level, memory fast data processing to put more data in real-time business decisions, to scale. Moreover, all these approaches do not deal the modeling step of RTDW and ETL real time before the implementation task. Therefore, we can use the Real-Time Unified Modeling Language (Real Time UML) standard to model our data warehouse real time before building Talent Real Time Big Data platform.

7 Conclusion

In this paper, we deal with a survey about data warehousing real time by indicating the problems and the solutions to move closer to this RTDW. This survey focuses on the literature review of existing data warehouse architecture. Then, it details the changes in the ETL process to deal with real time data warehouse systems, and treats the integration data in the RTDW. As future work, we will propose a general architecture that includes the construction of a RTDW and we will model the RTDW with Real-Time UML Standard. We think about a loading strategy to RTDW and real-time ETL process taking into account the specificities of data collected in real time. We will develop a RTDW under Talend Real Time Big Data.

References

1. Basu, R.: Challenges of real-time data warehousing (2003). www.dmreview.com
2. Ferreira, N., Martins, P., Furtado, P.: Near real-time with traditional data warehouse architectures: factors and how-to. In: Proceedings of the 17th International Database Engineering & Applications Symposium, IDEAS'13, New York, NY, USA, pp. 68–75. ACM (2013). http://doi.acm.org/10.1145/2513591.2513650
3. Langseth, J.: Real-time data warehousing: Challenges and solutions (2004). http://dssresources.com
4. Lebdaoui, I., Orhanou, G., ElHajji, S.: Data integrity in real-time data warehousing. In: Proceedings of the World Congress on Engineering, vol. 3 (2013)

[1] https://fr.talend.com/products/real-time-big-data.

5. Lebdaoui, I., Orhanou, G., Elhajji, S.: An integration adaptation for real-time datawarehousing. Int. J. Softw. Eng. Appl. **8**(11), 115–128 (2014)
6. Levy, E.: What is real-time data warehousing? (2005). www.dmreview.com
7. Obali, M., Dursun, B., Erdem, Z., Grr, A.K.: A real time data warehouse approach for data processing. In: Signal Processing and Communications Applications Conference SIU 2013 21st, pp. 1–4 (2013)
8. Santos, J.R., Bernardino, J.: Real-time data warehouse loading methodology. In: Proceedings of the 2008 International Symposium on Database Engineering & Applications, IDEAS'08, New York, NY, USA, pp. 49–58. ACM (2008). http://doi.acm.org/10.1145/1451940.1451949
9. Vassiliadis, P., Simitsis, A.: Near real time ETL, pp. 1–31. Springer, Boston (2009). http://dx.doi.org/10.1007/978-0-387-87431-9_2
10. Wibowo, A.: Problems and available solutions on the stage of extract, transform, and loading in near real-time data warehousing. In: 2015 International Seminar on Intelligent Technology and Its Applications (ISITIA), pp. 345–350, May 2015

Intelligent Traffic Congestion Prediction System Based on ANN and Decision Tree Using Big GPS Traces

Wiam Elleuch(✉), Ali Wali, and Adel M. Alimi

REGIM-Laboratory: Research Groups in Intelligent Machines,
National Engineering School of Sfax (ENIS), BP 1173, 3038 Sfax, Tunisia
{wiam.elleuch.tn,ali.wali,adel.alimi}@ieee.org

Abstract. Road traffic jam is one of the major problems related to transportation field in big cities around the globe. The main purpose of our article is providing drivers with an intelligent system to predict the congestion states of the roads. In this paper, we present the architecture of our intelligent congestion prediction system based on the ANN and the data fusion. Our system does not take into account only the historical GPS data but also the real time unpredictable events which have impacts on traffic jams such as accidents. ANN has demonstrated its efficiency in forecasting traffic congestion. The fusion of predicted congestion state, the real time GPS information and the anomalous events using decisional tree has more improved the results. A real time mobile application is provided to drivers in order to help them to discover the traffic state of their destination. The model has been evaluated and validated using big GPS datasets gathered from vehicles circulating in very crowded urban city in Tunisian territory.

Keywords: Big GPS data · Smart congestion prediction system · Artificial neural networks · Real time mobile application · Decision tree

1 Introduction

Growth in the number of vehicles in the last century is one of the most important factors leading to serious traffic jam problems in practically all the countries of the world and especially the developing ones. Traffic congestion can result in several problems such as reducing the traffic system efficiency and increasing the travel time and the delays to reach the destination. In addition, it can harm the drivers' healths. In one hand, vehicles spend more than the needed time in the roads to arrive to their target due to the traffic jams. Therefore, cars and other means of transportation are found as the main contributor in air pollution in 2013 by emitting carbon monoxide and several dangerous gases [1]. In other hand, waiting in road jam queues increases the stress behind the wheels and causes heart attack for road users. Moreover, road jam increases the accidents

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_47

and as a consequence the number of deaths increases. Hence providing road users with a robust system able to predict the traffic congestion state in a particular road becomes a necessity.

Our aim in this paper is to provide an intelligent system able to forecast a real time traffic state based on Global Positioning Systems (GPS) traces. Our system should deal with structure of the roads and its different components. In fact, heterogeneous variables can intervene to define road congestion state such as traffic volume and density, speed of vehicles. All those components can influence the traffic congestion. To improve the efficiency of our system and guarantee better performances, we explore the neural networks capabilities in predicting congestion state. In fact, Artificial Neural Networks (ANN) is a well-known powerful methodology able to deal with complex problems with dependent inputs, and to bridge the human knowledge and the mathematical models. However, several unpredictable traffic components have important impacts on road congestion states such as the accidents. Therefore, we based our system model on fusion of the predicted congestion state from ANN and those real time unpredictable events using decisional module.

The remaining parts of the papers are organized as follow: Sect. 2 introduces a summary of previous related methods used in collection of data and some studies on the traffic flow prediction. Section 3 presents an overview of the collected big GPS database. Section 4 focuses on the architecture of our intelligent congestion prediction system and details the process of forecasting the traffic state based on the ANN and data fusion. Results will be presented in Sect. 5 as well as the mobile application provided to drivers. Finally, we conclude our paper in Sect. 6.

2 Literature Review

2.1 Review of Data Collection Techniques

Some initiatives in the literature to deal with congestion is the improvement of road traffic management using some Intelligent Transportation System (ITS) applications. According to our study, some interests focus on the measurement of road capacity, vehicle speeds and traffic flow using different methods and instrumentations such as the installation of fixed sensors like induction loop detectors [2] or video surveillance [3,4] with vision processing capabilities or the combination of two techniques [5]. Due to the poor funds destined for the transportation control and planning management and the high prices of installing and maintaining equipments to detect the traffic jams, another cheaper and more efficient alternative is the exploration of the installed GPS in the vehicles in circulation on the cities' roads which was used in several ITS applications [6–8].

2.2 Review of Traffic Forecasting Techniques

Several prediction techniques are used for forecasting traffic state variables, such as speed, volume, travel time, density. Literature study was performed to explore

the ANN usage for traffic flow prediction. In fact, the ANN is a very popular method used for short term traffic forecast since the nineties. Ledoux [9] introduced an urban traffic flow model based on neural networks. He concluded their efficiency in this domain based on several simulations. Dia [10] introduced a new object oriented neural network approach. This method presented basis modelling to establish complex interactions like mixture of supervised and unsupervised learning rules in one network. A combined approach based on Combined Neural Network (CNN) and Principal Component Analysis (PCA) for short term traffic prediction are presented in [11]. Several other studies [12–15] are carried out to integrate different ANN models with variety of inputs and they demonstrate the usefulness and efficiency of using ANN for forecasting traffic state.

3 Overview of the Collected GPS Traces

The GPS data were collected thanks to our tracking system described in [6] from several types of vehicles such as private and public cars, commercial tracks scattered across the 24 cities of Tunisia. The participating vehicles passed through urban and rural regions. Therefore, our data are collected from roads with different degrees of crowdedness and in different time of the day and night.

GPS data are composed of GPRMC frames containing essentially the actual position of the vehicle, its ID, its instantaneous speed and the timestamp of data transmission [16]. The number of the participating vehicles reaches 2853 and the number of the gathered GPS nodes reaches more than 70 millions between January, 2013 and July, 2015 and the number is increasing incrementally. Data are transmitted periodically to our servers each minute whenever the vehicles ignitions are switched on.

4 Smart Congestion Prediction System

To provide drivers with current road traffic state in real time, we designed our intelligent system based on two main phases illustrated in Fig. 1. The off-line phase is composed of the preprocessing of raw GPS database and the prediction of congestion state from historical data. The second phase is the on-line step relying on the data fusion of the detected anomalous events in roads, average speed of vehicles in the required road in real time and the predicted congestion state.

4.1 Preprocessing Module

Raw GPS data should be preprocessed to be easily handled in further analysis. The preprocessing module is composed of three parts. A step of features extraction is performed in order to extract the most relevant information from the raw GPS traces such as speed and timestamp. A decoding step is needed to convert the data into understandable information because transmitted frames received

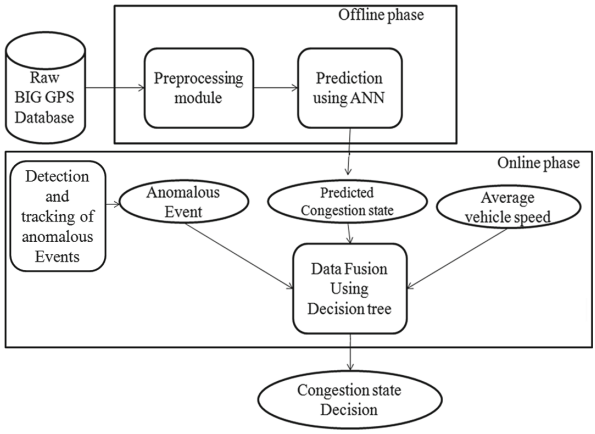


Fig. 1. Overview of our smart congestion prediction system

from GPS systems are encrypted. However, GPS systems can undergo several types of deformations in the transmission process. In order to avoid the possible errors and outliers, we used a thresholding filtration algorithm [17]. To guarantee the precision of our traces, we performed a map matching of our BIG GPS data to the corresponding road segments of Google Maps [18], the well known digital map. A high accuracy is demonstrated compared to the roads provided in the digital map (Fig. 2).



Fig. 2. Map-matching of BIG GPS data to Google Maps after preprocessing

4.2 Training and Prediction Module

The purpose of this module is the prediction of the current state of the road based on the historical GPS data. We used ANN, one of the most popular machine

learning techniques. The ANN was chosen because it can handle several predictors with very complex relationships. Also, ANN is able to take into consideration spatial and temporal information at the same time. The underlying purpose behind this interesting choice is having knowledge about what happened in the preceding time could be valuable for forecasting what will happen in the future. Moreover, it is suitable to our system which deals with very big GPS database. Some basic components are shared between all neural networks. In fact, they are composed of the elementary processing units called "neurons". Their role is getting data from the source of information as input and calculates an output dependent of those inputs by the means of an internal transfer equation, known as activation function. These neurons are linked with each other using some weights. Generally, feedforward ANNs are composed of neurons grouped in layers and connection between neighbouring layers' elements. However, neurons in the same layer are not joined together. Feedforward ANNs contain essentially an input layer, one or more hidden layers and an output layer. Each of which contains a specific number of basic processing units. We used in this study, a feedforward network with a fully connected MultiLayers perceptron (MLP).

After performing different ANN architectures with different number of neurons in the hidden layers, we choose an architecture composed of 3 layers: input, hidden and output. The first layer contains as inputs the GPS data (position, average speed, timestamp). However, the output layer provides the estimated result to the problem which is in this paper the traffic states. The target values are gathered from many expert drivers by giving their opinions in the congestion levels corresponding to the input data. To ensure consistency of target values, we used a huge database composed of more than 5 millions of GPS data divided into training, cross validation and testing process in the ratio 60, 20 and 20 respectively.

Concerning the training phase, we used to train ANNs the Levenberg-Marquardt learning algorithm. The Table 1 depicts the repartition of our big dataset in each process.

Table 1. Repartition of BIG GPS data between training, validation and test

Process	Number of GPS data	Ratio %
Training	3560297	60
Validation	1186766	20
Test	1186766	20

In order to validate the precision of our model, we used the mean absolute percentage error (MAPE) and the root mean square error (RMSE), defined by Eqs. 1 and 2 respectively.

$$MAPE = \frac{1}{N} \sum_{t=1}^{N} \mid \frac{x_i - \hat{x_i}}{x_i} \mid \times 100 \tag{1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(x_i - \hat{x_i})^2} \tag{2}$$

where N is the number of testing nodes, x_i is the ground truth congestion state defined by expert drivers and $\hat{x_i}$ is the predicted congested state by our model.

4.3 Detection and Tracking of the Anomalous Events Module

This section describes the process of detection and tracking of the anomalous events. This module is very special and interesting because it relies on the cooperation between drivers from one side and our system from another. Drivers are directly interactive via our system by warning each other about an anomalous event or incident such as accidents or road works or any strange act which can directly influence the road traffic state. We afford drivers a mobile application to send any special incident faced in the road to aware our system and obviously other drivers who wish to reach their destination via this road.

4.4 Data Fusion Module

Prediction of the future depending on the past information can give a good result. However, sometimes, some events can suddenly occur and disturb the system such as the accidents. In order to improve the prediction of our system, we apply a data fusion method to incorporate those unpredictable events. Also, the real time speeds of vehicles are relevant variables in forecasting of congestion state. The decisional module is based on the decision tree algorithm. The decision depends on the predicted congestion state from historical database of a road in an instant t, the gathered anomalous events in real time in the same road as well as the average speed of vehicles in motion in the required road in (t−5) min. After performing the training of the decision tree, the data fusion module finally provides the decision about the traffic state.

5 Experiments and Results

5.1 Area of Study

In our experiment, we focus our study on the region of Sfax, the second biggest city in Tunisia after the capital. In fact, it was chosen because it is one of the most congested cities in Tunisia. Also, the largest number of GPS datasets existing in our database are centralized in the region of Sfax.

The study area is mainly composed of 9 highways. All the highways are linking the center of Sfax to the famous dual carriageway of Habib Bourguiba as it is shown in the Fig. 3.



Fig. 3. Study area: Sfax road network

5.2 Results and Discussion

In order to obtain the optimal neural network configurations, several criteria were tested carefully. We tested in our experiment the impact of variation of the hidden neurons in the hidden layer from 1 to the 11 nodes (twice the input neurons plus 1 as introduced in the literature).

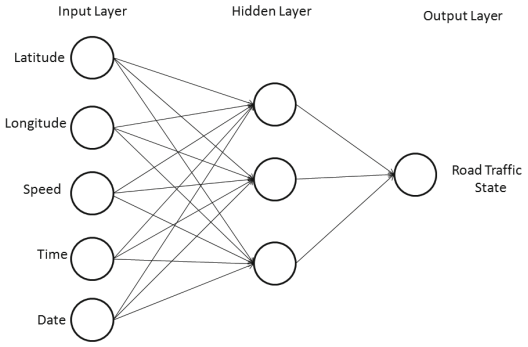


Fig. 4. Structure of the neural network

By comparing RMSE and MAPE, the needed hidden nodes giving the best performance are defined. As it is presented in Fig. 4, the best configuration shows 3 neurons in the hidden layer. The following Table 2 contains the weights of each input. As it is shown in the Table 2 and according to the positive and negative values of weights, we can conclude that the most relevant inputs which has an important influence in the congestion classification are the speed, followed by the time, longitude, date and finally the latitude. This explain the importance of speed to predict the congestion state of a road. However, we remark that the

Table 2. ANN Inputs associated with their weights

	Weight 1	Weight 2	Weight 3
Latitude	−0.0520	−0.0017	−0.0108
Longitude	−0.9007	0.0173	0.1173
Speed	1.0876	36.9880	−6.2685
Date	−0.4371	0.0074	−0.0286
Time	3.3045	0.0055	0.0342

Table 3. Samples of predictors in decision tree technique

Average speed (km/h)	Anomalous event	Predicted congestion state
0	1	1
10	1	0
20	0	1
20	1	1
30	0	0
50	0	0
100	0	0

latitude has not an influence in the prediction because the latitude in the chosen region is nearly the same value.

Concerning the decision tree, we present in the Table 3 some samples of the predictors.

Our model has very high accuracy of 99.03% with RMSE 0.0178 and MAPE 0.0130 compared to the work of Posawang et al. in [19] which reached 94.99%.

5.3 Information System

Our developed application presents an interactive interface containing Google Maps [18] API. It allows the driver to add a notification in the road map and to describe the type of the encountered anomalous event as described in Fig. 5(a). Therefore, this incoming information are taken into consideration in our congestion prediction system and broadcast to corresponding drivers.

The road state displayed in the map are enabled corresponding either to the driver location gathered from its GPS coordinates or corresponding to a response of our system to a specific request performed by the driver. As a consequence, a real time state of the traffic will be displayed as illustrated in the Fig. 5(c). Therefore, our system guidance helps the drivers to have an idea about state of a road. So, they can choose another alternative to reach their destination and overcome the eventual congestion and the resulting time and money wastage.

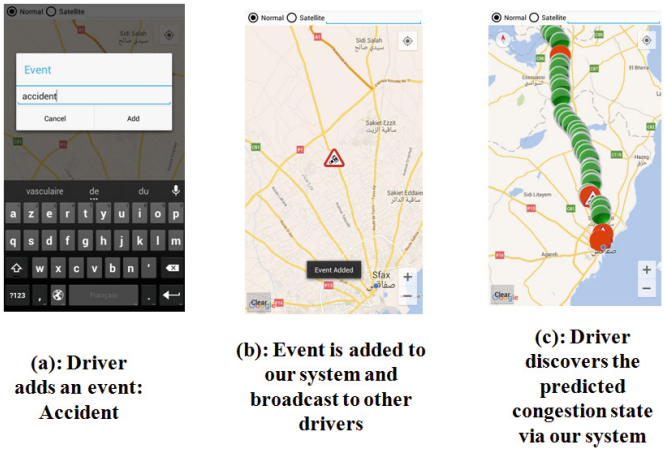


Fig. 5. Real time traffic congestion state via mobile application

6 Conclusion

This paper presents the architecture of our developed system which is able to predict the state of the congestion in real time.

Several contributions are provided in this paper including the big database gathered from Tunisian cities. In addition, we provide road users with an intelligent system able to forecast the future state of the needed road thanks to the developed mobile application.

Our system offers the drivers the ability of consulting the real time traffic state via a mobile application. Thanks to our system, drivers are able not only to discover the state of the road, but also the anomalous events that can eventually occur in the roads and have impacts on the traffic state.

Acknowledgement. This work was carried out within the framework of a PhD MOBIDOC (PASRI program), funded by the EU and managed by the ANPR. The authors would like also to acknowledge the partial financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program.

References

1. Union of Concerned Scientists, Vehicles, Air Pollution, and Human Health. http://www.ucsusa.org/clean-vehicles/vehicles-air-pollution-and-human-health
2. Gramaglia, M., Trullols-Cruces, O., Naboulsi, D., Fiore, M., Calderon, M.: Vehicular networks on two Madrid highways. In: 2014 Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), pp. 423–431. IEEE (2014)
3. Maze, T., Schrock, S.D., Kamyab, A.: Capacity of freeway work zone lane closures. Work **6**(8), 12 (2000)

4. Sarasua, W., Davis, W., Clarke, D., Kottapally, J., Mulukutla, P.: Evaluation of interstate highway capacity for short-term work zone lane closures. Transp. Res. Rec. J. Transp. Res. Board (1877), 85–94 (2004)
5. Greenberg, H.: An analysis of traffic flow. Oper. Res. **7**(1), 79–85 (1959)
6. Elleuch, W., Wali, A., Alimi, A.M.: Mining road map from big database of GPS data. In: 2014 14th International Conference in Hybrid Intelligent Systems (HIS), pp. 193–198. IEEE, December 2014
7. Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C.: Short-term traffic forecasting: where we are and where were going. Transp. Res. Part C Emerg. Technol. **43**, 3–19 (2014)
8. Yao, B., Chen, C., Cao, Q., Jin, L., Zhang, M., Zhu, H., Yu, B.: Short term traffic speed prediction for an urban corridor. Comput. Aided Civ. Infrastruct. Eng. (2016)
9. Ledoux, C.: An urban traffic flow model integrating neural networks. Transp. Res. Part C Emerg. Technol. **5**, 287–300 (1997)
10. Dia, H.: An object-oriented neural network approach to short-term traffic forecasting. Eur. J. Oper. Res. **131**(2), 253–261 (2001)
11. Zhang, X.L., He, G.G.: Forecasting approach for short-term traffic flow based on principal component analysis and combined neural network. Syst. Eng. Theory Pract. **27**, 167–171 (2007)
12. Gao, Y., Sun, S.: Multi-link traffic flow forecasting using neural networks. In: 2010 Sixth International Conference on Natural Computation, vol. 1, pp. 398–401. IEEE, August 2010
13. Pamua, T.: Road traffic parameters prediction in urban traffic management systems using neural networks. Transport Probl. **6**, 123–128 (2011)
14. Çetiner, B.G., Sari, M., Borat, O.: A neural network based traffic-flow prediction model. Math. Comput. Appl. **15**, 269–278 (2010)
15. Dunne, M.S., Ghosh, B.: Traffic flow predictions employing neural networks in a novel traffic flow regime separation technique. In: Proceedings of the ITRN 2011, vol. 31 (2011)
16. Elleuch, W., Wali, A., Alimi, A.M.: An investigation of parallel road map inference from big GPS traces data. Procedia Comput. Sci. **53**, 131–140 (2015)
17. Elleuch, W., Wali, A., Alimi, A.M.: Collection and exploration of GPS based vehicle traces database. In: 2015 4th International Conference on Advanced Logistics and Transport (ICALT), pp. 275–280. IEEE, May 2015
18. https://www.google.com/maps
19. Posawang, P., Phosaard, S., Pattara-Atikom, W., Polnigongit, W.: Perception-based road traffic congestion classification using neural networks. In: Proceedings of the World Congress on Engineering, vol. 1, pp. 1–3, July 2009

Forecasting Using Elman Recurrent Neural Network

Emna Krichene[3(✉)], Youssef Masmoudi[3], Adel M. Alimi[3], Ajith Abraham[1], and Habib Chabchoub[2]

[1] Machine Intelligence Research Labs, Scientific Network for Innovation and Research Excellence, Auburn, WA, USA
ajith.abraham@ieee.org
[2] GIAD Unit Research, FSEGS University of Sfax, Sfax, Tunisia
habib.chabchoub@gmail.com
[3] Research Groups in Intelligent Machines (REGIM), National School of Engineers (ENIS), University of Sfax, BP 1173, 3038 Sfax, Tunisia
{emna.krichene.tn,adel.alimi}@ieee.org, youssef.masmoudi@gmail.com

Abstract. Forecasting is an important data analysis technique that aims to study historical data in order to explore and predict its future values. In fact, to forecast, different methods have been tested and applied from regression to neural network models. In this research, we proposed Elman Recurrent Neural Network (ERNN) to forecast the Mackey-Glass time series elements. Experimental results show that our scheme outperforms other state-of-art studies.

Keywords: Recurrent neural networks · Forecasting · Time series · Machine learning

1 Introduction

An important aspect related to forecasting represents the subject of much research today. Thus, Forecasting is an important data analysis technique that aims to study historical data in order to explore and predict its future values.

Hence, a time series is defined as a set of observations *X(t)* measured at successive points in time. Furthermore, estimating future data has to analyze and model past values in order to identify the behavior of the series and so that to determine predictable values at *t+1*, *t+2*, etc… assuming that the future will be like the past.

In fact, over the last few decades, numerous researches have contributed to the literature of forecasting from regression to neural network models [12–16]. However, there is no single method or technique which can be properly applied for all fields because each one has its own advantages and limitations.

Thus, forecasting literature contains a large variety of methods that can be classified into two main families: (i) qualitative methods that rely on opinions of experts and their intuitions and (ii) quantitative methods that are based on quantitative historical data and mathematical models to generate forecasts. In the following, we present some methods in the scope of quantitative methods:

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_48

1. Moving Average (MA): where only the most recent observations are used to calculate the forecast.
2. Weighted Moving Average (WMA): this technique is similar to MA except that all values are logically weighted differently.
3. Exponential Smoothing: is a derivative of the WMA by assigning more weight to the more recent observations; this method is based on a fundamental assumption: each observation at time i depends on the previous observations and the variation between that forecast and the actual value of the series at that point.
4. Linear regression method: which is a model where a dependent variable y is explained linearly by a one or more explanatory variables x_i.

Recently, various methods-belonging to machine learning algorithms - have been presented to the task of forecasting; the most-used method is assuredly artificial neural networks (ANN) as they are a nonparametric and nonlinear methods. Thus, ANNs are able to model and to map complex problems and to train them in order to determine parameters and explore relationships between data [2, 11]. Hence, over the past three decades, ANNs have been considered as a promising and useful technique applied to a diversity of forecasting problems.

However, in ANN, we distinguish two learning methods: (i) supervised and (ii) unsupervised models. The most popular method in time series forecasting is the feed-forward multi-layer perceptron (MLP). This model, which is a supervised neural network, is based on combining several simple neurons to form a network. The set of these neurons are organized into three layer types:

1. The first layer is called the input layer and is composed of inputs' neurons which are stimulated by the received data to be processed.
2. The last layer is the output layer and is composed of outputs' neurons which provide a response (result) to the problem posed at the input layer.
3. And finally the layers between which are the hidden layers and are composed of neurons through which the information provided by the input layer is propagated and processed.

Hence, there are many types of neural networks but the basic principles are very similar. However, every problem needs to be modeled by having its own neural network architecture and network configuration since there is no single learning method or model which can be appropriately used for all fields.

Pursuant to [8], the major challenge in this field is to determine the ideal network architecture giving best results, because popularly feed-forward back-propagation network promotes successful results, but these results could be enhanced by using recurrence, in other words, by reusing previous values. Therefore, the major reason behind using recurrence is often done under the assumption that a network which remembers past or historical values and feedbacks them may be able to determine future ones as patterns may repeat in time.

In this context, we have focused our work on recurrent neural network (RNN) using a nonlinear chaotic system called Mackey-Glass, which represents a good benchmark test as its elements are hard to predict.

The rest of this paper is structured as follows: Neural network architectures are presented in Sect. 2. In Sect. 3, we discussed the learning algorithm. We presented our proposed approach and its results in Sect. 4. Finally, we concluded our paper in Sect. 5.

2 Neural Network Architectures

This Section is divided into two parts: In the first part, we introduce the main principle of the Multi-Layer Perceptron. In the second part we explain the architecture of the Recurrent Neural Network as well as its different architecture types.

2.1 Multi-layer Perceptron

Typically, the Multi-Layer Perceptron (MLP) architecture includes three types of layers: the input layer, the hidden layer and the output layer; the data propagates strictly in only one direction: from input to output layer so that it is so-called feed-forward. The data processing can expand over different layers, but no feedback connections are permitted.

Mathematically, the artificial neurons are combined in order to process information: each neuron in a particular layer is connected with all neurons in the next layer and then summed and multiplied by the appropriate weight w_{ij} which characterizes the connection and the degree of importance between the i^{th} and the j^{th} neuron. Finally, the nonlinear mapping function F controls the amplitude of the resulting output Y which is determined by (1). It holds that:

$$\mathrm{Y}_i = F(\sum\nolimits_{j=1}^{i} w_{ij} x_j). \tag{1}$$

The aforementioned function F is called the activation function. We note that it is very common to use a sigmoid as the transfer function for all the neurons; thus, the transfer function can be computed according to (2).

$$\mathrm{F(X)} = \frac{1}{1 + e^{-cx}}. \tag{2}$$

where c is the adaptive gain parameter of sigmoid function. Thus, during the learning process, one parameter has to be adjusted: the gain parameter c in addition to the appropriate weight w_{ij}.

2.2 Recurrent Neural Network

An important aspect related to recurrent neural network (RNN) represents the subject of much forecasting research today. Among them, authors in [3] proposed to apply RNN and MLP as forecasting tools to be then compared. In the same context, researchers in [9] applied two types of RNN (i.e. [4, 6]). Then they propose to compare the obtaining result to other statistical forecasting methods. In [18], authors applied two different architectures of RNN [6] as a predictor method to the consumer price index (CPI) of

education, recreation, and sports data in Yogyakarta. In the same context, authors in [16] applied a RNN architecture to forecast wave data; experimental results show that the proposed architecture gives more accurate results when the input data has been smoothed. However, unlike the MLP architecture, the RNN is characterized by a dynamic neuron model which serves as a memory; otherwise RNNs uses feedback connections from one or more of its neurons to enable signals from one layer to be sent back to a previous layer [1].

Hence, a set of additional context neurons are added to the input layer that receive input from either the output or the hidden layer neurons. These feedback connections to the context neurons have a fixed weight of unity shown in Fig. 1.

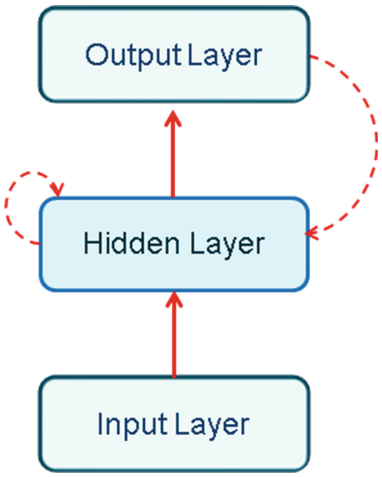


Fig. 1. Typical Recurrent neural network composed of three layers

According to the characteristics of the network connections, connecting several neurons together forms different layers. In fact, RNN has at least one context-layer which serves as the memory of the network [9].

Thus, there are different types of RNN's architectures such as Jordan, Hopfield and Elman networks.

In the following, we will discuss the difference among those architectures.

(a) *Jordan Recurrent Neural Networks (JRNN)*

One of the most known recurrent neural networks is Jordan RNN (JRNN) [6] where three layers are used, with the addition of a context-layer which receives input from the output-layer. This added layer works as a dynamic memory as it saves the previous states of the output-layer.

(b) *Hopfield Recurrent Neural Networks (HRNN)*

Another type of the wide range of recurrent neural networks seems to be interesting to describe is the Hopfield RNN which was developed by [5]. Merely, its principle can be described as follows: all neurons influence each other by updating their activations

values so that they are all both inputs and outputs. We note that in HRNN, the connections between neurons are symmetric and bidirectional.

(c) *RNN with Hidden and Input-Context (Multi-recurrent)*

Further, it is possible to have a context-layer for every layer in order to save past inputs. This structure is later called multi-recurrent [10].

(d) *Elman Recurrent Neural Networks (ERNN)*

Finally, another type of RNN seems to be interesting to discuss is the Elman RNN (ERNN) [4]. The main difference between its architecture and those of JRNN networks is that the context-layer receives input not from the output-layer but from the hidden-layer. The usefulness of this type of recurrence is that ERNN are able to underline the relationships between future values and past ones even if it is hard to learn from and so that to generalize.

Due to their specific structure, ERNN can be a good alternative model to forecast time series even if their elements are hard to predict. However, its structure is believed to have a more powerful ability than other typical RNN architectures to learn and forecast the given data [18]. This point motivates us to explore it in advance in our work.

3 Learning Algorithms

Before an ANN can be used to solve any problem, it must be learned to do so. Basically, learning is the process of adjusting network' parameters which play the key role within the prediction phase in ANN. Hence, the learning process is described as follows: After computing the output by the neural network, a corresponding desired value is transmitted to the output and an error is calculated. Pointing out that an error is computed as the difference of the target and the system output. The resulting error represents an estimation of the predictions quality of the trained network and is sent back to the system to adjust its parameters. Thereby, this process is repeated in such way the obtained output and the desired one are as close as possible and then is accepted. This formalism is called the learning process.

We note that both MLP and RNN adopt the same process of learning algorithm except that in the case of RNN, hidden contexts have the ability to be processed in the same way as well as hidden and output layers.

Hence, the input layer is composed of two parts: the first part is the real input representing the input data, and the second part is the context input which is a copy of the hidden units from the previous time step. For that reason, RNN seems dynamically very rich in the way that it is able to save previous states to use them in the future.

Graphically, this architecture is more described in Fig. 2.

Mathematically, the aforementioned Sigmoid function (see (2)) is used as the activation function for both the hidden layer and the output layer. As shown in Fig. 3, the output $y_i(t)$ can be calculated using the weighted sum of its input as expressed in (3),

$$y_i(t) = F\left(\sum_{j=1}^{i} w_{ji} h_j(t)\right). \tag{3}$$

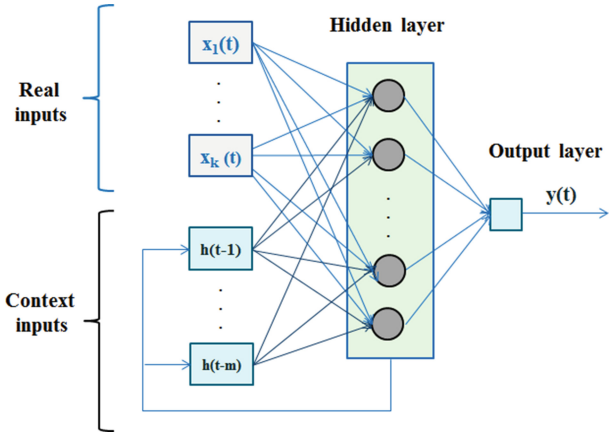


Fig. 2. Architecture of elman recurrent neural network

where F is the Sigmoid function (2), w_{ji} is the connection's weight between the j^{th} hidden unit h_j and the i^{th} output.

In the same way, the output of the j^{th} hidden unit h_j is so calculated except that it computes context units in addition to the true inputs as illustrated in (4).

$$h_j(t) = F(\sum_{i=1}^{k} w_{ij}x_i(t) + \sum_{m=1}^{n} v_{mj}h_m(t-1)). \tag{4}$$

where w_{ij} is the connection strength between the j^{th} hidden unit h_j and the i^{th} real input x_i and v_{mj} is the connection's weight between the j^{th} hidden unit h_j and the m^{th} hidden past value at time $t-1$.

Basically, the connection strengths between hidden past values and their corresponding actual ones are fixed to one as we suppose retaining past values without applying any modification; however experiments show that updating connection strengths between context units and hidden units v_{kj} in the same way as w_{ij} gives better results than fixing them at *1*.

Thereafter, serving by the back-propagation learning algorithm, all connection weights in the network are simply computed and updated according to (5),

$$\Delta w_{ij} = \alpha \frac{\partial E}{\partial wij}. \tag{5}$$

where α is the learning rate and E is the calculated error between the target output y_i and the estimated one $\hat{y}_i$. It holds that:

$$E = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i). \tag{6}$$

We note that the back-propagation algorithm aims to minimize E and adjusts the connection weights depending on. Therefore, it is simple to calculate the new weight which is updated as follows

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}. \tag{7}$$

In the same way, the connection weights v_{kj} are so updated.

4 Experimental Work

This section is divided into two parts: In the first part, we present a brief introduction about the data used in our study. In the second part we explain our architecture and we compare our results to the work of [3, 17] as they used the same benchmark which would allow them to be compared and evaluated.

4.1 Data

This study applied the Elman RNN on a set of time series provided from the Mackey-Glass Chaotic Time series. Being proposed by [7], the Mackey-Glass time series represents a good benchmark for testing the performance of the proposed forecasting model as its elements are hard to forecast. It is defined by a simple dynamic equation that can be written as:

$$x(n-1) = \frac{1}{1+b}[x(n-1) + \frac{ax(n-\tau)}{1+x^{10}(n-\tau)}]. \tag{8}$$

where a and b are constants, and τ represents the time delay.

In this paper, we assume that a and b are set to *0.2* and *0.1* respectively, $\tau = 17$ and time step size $t_0 = 0.1$ s.

Our experiments aim to forecast properly the future values of the time series using the actual present value in addition to p past others. As proposed in [3], the simplest and standard method for this type of forecasting is to convert the dynamic equation into a mapping function that can be expressed as:

$$x(t+m) = f(x(t), x(t-\tau), x(t-2\tau), \ldots, x(t-p\tau)). \tag{9}$$

where m is the future instant we want to forecast in, and p the number of the time delay denoted τ.

In this study, we choose to predict the 6^{th} future value using four past observations and the actual point.

4.2 Results

Since ANN is very flexible, there is no method permitting building the reliable architecture from the first test. Thus, achieving a correct neural model requires selecting a

large number of parameters experimentally over trial and error such as the number of iterations, the number of hidden units, weight initialization and the learning rate, etc… Therefore, building the architecture begins with identifying both inputs and outputs.

To start, strength connections are randomly initialized in the interval [−0.001, 0.001]. To facilitate the learning process, the output of the Mackey-Glass was normalized in [0, 1], which is the range of the sigmoid function. Hence, as we aim to forecast only one future value by using four past values in addition to an actual point, it seems clear that we need to define only one neuron in the output layer. Whereas we need five neurons in the input layer: one actual input neuron plus four past values neurons. Once input layer and output layer were defined, trials and tests start to define the proper number of hidden neurons. After performing different tests (5, 7, 9, 12, 50 and 100 hidden neurons), the best result was given with using 100 hidden neurons.

Experimentally, and in order to improve the performance of our architecture with new cases, we divided the data into two sets: the first one is the training sets which represents 50% from the data (500 samples) and the second one is the test sets and it represents also 50% from the data.

During the training of our network, weights were calculated in each iteration using the back-propagation learning algorithm. The aim is to minimize the normalized root mean square error (NRMSE) from iteration to another until reaching an error goal already fixed at the beginning of the training as the satisfied criterion. The formula of NRMSE is expressed in (10),

$$NRMSE = \frac{1}{\varphi}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \tilde{y}_n)^2}. \tag{10}$$

where $\varphi = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - O)^2}$ and $O = \frac{1}{N}\sum_{n=1}^{N} y_n$

where y_n denotes the actual output and $\tilde{y}_n$ represents the estimated one.

To obtain a more reliable and accurate result, network parameters such as the learning rate, the number of the hidden units and the context neurons were memorized only in the case of new lower error is performed. Results are shown in Table 1.

Table 1. Prediction of Mackey-Glass time series using elman recurrent neural network

	Number of hidden neurons	Learning step	NRMSE learning	NRMSE test
Our results	100	13928	0.0165	0.047
[3]	7 and 10	14770	0.035	0.087
[17]	12	1500	0.3	0.5

During the training phase, different parameters' values were experimented; Epoch was fixed at *30000* and minimum error goal at *0.0165* but the best result was obtained at the *13928* iteration. Thus, a system performance cannot be concluded only from the training accuracy, but also on examining new data that are not experimented during the learning phase; hence a new set of data (*500* samples) was given to our system to test the accuracy of our model. Therefore, it can be seen that our architecture outperforms

the performance of both [3, 17] during the learning phase and the test phase as illustrated in Table 1.

Figure 3 depicts the example of 500 samples from the Mackey-Glass time Series predicted by our network architecture. We remark that the proposed scheme gives accurate results as they are predicted efficiently.

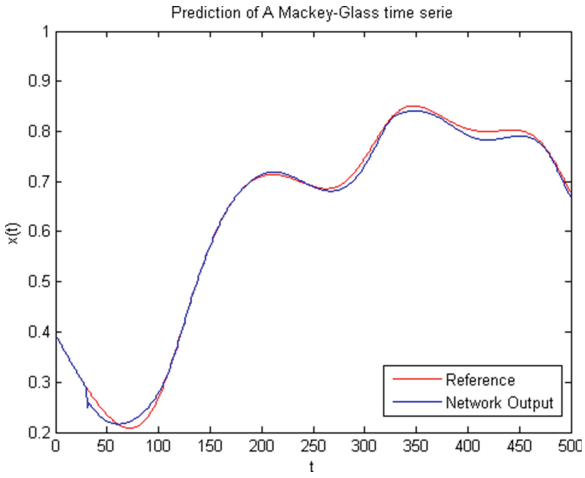


Fig. 3. Prediction of Mackey-Glass time series using elman recurrent neural network (test results).

5 Conclusion

The task of Forecasting was examined with different methods where Artificial Neural Network is one of them. Otherwise, there are many types of ANN such as Recurrent Neural Networks architectures. In this study, Elman RNN was used to predict one of the most known time series which is Mackey-Glass time series. For this purpose, different architectures were trained in order to adjust network' parameters which play the key role within the task of Forecasting. However, testing results show that comparing to [3, 17], our architecture gives us more accurate results.

Hence, it is interesting to mention that the best results were obtained by initializing randomly the weights of the context units.

In summary, this paper justifies the assumption aforementioned at the beginning of this paper which supposes that RNN can be a good alternative model to handle hard time series as it is well suited to memorize past values to use them as memory within the training phase.

Acknowledgments. The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program.

References

1. Ammar, B., Cherif, F., Alimi, A.M.: Existence and uniqueness of pseudo almost-periodic solutions of recurrent neural networks with time-varying coefficients and mixed delays. IEEE Trans. Neural Netw. Learn. Syst. **23**(1), 109–118 (2012)
2. Cao, Q., Ewing, B.T., Thomson, M.A.: Forecasting wind speed with recurrent neural network. Eur. J. Oper. Res. **221**, 148–154 (2012)
3. Danko, B., Tomislav, B., Dubravko, M., Josip, K., Branko, N.: A comparison of feed-forward and recurrent neural networks in time series forecasting. In: IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr) (2012)
4. Elman, J.L.: Finding structure in time. Cogn. Sci. **14**, 179–211 (1990)
5. Hopfiled, J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Nat. Acad. Sci. USA **79**, 2554–2558 (1982)
6. Jordan, M.: Attractor dynamics and parallelism in connectionist sequential machine. In: Proceedings of the Eighth Conference of the Cognitive Science Society (1986)
7. Lapedes, A.S., Farber, R.: Nonlinear signal processing using neural networks: prediction and system modeling, Los Alamos National Laboratory, Los Alamos, New Mexico, Technical report (1987)
8. Lawrence, R.: Using neural networks to forecast stock market prices, Department of Computer Science, University of Manitoba, Winnipeg, Canada, Technical report, December 1997
9. Mortiz, M., Stefan, L., Stefan, V.: Sales forecasting with partial recurrent neural networks: Empirical Insights and Benchmarking results. In: 48th Hawaii International Conference on System Sciences (2015)
10. Ulbricht, C.: Multi-recurrent networks for traffic forecasting. In: Proceedings of the Twelfth National Conference on Artificial Intelligence (AAI-94), Seattle, USA, vol. 2, pp. 883–888 (1994)
11. Wallace, M.P.: Neural networks and their applications to finance. Bus. Intell. J. **1**, 67–77 (2008)
12. Mustfaraj, G., Lowry, G., Chen, J.: Prediction of room temperature and relative humidity by autoregressive linear and non-linear neural network models for an open office. Energy Buildings **43**(2011), 1452–1460 (2011)
13. Guo-Rui, J., Pu, H., Yong-Jie, Z.: Wind speed forecasting based on support vector machine with forecasting error estimation. In: Proceedings of the Sixth International Conference on Machine Learning and Cybernetics (2007)
14. Brooks, C.: Introductory Econometrics for Finance. Cambridge University Press, Cambridge (2002)
15. Florance, M.M., Sawicz, M.S.: Positioning sales forecasting for better results. J. Bus. Forecast. **12**, 15–17 (1993)
16. Kostas, H., Parth, P., Shainee, K., Rick, S.B., Alberto, J.L.: Multi-step forecasting of wave power using a nonlinear recurrent neural network. In: IEEE PES General Meeting (2014)
17. Park, D.C.: A time series data prediction scheme using bilinear recurrent neural network. In: 2010 International Conference on Information Science and Applications, Seoul, pp. 1–7 (2010)
18. Wutsqa, D.U., Kusumawati, R., Subekti, R.: The application of elman recurrent neural network model for forecasting consumer price index of education, recreation and sports in yogyakarta. In: 10th International Conference on Natural Computation (2014)

Multi-objective Particle Swarm Optimisation for Robust Dynamic Scheduling in a Permutation Flow Shop

Mohanad Al-Behadili[1,2(✉)], Djamila Ouelhadj[2], and Dylan Jones[2]

[1] University of Basra, Basra, Iraq
mohanad.saad@uobasrah.edu.iq
[2] Department of Mathematics, University of Portsmouth, Lion Gate Building, Portsmouth PO1 3HF, UK
{djamila.ouelhadj,dylan.jones}@port.ac.uk

Abstract. This paper proposes a multi-objective optimisation model and particle swarm optimisation solution method for the robust dynamic scheduling of permutation flow shop in the presence of uncertainties. The proposed optimisation model for robust scheduling considers utility, stability and robustness measures to generate robust schedules that minimise the effect of different real-time events on the planned schedule. The proposed solution method is based on a predictive-reactive approach that uses particle swarm optimisation to generate robust schedules in the presence of real-time events. The evaluation of both the optimisation model and solution method are conducted considering different types of disruptions including machine breakdown and new job arrival. The obtained results showed that the proposed model and solution method gives better results than a bi-objective model that considers only utility and stability measures [1] and the classical makespan model.

1 Introduction

Scheduling deals with the assignment of a set of jobs to a set of machines in a reasonable amount of time in order to optimise one or more objectives [2]. In the flow shop scheduling problem (FSP) all jobs have to follow the same route. FSP that do not allow sequence changes between machines is called permutation flow shop scheduling problem (PFSP). In the PFSP, the minimisation of the completion time or makespan is the most common criterion that has been studied in the literature. It has been proven that, the PFSP is an NP-hard problem [3]. The most traditional classifications of the FSP are static, stochastic and dynamic scheduling environments [4]. In dynamic scheduling environments, random disruptions may interrupt the system, which could change the scheduled production plans. Noticeable studies of dynamic scheduling are given by [5–7], while a comprehensive survey is given by [8]. Most rescheduling strategies in the literature to handle the real-time events consider the following two goals [9]; make the schedule feasible again, and improve the efficiency of the schedule in

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_49

order to minimise the deviation of the current solution from the pre-planned one. To minimise the deviation between the performance measure values of the baseline and realised schedules, [10–12] propose a bi-objective function and consider the robustness measure as a linear combination of utility and stability measures. Rahmani and Heydari [13] propose a multi-objective optimisation model that considers utility, stability and robustness for FSP with unexpected new job arrival and uncertain processing times. The model was designed for the case of n-jobs and 2-machines. An exact method was applied to obtain the solution for small instances. Most of approaches designed for rescheduling are mainly focused on machine breakdowns, new job arrivals, etc. However, the majority of the existing work addresses these disruptions only independently [14–19]. Katragjini et al. [1] introduce a novel benchmark for PFSP that considers different types of disruptions during the time horizon. The authors implement a predictive-reactive approach for PFSP with three different random disruptions that interrupt the initial schedule simultaneously. These disruptions are machine breakdown, new job arrival and job ready time variations. To keep the solution efficient and stable, Katragjini et al. [1] proposed a bi-objective optimisation model, which takes into account the makespan and instability measures. Particle swarm optimisation (PSO) is a stochastic optimisation technique that was introduced by Kennedy et al. [20]. The PSO algorithm has many advantages making it more preferable to be proposed for dynamic combinatorial optimisation problems. Some advantages are the dynamic nature of the algorithm and its simplicity of implementation [21–23]. Other features are the use of self-information, individual best information and global best information to generate effective and optimal results, as well as the fast convergence speed of the swarm [24]. The main advantage of the PSO algorithm is that, it requires fewer parameters to be adjusted in comparison to other meta-heuristic methods [25]. The PSO algorithm was applied successfully to PFSP for example see in [26,27]. The contribution of this paper is to introduce a multi-objective optimisation model that considers utility, stability and robustness measures for robust dynamic PFSP under different types of real-time events and to apply an efficient predictive-reactive approach with PSO algorithm. The remainder of the paper is structured as follows; in Sect. 2, the optimisation model for robust scheduling is presented. In Sect. 3, the proposed predictive-reactive based PSO algorithm is developed. Section 4 shows numerical experiments and results that illustrate the performance of the proposed methodology, with a discussion of its results. Finally, conclusions and future work are presented in Sect. 5.

2 Multi-objective Optimisation Model

We propose a new multi-objective optimisation model based on the robust scheduling model introduced by [11,13]. This model has been extended for the case of n jobs and m machines for PFSP, considering three important measures: makespan measure (utility), stability and robustness. The new model for PFSP is given in (1).

$$Min\ MSR = \alpha U_n(S^*) + \beta I_n(S^*) + \gamma R_n(S^*) \tag{1}$$

where S^* refers to the new schedule after the time of disruption t_D and $\alpha + \beta + \gamma = 1$.

$U_n(S^*) = \sum_{\acute{j}} CR_{m\acute{j}}$ the real makespan in real scheduling.

$I_n(S^*) = \sum_i \sum_{\acute{j}} |CR_{i\acute{j}} - CP_{i\acute{j}}|$ the stability measure where $CR_{i\acute{j}}$ is the real completion time and $CP_{i\acute{j}}$ is the predicted completion time of job $\acute{j}$ on machine i according to the initial schedule.

$R_n(S^*) = |\sum_{\acute{j}} CR_{m\acute{j}} - \sum_{\acute{j}} CP_{m\acute{j}}|$ the robustness measure where $\sum_{\acute{j}} CP_{m\acute{j}}$ is the predictive makespan according to the initial schedule. n the number of jobs, m the number of machines, i index for machines $\{1, 2, ..., n\}$, $\acute{j}$ index of jobs that have not been executed on any machine yet and the newly arrived job at the time of disruption. We define the partial fixed sequence including the jobs that have already been processed or are in progress on the first machine before the time of disruption by π_{BD}. Similarly, the permutable subsequence containing the jobs after the time of disruption whose succession order can be modified is denote by π_{AD}. The three objective functions of model (1) are normalised to enable a reasonable comparison with the bi-objective model of [1] and the classical single objective makespan model. The normalised objective function (NMSR) in Eq. (1) is as follows:

$$NMSR = \alpha NU_n(S^*) + \beta NI_n(S^*) + \gamma NR_n(S^*) \tag{2}$$

where $NU_n(S^*)$, $NI_n(S^*)$ and $NR_n(S^*)$ represent the normalised makespan, stability and robustness respectively. The calculations of normalised makespan $NU_n(S^*)$ and normalised stability $NI_n(S^*)$ are given in details in [1], while the normalised robustness function is defined as follows:

$$NR_n(S^*) = \frac{R_n(S^*) - Min(R_n)}{Max(R_n) - Min(R_n)} \tag{3}$$

where $R_n(S^*)$ is the robustness after the time of disruption t_D. $Max(R_n)$ and $Min(R_n)$ are the upper and lower robustness bound, respectively. To obtain these bounds, we solve the problem three times where $\alpha = 1, \beta = 0, \gamma = 0$ for the first solution, $\alpha = 0, \beta = 1, \gamma = 0$ for the second solution and $\alpha = 0, \beta = 0, \gamma = 1$ for the last solution. The results are then given in the following 3×3 matrix form:

$$\begin{array}{c} \\ U_n^* \\ I_n^* \\ R_n^* \end{array} \begin{array}{c} \begin{array}{ccc} U_n & I_n & R_n \end{array} \\ \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix} \end{array} \tag{4}$$

The values of $Max(R_n)$ and $Min(R_n)$ are then calculated as follows: $Max(R_n) = max\{a_{1,3}, a_{2,3}, a_{3,3}\}$ and $Min(R_n) = min\{a_{1,3}, a_{2,3}, a_{3,3}\}$.

3 Proposed Solution Methods

3.1 A Robust Predictive-Reactive Rescheduling Approach

In this paper we consider two different real-time events that may occur at the same time or separately during the time horizon. These events are machine breakdown and new job arrival, also the number of disruptions could vary for each instance. A predictive reactive approach with an evolutionary PSO algorithm is applied to deal with large size PFSP in the presence of real-time. The predictive reactive approach starts with a robust predictive solution and then it applies rescheduling at the time of disruption. The rescheduling phase uses the PSO algorithm for the partial subsequence of jobs π_{AD} (the sequence of jobs that have not been executed on the first machine yet after the time of disruption) with the goal to minimise the partial makespan.

3.2 Particle Swarm Optimisation (PSO)

PSO is a population-based evolutionary computational method [20]. The algorithm simulates a social behaviour such as swarm bird migration. The basic elements of the PSO algorithm are summarised below where $j = \{1, 2, ..., n\}$, $i = \{1, 2, ..., \rho\}$, ρ is the number of particles and n is the number of jobs:

Particle: The i^{th} particle X_i^t in the swarm at iteration t is defined as $X_i^t = \{x_{i1}^t, x_{i2}^t, ..., x_{in}^t\}$ where x_{ij}^t is the position value of particle i of job j at iteration t.
Particle velocity: The i^{th} particle velocity V_i^t at iteration t is defined as $V_i^t = \{v_{i1}^t, v_{i2}^t, ..., v_{in}^t\}$ where v_{ij}^t is the velocity of particle i of job j at iteration t.
Population: The set consisting of ρ particles in the swarm at iteration t is called a population po^t where $po^t = \{X_1^t, X_2^t, ..., X_\rho^t\}$.
Permutation: The permutation of job sequence implied by the particle X_i^t is defined as $\pi_i^t = \{\pi_{i1}^t, \pi_{i2}^t, ..., \pi_{in}^t\}$ where π_i^t is the assignment of job j of the particle i in the permutation at iteration t.
Inertia Weight: The inertia weight w^t is used to control the impact of the velocities from the previous step on the current velocity.
Personal Best: The personal best P_i^t represents the best position of particle i with the best fitness at iteration t where $P_i^t = \{p_{i1}^t, p_{i2}^t, ..., p_{in}^t\}$ and p_{ij}^t is the position value of the i^{th} personal best with respect to j. In a minimisation problem with the objective function $f(\pi_i^t)$, the personal best P_i^t of the i^{th} particle in the swarm can be obtained such that $f(\pi_i^t) \leq f(\pi_i^{t-1})$ where π_i^t is the corresponding permutation of P_i^t and π_i^{t-1} is the corresponding permutation of P_i^{t-1}. The fitness function of P_i^t is simplified as f_i^{pb} instead of $f(\pi_i^t)$.
Global Best: The global best among all the swarm of particles achieved so far is called global best G^t. It is defined as $G^t = \{g_1^t, g_2^t, ..., g_n^t\}$ where g_i^t is the position value of G^t. To obtain the global best we use the criterion $f(\pi^t) \leq f(\pi_i^t)$ where π^t and π_i^t are the corresponding permutation of global

best G^t and personal best P_i^t respectively. For simplicity we use f^{gb} for the fitness function of the global best instead of $f(\pi^t)$.
Termination Criterion: The search process of PSO algorithm is terminated after a maximum number of iterations.

PSO particles are used to solve the PFSP as follows; for n jobs in the problem each particle $X_i^t = \{x_{i1}^t, x_{i2}^t, ..., x_{in}^t\}$ corresponds to a continuous position value. The vector of position values is transformed to the permutation of jobs according to the SPV rule [28]. The initial population of particles generated randomly and the initial position values for the particle are generated randomly as follows: $x_{ij}^0 = x_{min} + r_1(x_{max} - x_{min})$ where $[x_{min}, x_{max}] = [0, 4]$ and $r_1 \in (0, 1)$ is a uniform random number. Initial velocities are established similarly as follows: $v_{ij}^0 = v_{min} + r_2(v_{max} - v_{min})$ where $[v_{min}, v_{max}] = [-4, 4]$ and $r_2 \in (0, 1)$ is a uniform random number. The fitness function is given as $f(\pi_i^t) = NMSR$ and it is rewritten as f_i^t in short. The steps of the PSO algorithm are summarised as below:

Algorithm 1. Particle Swarm Optimisation Algorithm

1. Set initial iteration $t = 0$ and $\rho = 2 \times n$.
2. Generate ρ initial particles $X_i^0 = \{x_{i1}^0, x_{i2}^0, ..., x_{in}^0\}$ where x_{ij}^0 is selected randomly from the range $[0, 4]$.
3. Generate ρ initial velocities $V_i^0 = \{v_{i1}^0, v_{i2}^0, ..., v_{in}^0\}$ where v_{ij}^0 is chosen randomly from the range $[-4, 4]$.
4. Use the SPV rule to detect $\pi_i^0 = \{\pi_{i1}^0, \pi_{i2}^0, ..., \pi_{in}^0\}$ of particle X_i^0.
5. Apply the fitness function f_i^0 to evaluate each particle i in the swarm.
6. For each particle set $P_i^0 = X_i^0$ where $P_i^0 = \{p_{i1}^0, p_{i2}^0, ..., p_{in}^0\}$, $p_{i1}^0 = x_{i1}^0, p_{i2}^0 = x_{i2}^0, ..., p_{in}^0 = x_{in}^0$ together with its best fitness function $f_i^{pb} = f_i^0$.
7. Detect the best fitness value in the whole swarm such that $f_L^0 = min\{f_i^0\}$ with its corresponding particle X_L^0. Set global best to $G^0 = X_L^0$ such that $g_1^0 = x_{L1}^0, g_2^0 = x_{L2}^0, ..., g_n^0 = x_{Ln}^0$ with its fitness value $f^{gb} = f_L^0$.
8. Update the iteration $t = t + 1$.
9. Update the inertia weight $w^t = w^{t-1} \times \beta_{PSO}$ where β_{PSO} is the decrement factor.
10. Update the velocity as follows $v_{ij}^t = w^{t-1}v_{ij}^{t-1} + c_1r_1(p_{ij}^{t-1} - x_{ij}^{t-1}) + c_2r_2(g_j^{t-1} - x_{ij}^{t-1})$ where c_1 and c_2 are social and cognitive parameters and $r_1, r_2 \in (0, 1)$ are uniform random numbers.
11. Update position values $x_{ij}^t = x_{ij}^{t-1} + v_{ij}^t$.
12. Use the SPV rule to detect the permutation $\pi_i^t = \{\pi_{i1}^t, \pi_{i2}^t, ..., \pi_{in}^t\}$.
13. Use the permutation to evaluate particles by checking if there is any improvement at iteration t for the personal best. That is, if $f_i^t < f_i^{pb}$, then P_i^t is updated as $P_i^t = X_i^t$ and $f_i^{pb} = f_i^t$.
14. Find the minimum value of P_i^t as $f_L^t = min\{f_i^{pb}\}$, $L \in \{i | i = 1, 2, ..., \rho\}$. If $f_L^t < f^{gb}$, then update the global best as $G^t = X_L^t$ and $f^{gb} = f_L^t$.
15. If the number of iteration exceeds the maximum number of iterations, then stop; otherwise go back to step.

4 Computational Results

We have conducted extensive experiments to evaluate the performance of the proposed optimisation model and solution methods. The proposed methodology has been implemented in Java on Windows7 64 bit. An Intel Core i5, 2.30 GHz processor and 6 GB of RAM was used. In PSO algorithm the population size is set to $2 \times n$, the initial inertia weight is set up to $w^0 = 0.9$ and no less than 0.4, for w the decrement factor β_{PSO} is taken as 0.975; the acceleration coefficients are set to $c_1 = c_2 = 2$ [28]. The computational experiments were carried out on the Taillard benchmark instances [29], which are grouped in 12 sets of 10 instances, ranging from 20 jobs and 5 machines to 500 jobs and 20 machines. All instances have been run independently ten times, and the average is computed. To evaluate the performance of the proposed model and the solution method, thirteen different set of weights (α, β, γ) have been chosen, these sets represent the relative importance of each objective in model (1). The selection of these weights are based on the practical weight sensitivity algorithm detailed in [30]. Three weights are used to normalise the model in Eq. (1) as described in Sect. 2. These weights are as follows; $(0.999, 0.001, 0.001)$, $(0.001, 0.999, 0.001)$, $(0.001, 0.001, 0.999)$. The remaining 10 set of weights are examined for this experiment. These sets are arranged as follows: $W_1 = (0.333, 0.0.333, 0.333), W_2 = (0.666, 0.166, 166)$, $W_3 = (0.499, 0.499, 0.002), W_4 = (0.416, 0.416, 0.166), W_5 = (0.166, 0.666, 0.166)$, $W_6 = (0.002, 0.499, 0.499)$, $W_7 = (0.166, 0.416, 0.166)$, $W_8 = (0.166, 0.166, 0.666)$, $W_9 = (0.499, 0.002, 0.499)$, $W_{10} = (0.416, 0.166, 0.416)$. On the other hand, the bi-objective model of [1] is examined for the first components of W_1 to W_{10}. These components are as follows:

$$\alpha = 0.333, 0.666, 0.499, 0.416, 0.166, 0.002, 0.166, 0.166, 0.499, 0.416.$$

Once the best value is found, we calculate the average relative percentage deviation (RPD) over 10 of Taillard instances with the same number of jobs and machines as follows:

$$RPD = \frac{M - Best_{Sol}}{Best_{Sol}} \times 100 \tag{5}$$

where M is the solution obtained by the proposed model and PSO algorithm. $Best_{Sol}$ is the average of 10 Taillard instances of lower bound solution from the same size. Table 1 shows the RPD for some Taillard instances according to the weights where **MSR** represents the solution obtained from our proposed model, while **bi-obj** is the solution obtained by bi-objective model [1].

In this table **Ta** represents the instance of n jobs and m machines. According to Table 1, the objective functions are sensitive to different sets of weights. The results show that the stability and robustness measures play an important role in obtaining better solution. As shown on Table 1, the weights W_6, W_7 and W_8 produce better solution in comparison with other weights. Furthermore, the solution corresponding to the weight W_8 has the lowest RPD values, this shows that giving more priority to the robustness measure produces better solutions.

Table 1. RPD for MSR and bi-obj models.

	Ta	20 × 5	50 × 10	100 × 10	200 × 20	500 × 20	Total
W_1	MSR	29.446	22.872	23.259	22.559	17.737	115.873
	bi-obj	34.345	22.311	30.075	22.860	20.764	130.355
W_2	MSR	29.381	22.548	26.133	22.734	17.765	118.561
	bi-obj	31.384	22.241	30.012	21.896	23.592	129.125
W_3	MSR	29.062	23.539	26.151	22.518	17.960	119.230
	bi-obj	31.608	23.612	21.039	20.427	22.355	130.217
W_4	MSR	29.512	22.775	23.805	22.973	17.690	116.755
	bi-obj	31.945	24.781	32.556	21.367	20.482	131.131
W_5	MSR	30.796	23.369	24.082	22.600	17.440	118.287
	bi-obj	32.152	23.562	31.964	24.103	22.382	134.163
W_6	MSR	26.871	21.771	21.687	19.088	12.117	101.534
	bi-obj	31.278	25.360	34.158	23.554	20.359	134.709
W_7	MSR	28.163	22.358	20.920	19.082	13.794	104.317
	bi-obj	32.174	22.982	33.242	24.001	23.923	136.322
W_8	MSR	26.609	21.461	20.831	18.843	12.277	100.021
	bi-obj	31.769	23.375	32.380	23.909	24.521	135.954
W_9	MSR	29.422	22.184	23.648	22.889	17.177	115.320
	bi-obj	32.807	22.308	32.409	20.427	22.800	130.751
W_{10}	MSR	29.798	22.451	24.007	20.303	17.031	113.590
	bi-obj	32.611	23.332	33.679	22.830	22.763	135.215

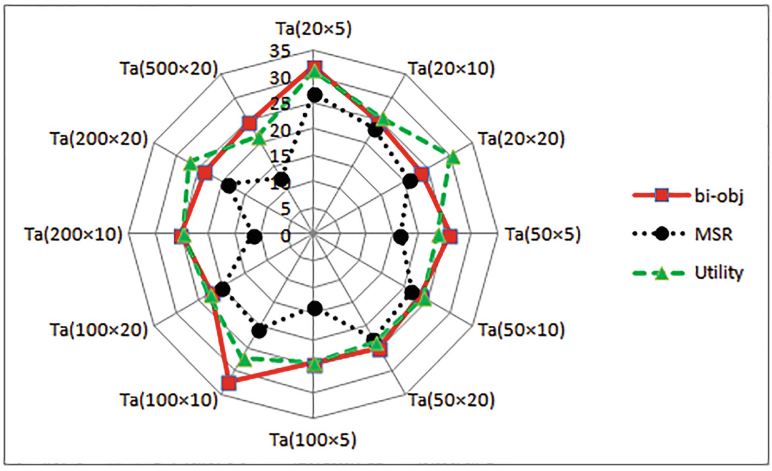


Fig. 1. RPD for models with the weight W_8.

Table 2. Analysis of variance.

Groups	Count	Sum	Average	Variance
MSR	12	242.196	20.183	22.17
bi-obj	12	308.617	25.718	17.768
Utility	12	307.253	25.604	9.702

ANOVA						
Source of variation	SS	df	MS	F	P-value	F crit
Between groups	240.165	2	120.082	7.257	0.002	3.284
Within groups	546.053	33	16.547			
Total	786.218	35				

For this reason, the weight W_8 is selected in this experiment. Figure 1, shows that MSR has the best RPD, compared to bi-obj and Utility. It is clear that the solution obtained by MSR has the lowest RPD in general compared with the bi-objective and utility models.

To further investigate the impact of the proposed models on the dependent variable RPD, the single factor mean of an Analysis of Variance (ANOVA) is performed. The results of the MSR model corresponding to the weight W_8, bi-obj model with $\alpha = 0.166$ and Utility model are statistically tested by ANOVA. For this analysis, the null and alternative hypotheses in ANOVA are:

H_0 : all means are same.
H_A: at least one mean is different.

The F-ratios and the p-values are reported in Table 2. Since the p-value is smaller than 0.05, we reject the null hypothesis of no difference, and conclude these fac-

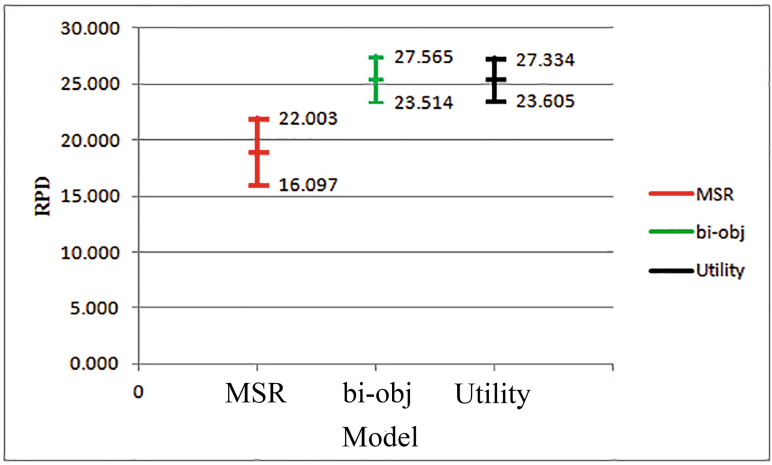


Fig. 2. 95% Tukey confidence intervals for all models.

tors have a statistically significant effect on RPD at the 95% confidence level. The ANOVA F-test only permits us to reject the null hypothesis, but it does not give indication about which groups have different mean values. Hence, the 95% confidence interval is used to specify which of the mean RPD differences are statistically significant. Figure 2 shows that MSR model has significance difference in comparison with the other two models.

5 Conclusion

This paper proposes a model that considers utility, stability and robustness to obtain robust and stable schedules of PFSP subject to different real-time events. Although we proposed a predictive-reactive approach based on generating robust schedule and reacting to every disturbance. Furthermore, PSO is applied for rescheduling, which shows the ability to deal with dynamic nature of this problem successfully. On the other hand, the proposed robust model has been compared with the bi-objective and the classical makespan model. Statistical ANOVA analysis have been conducted to compare the efficiency of these models. The ANOVA results revealed that the MSR model performed best. Also the sensitivity analysis indicated that giving higher priority to robustness leads to better solutions. For future work the current proposed model can be explored to handle other types of real-time events. Another interesting direction is to adapt the proposed model to other scheduling problems such as flexible flow shop and job shop scheduling.

References

1. Katragjini, K., Vallada, E., Ruiz, R.: Flow shop rescheduling under different types of disruption. Int. J. Prod. Res. **51**(3), 780–797 (2013)
2. Pinedo, M.: Scheduling: Theory, Algorithms, and Systems, 4th edn. Springer, New York (2012)
3. Graham, R.L., Lawer, E.L., Lenstra, J.K., Rinnooy Kan, A.H.: Optimisation and approximation in deterministic sequencing and scheduling: a survey. Ann. Discret Math. **5**, 287–326 (1979)
4. War, D.: The classification of scheduling problems under production uncertainty. Res. Logist. Prod. **4**(3), 245–255 (2014)
5. Sabuncuoglu, I., Goren, S.: Hedging production schedules against uncertainty in manufacturing environment with a review of robustness and stability research. Int. J. Comput. Integr. Manuf. **22**(2), 37–41 (2009)
6. Suwa, H., Sandoh, H.: Online Scheduling in Manufacturing: A Cumulative Delay Approach. Springer, London (2013)
7. Aytug, H., Lawley, M., McKay, K.: Executing production schedules in the face of uncertainties: a review and some future directions. Eur. J. **161**, 86–110 (2005)
8. Ouelhadj, D., Petrovic, S.: A survey of dynamic scheduling in manufacturing systems. J. Sched. **12**(4), 417–431 (2008)
9. Vieira, G.E., Herrmann, J.W., Lin, E.: Rescheduling manufacturing systems: a framework of strategies, policies, and methods. J. Sched. **6**(1), 39–62 (2003)

10. Cowling, P., Johansson, M.: Using real time information for effective dynamic scheduling. Eur. J. Oper. Res. **139**(2), 230–244 (2002)
11. Cowling, P., Ouelhadj, D., Petrovic, S.: A multi-agent architecture for dynamic scheduling of steel hot rolling. J. Intell. Manuf. **14**, 57–470 (2003)
12. Cowling, P.I., Ouelhadj, D., Petrovic, S.: Dynamic scheduling of steel casting and milling using multi-agents. Prod. Plan. Control **15**(2), 178–188 (2004)
13. Rahmani, D., Heydari, M.: Robust and stable flow shop scheduling with unexpected arrivals of new jobs and uncertain processing times. J. Manuf. Syst. **33**(1), 84–92 (2014)
14. Church, L.K., Uzsoy, R.: Analysis of periodic and event-driven rescheduling policies in dynamic shops. Int. J. Comput. Integr. Manuf. **5**(3), 153–163 (1992)
15. O'Donovan, R., Uzsoy, R., McKay, K.N.: Predictable scheduling of a single machine with breakdowns and sensitive jobs. Int. J. Prod. Res. **37**(18), 4217–4233 (1999)
16. Vieira, G.E., Herrmann, J.W., Lin, E.: Predicting the performance of rescheduling strategies for parallel machine systems. J. Manuf. Syst. **19**(4), 256–266 (2000)
17. Hall, N.G., Potts, C.N.: Rescheduling for new orders. Oper. Res. **3**(52), 440–453 (2004)
18. Rangsaritratsamee, R., Ferrell, W.G., Kurz, M.B.: Dynamic rescheduling that simultaneously considers efficiency and stability. Comput. Ind. Eng. **46**(1), 1–15 (2004)
19. Kopanos, G.M., Capo, E., Espun, A., Puigjaner, L.: Costs for rescheduling actions: a critical issue for reducing the gap between scheduling theory and practice. Ind. Eng. Chem. Res. **47**(22), 8785–8795 (2008)
20. Kennedy, J., Eberhart, R.: Particle swarm optimisation. In: Proceedings of International Conference on Neural Networks. ICNN 95, vol. 4, pp. 1942–1948 (1995)
21. Blackwell, T.: Particle swarm optimisation in dynamic environments. Evol. Comput. Dyn. Uncertain Environ. **51**, 29–49 (2007)
22. Li, X., Branke, J., Blackwell, T.: Particle swarm with speciation and adaptation in a dynamic environment. In: Proceedings of the 8th GECCO, p. 51 (2006)
23. Zhang, Y., Wang, S., Ji, G.: A comprehensive survey on particle swarm optimisation algorithm and its applications, vol. 2015, p. 38 (2015)
24. Blum, C., Merkle, D.: Swarm Intelligence: Introduction and Applications. Springer, Heidelberg (2008)
25. Eslami, M., Shareef, H., Khajehzadeh, M., Mohamed, A.: An effective particle swarm optimisation for global optimisation. Comput. Intell. Syst. **316**, 267–274 (2012)
26. Lian, Z., Gu, X., Jiao, B.: A novel particle swarm optimisation algorithm for permutation flow-shop scheduling to minimise makespan. Chaos Solitons Fractals **35**(4), 851–861 (2008)
27. Ramanan, T., Iqbal, M., Umarali, K.: A particle swarm optimisation approach for permutation flow shop scheduling problem. Int. J. Simul. Multidiscip. Des. Optim. **5**, A20 (2014)
28. Tasgetiren, M.F., Sevkli, M., Liang, Y.-C., Gencyilmaz, G.: Particle swarm optimization algorithm for permutation flowshop sequencing problem. In: Dorigo, M., Birattari, M., Blum, C., Gambardella, L.M., Mondada, F., Stützle, T. (eds.) ANTS 2004. LNCS, vol. 3172, pp. 382–389. Springer, Heidelberg (2004). doi:10.1007/978-3-540-28646-2_38
29. Taillard, E.: Benchmarks for basic scheduling problems. Eur. J. Oper. Res. **64**, 278–285 (1993)
30. Jones, D.: A practical weight sensitivity algorithm for goal and multiple objective programming. Eur. J. Oper. Res. **213**(1), 238–245 (2011)

Mining Perfectly Rare Itemsets on Big Data: An Approach Based on Apriori-Inverse and MapReduce

F. Padillo, J.M. Luna, and S. Ventura(✉)

Department of Computer Science and Numerical Analysis,
University of Cordoba, Rabanales Campus, 14071 Cordoba, Spain
{fpadillo,jmluna,sventura}@uco.es

Abstract. Association rule mining is one of the most common data mining techniques used to identify and describe interesting relationships between patterns from large quantities of data. Whereas many researches have been focused on the extraction of these patterns which appear frequently to obtain general information, in some scenarios it could also be interesting to extract unexpected phenomena. Rare association rule mining is a recent field aiming to discover sporadic rules having a low frequency of appearance but high confidence of occurring together. This field is really useful over Big Data where abnormal endeavor are more interesting than normal behavior. In this sense, our aim is to propose a new algorithm to obtain rare association rule on Big Data using MapReduce by means of Spark and Hadoop. The experimental study includes more than 30 datasets revealing alluring results in efficiency when more than 60,000 million of instances and file sizes of 500 GBytes are considered.

Keywords: Spark · Hadoop · Big Data · Association rules · Rare itemsets

1 Introduction

The interest in data gathering has been growing exponentially for the last decade, and the estimations confirm that it will continue growing faster than ever. To extract nontrivial information from this raw data many different techniques have been proposed, all of them cataloged under the term Knowledge Discovery in Databases (KDD).

Pattern Mining (PM) is one of the most prestigious techniques from KDD, where the useful information is represented by means of patterns. This representation enables to define subsequences, substructures or itemsets that represent any type of homogeneity and regularity in data [1]. On this way, general tendency is described since patterns are witnesses of recurrent phenomena.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_50

However, in many situations it may be interesting to extract unexpected phenomena, that is, patterns that do not occur frequently or rare. These corresponding to unexpected situations, possibly contradicting beliefs in the domain [2,3]. Rare itemsets are related to exceptions and thus may convey information of high interest for the expert in the domain. For instance, a rare but fatal disease could be missed out due to the frequency of incidences not reaching a threshold of occurrence. The extraction of rare phenomena is really useful on Big Data, where the normal behavior is known and the aim is to extract abnormal movements.

At this point, our aim is to propose a new algorithm for mining rare itemsets on Big Data based on traditional algorithms and using emerging paradigms of distributed computing. In this regard, MapReduce [4] has been considered since it has proved to obtain excellent results working on Big Data. In concrete, Hadoop and Spark have been selected among others because they are considered as the most prestigious ones. The goal is to propose an adaptation of one distinguished algorithm for obtaining rare itemsets but over MapReduce paradigms without losing accuracy in the solutions. In this regard, Apriori-InverseMR (Apriori-Inverse using MapReduce) is proposed as an iterative algorithm designed to be used in Hadoop and Spark based on the well-known Apriori-Inverse algorithm [2]. A detailed experimental study was carried out to prove the efficiency and scalability of our proposal. The results reveal that our proposals outperform with respect to traditional algorithms using 30 real world datasets and they are run on 60,000 million of instances efficiently.

This paper is structured as follows. Section 2 presents the most relevant definitions. Section 3 describes the proposed algorithm. Section 4 describes the experimental analysis. Finally, some concluding remarks are stated in Sect. 5.

2 Background

In this section, the pattern mining task is formally defined. Then, MapReduce paradigm is described.

2.1 Pattern Mining

Pattern Mining (PM) aims to extract hidden knowledge from raw data by means of a versatile representation (Pattern). A pattern is defined as a set of items that represents any type of homogeneity and regularity in data, denoting intrinsic and important properties of data [1]. Formally, let $\mathcal{I} = \{i_1, i_2, ..., i_n\}$ be a set of items, a pattern P is formally defined as $P = \{i_j, ..., i_x\} \subseteq \mathcal{I}, j \geq 1, x \leq n$. Thus, for $P = \{i_1, i_2, ..., i_s\} \subseteq \mathcal{I}$, its size is defined as $|P| = s$. Additionally, given a set of all transactions $\mathcal{T} = \{t_1, t_2, ..., t_m\}$ in a dataset, the support of a pattern P is defined as the number of transactions that P satisfies, i.e., $support(P) = |\{\forall t_l \in \mathcal{T} : P \subseteq t_l\}|$. A pattern P is considered as frequent iff $support(P) \geq threshold$. Furthermore, in some cases the information represented by a pattern could not be enough and a more descriptive expression is required. In this regard,

Agrawal et al. [5] proposed to use the concept of association rule as a way of describing correlations within patterns of potential interest. They are formally defined as an implication of the form $X \rightarrow Y$ where $X \subset \mathcal{I}$, $Y \subset \mathcal{I}$, and $X \cap Y = \emptyset$. In order to measure how the strong is its implication, the measure confidence is used, which is defined as $confidence(X \rightarrow Y) = \frac{support(X \rightarrow Y)}{support(X)}$. In order to extract these association rules, many different approaches have been proposed, although almost all of them are based on the Apriori [5] algorithm.

Not only the extraction of frequent patterns is interesting, but also the discovery of abnormal or unusual behavior. The extraction of reliable and rare relations [1] is considered as a subset of PM and also an area of interest for further exploration. In this regard, perfectly rare itemsets [2] are defined as these rules whose $confidence(X \rightarrow Y) \geq threshold_{confidence}$ and $\forall x : x \in (X \cup Y), support(x) < threshold_{support}$. Other definitions of rare patterns have been proposed [6], however they are not as sporadic as perfectly rare itemsets. Proposed approaches on the rare itemset field are based on slight adaptations of classic algorithms where the threshold is set to a very low value, allowing to obtain exceptional behavior but imposing a search space bigger than needed. That is known as *rare item problem* [2], meaning that using classic algorithms are unlikely to generate rare association rules. For instance, in the security field, normal behavior is very frequent, whereas abnormal or suspicious behavior is less frequent but of high interest. Aiming to solve this issue, Apriori-Inverse [2] is proposed as an adaptation of Apriori to obtain perfectly rare itemsets, enabling to obtain rules with very low support but high confidence of occurring together with a lower computational cost. Despite Apriori-Inverse is an efficient algorithm, its performance is hampered when large amount of data are considered. In order to speed up the process some adaptations have to done to be used over new paradigms of distributed computing [7,8].

2.2 MapReduce

MapReduce [4] is a recent paradigm of distributed computing. In this paradigm, applications are composed of two main phases: map and reduce. In the map phase, each mapper processes a sub-set of input data and produces $\langle k, v \rangle$ pairs. Then, an intermediate step is carried out, known as shuffle phase, that merges all the values associated with the same key k. Finally, the reducer takes this new list as input to produce the final values. Many implementations have been developed following this paradigm, but some of them have stood out [9].

Apache Hadoop is one of the first pioneering open source implementations of MapReduce. The main drawback of Hadoop is that it imposes an acyclic data flow graph, and there are applications that cannot be modeled efficiently using this kind of topology [9]. Moreover, the communication among mappers and reducers are performed using disk, in this sense the output of mappers are written on disk, to be read from the reducers. This operation could cause problems of I/O, when the number of $\langle k, v \rangle$ are extremely large. To solve these downsides, a novel solution was proposed known as Spark [9]. This new proposal is eminently designed to be used in iterative and interactive algorithms. To speed

up the process of tackling huge amounts of data, it introduces an abstraction called Resilient Distributed Datasets (RDDs). It represents a read-only collection of objects partitioned across a set of machines stored in main memory.

3 Apriori-Inverse Using MapReduce on Big-Data

Our proposal is based on a level-wise search strategy whose goal is to obtain perfectly rare itemsets as well as rules containing these itemsets on Big Data, known as Apriori-Inverse-MR (Apriori-Inverse using MapReduce). Our approach follows two different steps: (1) to obtain patterns whose support are lower than a minimum support threshold but higher than a maximum support threshold; (2) to mine association rules using the previously patterns whose confidence are higher than a threshold.

Step I: Obtaining perfectly rare itemsets

In this stage perfectly rare itemsets are obtained. In order to do so, different iterations are required where in each one a different size of itemset is mined. Each iteration implies a MapReduce operation, working as follows.

Step I.1. Map phase. Each mapper receives two kind of inputs: (1) a chunk of the database; (2) the perfectly rare itemsets obtained in the previous iteration. Being its goal to produce the different itemset of size s joining $s-1$ itemsets which shares its last $s-2$ items and evaluating these in its chunk. Producing a set of $\langle k, v\rangle$ pairs where k is the pattern and v its support. Figure 1 shows an example where three different iterations have been carried out. In the last iteration ($s=3$), the itemsets bc and bf are joined since they share $s-2$ items, producing bcf.

Step I.2. Reduce phase. Different reducers are considered, where the aim of each one is to analyze a set of $\langle k, v\rangle$ pairs whose k are the same, adding their v to obtain the support of each pattern in the whole dataset. After that, the $\langle k, v\rangle$ pair is produced as output if it satisfies the minimum and maximum support threshold values. For instance, in Fig. 1 could be appreciated that in the first iteration the itemset a is discarded since it does not satisfy the maximum support of 2.

Step II: Mining association rules using the perfectly rare itemsets

In this step the association rules are mined using the previously obtained perfectly rare itemsets. In order to create association rules a MapReduce phase is used, working as follows.

Step II.1. Map phase. All the previously obtained perfectly rare itemsets are used as input for this phase. Different mappers are considered in function of the number of rare itemsets. The aim of these mapper is two-fold: (1) Producing a

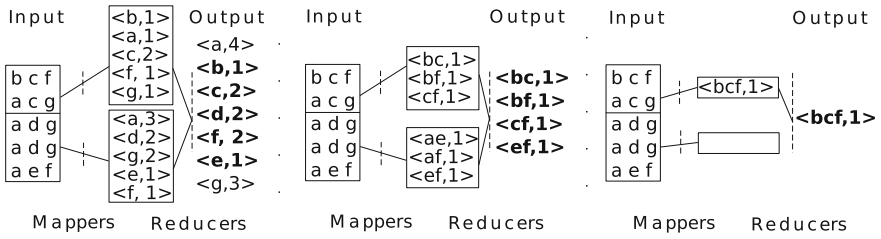


Fig. 1. Example of Step I where perfectly rare itemsets are produced. Results are a set of $\langle k, v \rangle$ pairs, where k is a rare itemset and v is its support. A threshold of minimum and maximum support of 1 and 2 respectively have been used, bold typeface denotes to the solutions returned.

set of $\langle k, v \rangle$ pairs where k is the perfectly rare itemset, and v is the support; (2) For each rare itemset to produce as many rules as possible, generating a set of $\langle k, v \rangle$ pairs where k is the antecedent of the association rule, and v is a tuple of two values with the form of $\langle consequent, supportRule \rangle$. This functionality is depicted in Fig. 2.

Step II.2. Reduce phase. For each different antecedent a unique reducer is considered, where the aim of each one is to analyze a set of $\langle k, v \rangle$ pairs whose k are the same. For the sake of better understanding, Fig. 2 shows an example. Six different reducers have been considered since six different antecedents exists. The first one only analyzes $\langle k, v \rangle$ pairs where $k = f$, the second one

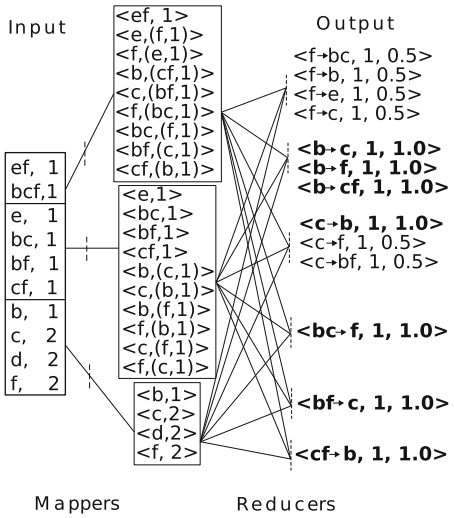


Fig. 2. Example of Step II where association rules are produced. Results are a set of $\langle k, v \rangle$ pairs, where k is an association rule and v is a tuple of two different values with the form of $(supportRule, confidence)$. A threshold of confidence of 0.8 has been considered, bold typeface denotes to the solutions returned.

only analyses $\langle k, v\rangle$ pairs where $k = b$, and so on. The first reducer produces the rule $f \rightarrow c$ using $\langle f, 2\rangle$ to know the support of the antecedent of the rule and $\langle f, (c, 1)\rangle$ to calculate the support of this rule. Producing $\langle f \rightarrow c, 1, 0.5\rangle$, where 1 is the support of the rule and 0.5 is the confidence of the rule, i.e., $confidence = \frac{supportRule}{supportAntecedent}$. As the confidence is lower than a threshold this rule is discarded.

4 Experimental Study

The aim of this section is to compare the scalability as well as the performance of our proposal with other well-known algorithms on the field. The purpose of this study is two-fold:

- To compare the performance of the standard Apriori algorithm with Apriori-Inverse to obtain perfectly rare itemsets.
- To analyze the computational time for a varied set of implementations of our proposal, comprising both sequential and MapReduce algorithms. This study also includes an evaluation of our proposal when truly Big Data is considered.

It is worth noting that the runtime obtained in this experimental analysis is the average runtime obtained for 10 different runs, even though all of them obtained the exact same results. It has been carried out in this way to alleviate the variations in the runtime derived from the resource management.

All the experiments were run on a HPC cluster comprising 12 compute nodes. Each of the nodes comprised two Intel Xeon E5645 CPUs with 6 cores at 2.4 GHz and 24 GB DDR memory. Cluster operating system was Rocks cluster 6.1×64. As for the specific details of the software used, the experiments have been run on Spark 2.0.

4.1 Apriori vs Apriori-Inverse to Obtain Perfectly Rare Itemsets

The aim of this section is to compare Apriori with Apriori-Inverse using 30 well-known real world datasets. All of them are freely available in KEEL repository. The number of instances ranges from 40 to $1,025,010$, whereas the search space varies from $2^3 - 1$ to $2^{14} - 1$ itemsets. As both algorithms only work with discrete values, those datasets having continuous attributes have been discretized using equal-width and equal-frequency with 3, 5 and 10 bins [10].

Table 1 shows the results obtained. For space limitations, only results using 5 bins have been included. The rest of results for different bins are available at http://www.uco.es/grupos/kdis/kdiswiki/Apriori-Inverse. A minimum and maximum support threshold of 0.001 and 0.5 have been considered, as well as a threshold of 0.7 for confidence. As it could be appreciated the number of rules and patterns highly varies among datasets, it is due to the different distribution of the data. In function of the distributions, the thresholds will be able to prune more search space than in other cases. Even more, in the same dataset using

Table 1. Results obtained after running Apriori and Apriori-Inverse using a minimum and maximum support threshold of 0.001 and 0.5 respectively, as well as a confidence threshold of 0.8.

Dataset	Discretization	Apriori			Apriori-Inverse		
		Time (Minutes)	Patterns	Rules	Time (Minutes)	Patterns	Rules
Bolts	Width 5	0.009	5859	52474	0.006	3190	25005
	Frequency 5	0.009	6845	66036	0.009	6845	66036
Wine	Width 5	0.03	326042	3266006	0.02	4397	3756
	Frequency 5	0.04	274857	1361958	3.8	274857	1361958
Wankara	Width 5	0.31	124832	2000431	0.1	50154	575362
	Frequency 5	0.69	202970	3915275	0.33	131748	2252549
Wizmir	Width 5	0.25	98496	819910	0.015	8944	20490
	Frequency 5	0.67	194534	960690	0.12	65360	164971
Abalone	Width 5	0.1	41513	282224	0.04	20906	88304
	Frequency 5	0.18	64974	263945	0.18	64974	263945
Concrete	Width 5	0.07	46040	641778	0.01	11541	89567
	Frequency 5	0.11	59978	699591	0.07	45643	501260
Delta_elv	Width 5	0.05	12781	6925	0.03	4372	561
	Frequency 5	0.09	20435	7631	0.07	12426	514
Breast	Width 5	0.002	92544	1102427	0.04	36260	339144
	Frequency 5	0.003	99673	1060238	0.04	38450	345414
Dee	Width 5	0.08	36593	411422	0.08	36593	411422
	Frequency 5	0.08	37149	413358	0.08	37149	413358
Friedman	Width 5	0.06	3505	288	0.05	3505	288
	Frequency 5	0.05	3522	295	0.05	3522	295
Glass	Width 5	0.08	58869	1278912	0.005	3676	40295
	Frequency 5	0.33	134419	2872953	0.07	53385	838673
Pima	Width 5	0.11	63395	441475	0.01	13177	50598
	Frequency 5	0.33	114429	891076	0.33	114429	891076
Bupa	Width 5	0.005	2755	6822	0.001	659	1375
	Frequency 5	0.008	8281	14103	0.008	8281	14103
Ecoli	Width 5	0.01	17961	104831	0.006	4730	11168
	Frequency 5	0.18	19797	113915	5201	12003	0.005
Stock Price	Width 5	0.005	150583	2355072	0.004	150583	2355072
	Frequency 5	0.01	212090	3084905	0.01	212090	3084905
Tic-Tac-Toe	-	1.16	250985	1220682	0.56	166555	733404
Flare	-	0.16	87649	1331844	0.005	1690	4531
Car	-	0.008	6271	0	0.008	6271	0
Mammography	Width 5	0.004	3841	7743	0.003	1395	1874
	Frequency 5	0.005	3782	6125	0.001	1300	1426
Marketing	Width 5	54.47	1334045	11078662	0.25	68375	115076
	Frequency 5	63.51	1433505	11024032	0.336	84311	153561
Nursery	-	0.54	88250	28738	0.6	88250	28738
Magic	Width 5	0.32	77514	454475	0.07	9543	9672
	Frequency 5	0.72	165732	245635	1.67	165732	232029
California	Width 5	0.05	3772	16700	0.01	422	366
	Frequency 5	0.001	17876	18761	0.04	6546	11142
ChessKrKp	-	0.14	9628	802	0.13	9628	802
Mv	Width 5	0.05	187428	230008	0.04	34886	14299
	Frequency 5	0.05	187428	199574	0.04	44200	23711
Adult	Width 5	147.98	2221020	127777894	0.30	19189	18436
	Frequency 5	150.1	2222378	68579457	1.02	70947	82513
Statlog (Shuttle)	Width 5	0.26	8751	250613	0.007	130	665
	Frequency 5	6.23	184900	936770	0.79	48076	132446
ColorMoments	Width 5	6.01	29703	71963	0.04	1612	1411
	Frequency 5	6.25	92964	82288	1.37	16970	122606
Transactions90k	-	0.5	427	0	0.5	427	0
Poker	Width 5	201.03	195742	116120	150	100874	0
	Frequency 5	270.07	218697	10403	161	147770	0

different discretization the number of patterns and rules could highly vary since the kind of discretization will determine the distribution of the data.

It is worth noting that Apriori-Inverse almost always obtains a low number of pattern and rules, only in the worst case scenario obtains the same number as Apriori (due to the distribution of data). This fact is cornerstone, since Apriori-Inverse enables to obtain only perfectly rare itemsets with a lower runtime than Apriori. It should be noted that Apriori-Inverse allows to obtain the exact same perfectly rare itemsets than Apriori using only a 0.54% of the runtime required by Apriori. Furthermore, in some cases where do no exist reliable unexpected phenomena (Car, Transactions90k and Poker), Apriori returns rules describing general behavior which is not the goal in this case. These situations where the number of rules obtained are 0 is due to although some patterns are rare, they do not occur together in enough examples, thus any rule could be obtained with enough confidence to determine a reliable relation.

Finally, it could also be appreciated that in large datasets the runtime required is more than 2 h (Poker dataset), meaning that it is impossible to run these sequential approaches in a reduced quantum of time. Therefore, some adaptations should be done to be run on Big Data efficiently.

4.2 Scalability

In this study the scalability of our proposal is analyzed. In this regard, a set of synthetic datasets have been used to analyze how the number of both instances and attributes affects the performance. These datasets comprises a search space from $2^8 - 1$ to $2^{40} - 1$ itemsets, and a number of instances from $2 \cdot 10^6$ to $6 \cdot 10^{10}$. Synthetic datasets have been used since the number of both instances and attributes can be easily changed to illustrate the scalability of the proposals. Only algorithms for obtaining perfectly rare itemsets have been considered in this study since it has been proved that obtain better results for mining rare itemsets than standard approaches. This study include three different approaches of the Apriori-Inverse algorithm, where all of them return the exact same results being the unique difference the approach followed to parallelize. To run these algorithms a maximum and minimum support threshold of 0.001 and 0.32 have been considered, as well as a threshold of confidence of 0.2. These values are set on this way since the data are randomly generated.

In this regard, Fig. 3 illustrates how the number of instances affects the performance of different approaches. As it could be appreciated when the number of instances varies from $2 \cdot 10^6$ to $2 \cdot 10^7$ the sequential approach is the most efficient. Apriori-InverseMR Hadoop is far from being efficient and Spark obtains a performance more similar to sequential approach. However, when this number of instances continues growing until $8 \cdot 10^7$, sequential performance's is hampered being an approach based on Spark more efficient.

Continuing this study, the number of instances is drastically increased up to $6 \cdot 10^{10}$. The sequential approach is meaningless since its runtime require several hours whereas our approaches only requires some minutes to be run achieving the exact same results. With this number of instances the performance of Hadoop

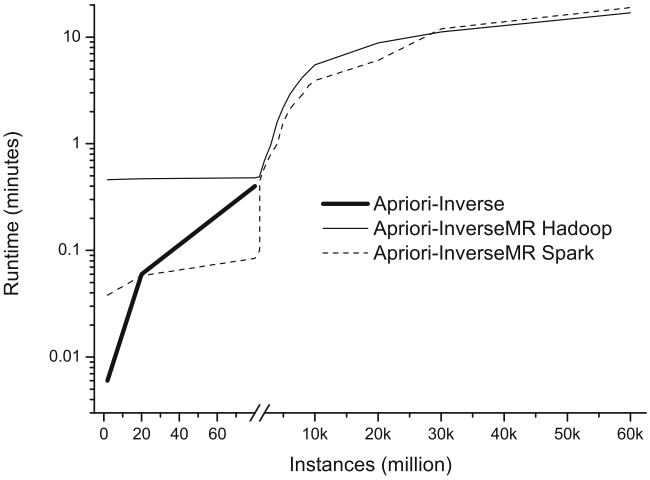


Fig. 3. Runtime of different implementations of Apriori-Inverse algorithm when they are run on datasets comprising a search space of $2^8 - 1$ itemsets and a number of instances from $2 \cdot 10^6$ to $6 \cdot 10^{10}$. A threshold of minimum and maximum support of 0.001 and 0.32 have been considered respectively, and a threshold of 0.2 for confidence.

and Spark is almost the same since the size of the data is bigger than the available RAM provoking that Spark cannot store the whole dataset in memory, but it also needs to use disk imitating the behavior of Hadoop. Thus, it could be concluded that sequential approach is recommended with only some million of instances. Spark should be used when the number of instances is bigger than 20 million, whereas Hadoop could be used when the data is as large that it cannot be stored in main memory.

As not only the number of instances affects the performance, it also has been considered interesting to analyze how the search space influences in it. On this matter, Fig. 4 illustrates the performance when the number of instances increases up to $2 \cdot 10^7$ and the search space varies from $2^8 - 1$ to $2^{40} - 1$ itemsets. It could be appreciated that sequential approach is highly affected by the search space taking more than 6 days to be run on search spaces of $2^{16} - 1$ itemsets. However, our proposals based on MapReduce achieves the exact same results using only 0.1% of the time required by sequential approach. Although all implementations increase exponentially with the search space, sequential approach does in a more significant way. Regarding to implementations based on MapReduce, their behavior are the same as the previous study, Spark achieves better performance than Hadoop always that the dataset can be stored in main memory.

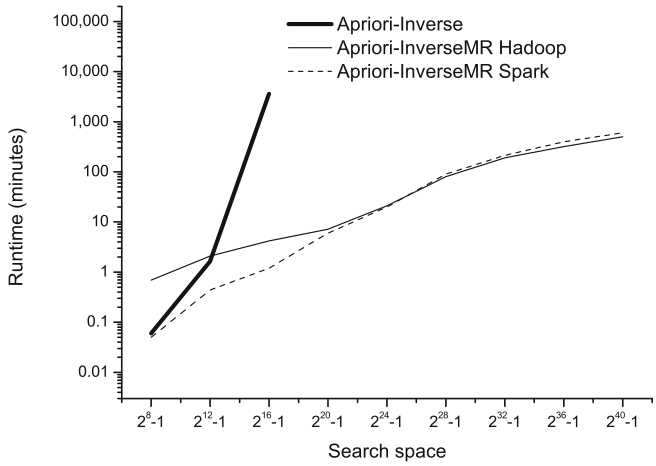


Fig. 4. Runtime of different implementations of Apriori-Inverse algorithm when they are run on datasets comprising a search space from $2^8 - 1$ to $2^{40} - 1$ itemsets and a number of instances of $2 \cdot 10^7$. A threshold of minimum and maximum support of 0.001 and 0.32 have been considered respectively, and a threshold of 0.2 for confidence.

5 Conclusions

In this paper, an adaptation of Apriori-Inverse have been proposed to be used over large amounts of data, relying on the MapReduce framework and the Hadoop and Spark open-source implementations. The aim of our proposal is to obtain perfectly rare associations on Big Data, enabling to describe rare cases, unexpected phenomena or disbelief in the domain.

Our proposals have been compared to another highly efficient algorithm in the field, showing that our proposal outperforms it when massive datasets are considered. The experimental stage revealed that the proposed algorithm perform really well when both the search space increases (a maximum search space of $2^{40} - 1$ items were used) and the number of instances drastically increases, achieving a good performance even for datasets comprising up to 60,000 million of different instances and file sizes of 500 GBytes.

Acknowledgments. This research was supported by the Spanish Ministry of Economy and Competitiveness, project TIN-2014-55252-P, and by FEDER funds.

References

1. Ventura, S., Luna, J.M.: Pattern Mining with Evolutionary Algorithms. Springer International Publishing, Switzerland (2016)
2. Koh, Y.S., Rountree, N.: Finding sporadic rules using apriori-inverse. In: Ho, T.B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 97–106. Springer, Heidelberg (2005). doi:10.1007/11430919_13

3. Romero, C., Luna, J.M., Romero, J.R., Ventura, S.: Mining rare association rules from e-Learning data. In: Proceedings of the 3rd International Conference on Educational Data Mining, EDM 2010, pp. 171–180 (2010)
4. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008). 50th anniversary issue
5. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. SIGMOD Rec. **22**(2), 207–216 (1993)
6. Szathmary, L., Napoli, A., Valtchev, P.: Towards rare itemset mining. In: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2007, Washington, DC, USA, pp. 305–312. IEEE Computer Society (2007)
7. Padillo, F., Luna, J.M., Ventura, S.: Subgroup discovery on big data: exhaustive methodologies using map-reduce. In: Proceedings of the 2016 IEEE Trustcom/BigDataSE/ISPA, Tianjin, China, pp. 1684–1691, August 2016
8. Padillo, F., Luna, J.M., Cano, A., Ventura, S.: A data structure to speed-up machine learning algorithms on massive datasets. In: Proceedings of the 11th International Conference on Hybrid Artificial Intelligent Systems, pp. 365–376 (2016)
9. Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I.: Spark: Cluster computing with working sets. In: Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, HotCloud 2010, Berkeley, CA, USA (2010)
10. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2011)

Patient-Specific Epilepsy Seizure Detection Using Random Forest Classification over One-Dimension Transformed EEG Data

Marco A. Pinto-Orellana(✉) and Fabio R. Cerqueira

Departamento de Informática, Universidade Federal de Viçosa, Viçosa, Brazil
pinto.marco@live.com, fabio.cerqueira@ufv.br

Abstract. This work presents a computational method for improving seizure detection for epilepsy diagnosis. Epilepsy is the second most common neurological disease. It impacts between 40 and 50 million of patients in the world. However, epilepsy diagnosis using electroencephalographic signals implies a long and expensive process involving medical specialists. The proposed system is a patient-specific system which performs an automatic detection of seizures from brainwaves applying a random forest classifier. Features are extracted using one-dimension reduced information from a spectro-temporal transformation of biosignals that pass through an envelope detector. The performance of the present method reached 97.12% of specificity, 99.29% of sensitivity, and a $0.77h^{-1}$ false positive rate. Therefore, the method hereby proposed has great potential for diagnosis support in clinical environments.

1 Introduction

Epilepsy is the second most common neurological disease in humans after stroke [1]. It is estimated between 40 and 50 million of patients suffering from this condition worldwide, which means 1% of the total world population [1–3]. Although there are several medical treatments, 30% of the patient population have not a positive response to medication [4] and require expensive and long diagnosis processes [5].

Seizures are the typical indicators for epilepsy diagnosis [2]. A seizure is an abnormal excessive and hyper-synchronized neural activity in the brain [2,4,5] nd it can be seen through electrical variations recorded from the whole brain mass or specific sections on the brain structure [6].

Among several imaging techniques of biosignals with medical relevance [7], electroencephalograms (EEGs) are some of the most relevant alternatives. EEGs are recordings of electrical time-dependent variations of the brain activity [7,8]. Although the electrical amplitude measurable in each neuron is small, the result of the synchronization in time and phase of signals from large neural networks during cognitive activities allows detecting the neuron electrical signals even at scalp with EEG [9]. It is a safe, non-invasive, and easy to assemble method [10].

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_51

Those advantages established EEG as the *de facto* standard, along with video monitoring, for epilepsy seizure diagnosis [11].

Medical recommendations for seizure diagnosis often include performing long brain activity recordings of patients [12]. Then, the obtained registers are analyzed by experts in the area relying on visual and subjective inspection [4,5]. This human dependency implies an expensive and time-consuming process prone to errors due to the stored data size [12].

It is important to notice that not all epileptiform waves appear during seizure intervals [6]. Paroxysmal activity is an abnormal synchronous discharge of a large ensemble of neurons and is strongly related to seizure processes. Such an activity can be detected in EEG and is usually confused with an epilepsy marker [2].

Several alternatives have been developed to recognize seizures in EEG [13,14]. A first distinction among the approaches is the specific domain of the EEG data on which each method focuses: Time series, frequency data, or spectro-temporal signals [15,16]. Another significant difference is the set of features that each method applies in the analysis. Some studies utilized statistics, chaos theory, or information theory parameters [15–17], while other efforts applied data transformations such as singular value decomposition (SVD) or principal component analysis (PCA) [18]. Furthermore, support vector machines (SVM), k-nearest neighbours (KNN), and artificial neural networks (ANN) are typically used as classifiers [13,14].

This study proposes a novel approach that presents a higher performance compared with other state of the art studies. The method herein described is a combination of several techniques of signal processing and machine learning: Short time Fourier transform, principal component analysis, maximum moving filter, and a random forest classifier.

2 Materials

In this research, the data source relies on the CHB-MIT electroencephalographic scalp database from the Physionet project [19]. This database stores epileptic seizures from several pediatric patients at the Children's Hospital Boston [20,21]. The data were collected for an experiment conducted for monitoring patients after withdrawal of the epileptic medication before a surgical intervention. A total of 24 datasets were built from 23 different subjects. Each dataset was labeled with a code from *chb01* to *chb24*. It should be noted that datasets *chb02* and *chb24* represent biosignals of the same patient but with a difference of 1.5 years between the recording dates. However, they are identified as different subjects in our experiments.

The complete data package is compounded of 686 files saved in the European data format (EDF) representing a total of 961.64 h of recording [19]. The seizure recording durations of the patients are in the range of 6 s (subject *chb16*) to 752 s (subject *chb11*). All signals were measured with a sampling frequency of 256 Hz with 16 bits of resolution. Nearly, all files were captured in 23 channels using the international 10-20 system of electrode positions and nomenclature.

3 Method

This study proposes to use a mapping of a spectro-temporal transform of the brain signals into a one-dimension space for being used as input to a classifier algorithm (Fig. 1). As a result, the complete signal analysis could be explained using four processes: Data preprocessing, dimensionality reduction, envelope detection, data regrouping, and classification.

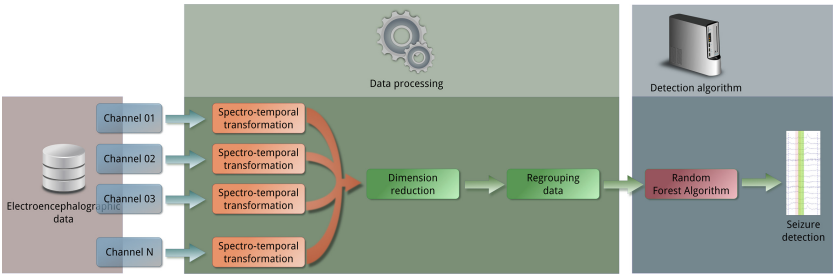


Fig. 1. Workflow of the seizure detection process: Data preprocessing, dimensionality reduction, envelope detection, data regrouping, and classification.

3.1 EEG Data Processing

Some characteristics are important to be explained about EEG signals. They exhibit several unique features: Nonstationary, nonlinear, and frequency-variant. The nonstationary property denotes that there is no period of time where a given signal is repeated [22,23]. Also, previous studies proved that variations in the information source on different conditions or activities are linked with alterations in the spectral response of the brain biosignals [23]. For this reason, the deviations that appear in the frequency domain are useful to be analyzed. The conventional alternative is a discrete Fourier transform or a fast Fourier transform [14]. However, those approaches are focused in stationary data series and cannot be directly applied in EEG [22].

A typical approach to analyze nonstationary signals is splitting them into several time intervals, or time windows, such that spectral features can be obtained, while partial time information is retained [2]. This method, known as short time Fourier transform (STFT), showed effectiveness to detect seizures [23,24]. Therefore, this mathematical tool was selected as the procedure to map EEG signals into a time-frequency domain [25]. For a given discrete function $x[n]$, STFT is calculated shifting a small sliding window $g[n]$ of size Δt over the time series to obtain the frequency spectrum in each time interval:

$$G\{x[n]\}[m,f] = \sum_{n=-\infty}^{\infty} x[n]\, g[n-m]\, e^{-j2\pi fn} \tag{1}$$

where n and m are sampling instants, and f is the signal frequency.

Due to the fact that the signal is not continuous, i.e., it is sampled at fixed periods with a frequency f_s, there are several limitations in the process of STFT calculation. Therefore, some constraints must be considered, because the time resolution Δt and frequency resolution Δf are related to the number of points Δn, as can be seen below:

$$\Delta t = \Delta n \cdot f_s \tag{2}$$

$$\Delta f = \frac{f_s}{2\Delta n} \tag{3}$$

Thus, for the datasets described previously, a time window $\Delta t = 39\ \mu s$ was used, which allowed a range of 10 frequency intervals with a resolution $\Delta f = 12.8 Hz$

3.2 Data Dimensionality Reduction

Electroencephalograms are multivariate data and their computed values present a high dimensionality, 230 features per instance in this research, which complicates data visualization and processing [26]. This problem, known as "the curse of dimensionality", can impact the classifier performance increasing misinterpretations or overinterpretation of the information [15]. Common procedures to reduce these inconveniences rely on selecting the most significant features. In this work, we chose principal component analysis for feature selection, as described in [14].

PCA was applied over the EEG multivariate time series maintaining only the component with maximum score. The aim of mapping to only one dimension is allowing to represent graphically the whole brain signal in one chart maintaining a relationship between the signal shape and the EEG epilepsy state.

3.3 Signal Envelope Detection and Regrouping Data

Previous experiments (not shown) have revealed several associations between the PCA reduced data and the seizure marks. This connection is apparently related to the PCA signal shape, as can be seen in Fig. 2. It should be noted that this behavior corresponds to amplitude modulated signals. As a consequence, it was applied an



Fig. 2. Graphical comparison between the PCA converted data and its envelope regarding the seizure marks in the datasets *chb01_04*, *chb03_03* and *chb03_04*.

envelope detection process over the PCA signal. There are several methods available for this goal: Hilbert transform, pass band filters, and hysteresis transform. However, after a trade-off analysis between processing time and resultant shape, the simplest off-line method was selected: The moving maxima filter.

The moving maxima filter process consists of creating a new series $h[n]$ from an original series $g[n]$, such that each point n represents the maximum value between $g[n]$ and $g[n + \Delta n]$:

$$h[n] = h'[n\Delta n] = \max_{n \le i \le n+\Delta n} \{g[i]\} \tag{4}$$

The goal of this filter, proposed by Zhang and Smith [27], is finding the envelope of the signal using amechanism similar to an amplitude demodulator, which depends only on the time period Δn, defined as 10 s in this work.

After this process, the enveloped data were reorganized in blocks of fixed size m. This size may influence the final system ability to detect a seizure state and it should be tested with several values. However, due to the limitations related to the data size, only two block sizes were tested: 30 and 70 s.

3.4 Machine Learning: Balancing and Classification

The last step in our method is to apply a machine learning algorithm. However, EEG datasets are typically imbalanced [12], i.e., the points marked as seizure comprise a very small set in comparison to the interictal intervals. As a result, firstly, the SMOTE algorithm was applied over the data for treating the imbalance problem. In [28], this oversampling technique was used effectively in other EEG analyses to improve recognition in the dataset.

After the application of the SMOTE method, a random forest classifier was configured and executed. Random forest is an ensemble learning algorithm that uses multiple decision trees as described in detail in [29].

Several studies have shown that random forest works satisfactorily with features extracted from EEG data with reasonably less training time [29–31]. Additionally, in the seizure classification context, a previous study reached average accuracies greater than 96% using random forest on the epilepsy dataset of the University of Bonn [25].

In this study, a random forest classifier was configured to work with 100 decision trees without prune restriction, and with $\log_2(N_{attributes}) + 1$ random features for each tree. We used the random forest implementation of the WEKA API 1.7.0 [32].

4 Results

The method proposed in this work was applied to the physiological signals stored in the CHB-MIT database. Every channel was processed with the STFT transform and PCA for reducing the multivariate set to only one time series. Next, each resultant time series passed through an envelope detector and, finally, the detected envelope was reorganized in blocks of fixed length to be processed by a random forest classifier.

Table 1. Performance evaluation parameters with block sizes of 30 and 70 s

Patient ID	Block size: 30 s					Block size: 70 s				
	ACC (%)	SEN (%)	SPE (%)	FPRe	T_{exec} (s)	ACC (%)	SEN (%)	SPE (%)	FPRe	T_{exec} (s)
chb01	96.55	94.38	98.38	2.27	60.73	99.62	99.24	99.95	0.16	87.91
chb02	87.46	84.40	90.05	12.13	61.45	97.26	96.38	98.00	2.42	82.6
chb03	96.50	93.74	98.84	1.53	47.02	98.76	97.38	99.94	0.08	77.22
chb04	94.41	91.14	97.20	15.35	369.31	99.54	99.06	99.96	0.23	509.67
chb05	96.25	95.61	96.79	4.33	33.21	99.82	99.73	99.91	0.13	38.45
chb06	89.79	86.57	92.52	17.51	125.64	99.54	99.07	99.93	0.16	147.88
chb07	88.56	85.95	90.78	21.68	137.53	99.27	98.69	99.77	0.53	185.51
chb08	90.89	86.17	94.85	3.55	27.75	96.41	93.50	98.85	0.78	42.49
chb09	**98.41**	**98.38**	**98.44**	**3.71**	**65.51**	**99.90**	**99.85**	**99.94**	**0.14**	**93.5**
chb10	94.80	91.89	97.28	4.75	75.23	98.38	97.25	99.34	1.14	115.82
chb11	96.39	94.21	98.24	2.12	51.01	98.81	97.83	99.63	0.44	86.89
chb12	**86.13**	**82.34**	**89.33**	**8.64**	**39.81**	95.14	93.39	96.62	2.72	50.16
chb13	92.62	90.15	94.71	6.03	55.9	97.30	95.83	98.54	1.65	78.79
chb14	86.73	84.26	88.83	10.03	40.29	98.08	97.16	98.86	1.02	47.16
chb15	89.13	83.62	93.78	8.57	74.6	**95.00**	**91.67**	**97.80**	**3.00**	**122.97**
chb16	87.40	83.37	90.82	6.02	30.34	97.66	95.74	99.29	0.46	35.31
chb17	93.84	89.55	97.47	1.75	27.36	97.77	95.34	99.83	0.12	47.74
chb18	89.30	88.00	90.40	11.82	59.82	97.42	96.39	98.29	2.08	76.07
chb19	96.08	97.10	95.22	4.95	35.8	99.85	99.80	99.89	0.12	37.34
chb20	95.76	92.70	98.36	1.56	32.47	99.21	98.40	99.90	0.09	46.39
chb21	95.62	92.29	98.46	1.75	46.77	99.11	98.16	99.93	0.08	80.56
chb22	89.28	84.69	93.17	7.32	50.82	98.40	96.69	99.85	0.16	84.12
chb23	94.28	91.19	96.90	2.88	31.32	98.25	96.43	99.78	0.20	53.77
chb24	92.76	91.79	93.58	4.71	26.2	98.61	97.88	99.23	0.56	32.29
Avg.	**92.46**	**89.73**	**94.77**	**6.87**	**66.91**	**98.30**	**97.12**	**99.29**	**0.77**	**94.19**
Min.	86.13	82.34	88.83	1.53	26.20	95.00	91.67	96.62	0.08	32.29
Max.	98.41	98.38	98.84	21.68	369.31	99.90	99.85	99.96	3.00	509.67

As mentioned previously, two possibilities in the re-sequencing process were considered: Blocks of 30 and 70 s. Different sizes showed an influence in the recognition capacity with the transformed signal.

All experiments performed, for each configuration, were assessed using 10-fold cross validation. Regarding the performance indexes, standard parameters were used: Accuracy (ACC), specificity (SPE), sensitivity (SEN), and false positive rate ($FPRe$) measured in samples per hour (h^{-1}) [4,5,12,15,17,18,33]:

$$\text{ACC} = \frac{TN + TP}{TN + TP + FP + FN} \tag{5}$$

$$\text{SPE} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{SEN} = \frac{TN}{TN + FP} \tag{7}$$

$$\text{FPRe} = \frac{FP}{\text{time (in hours)}} \tag{8}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, FN is the number of false negatives, and time is a unit of time analysis.

The obtained results, with the execution time T_{exec} of the entire 10-fold cross validation process, were compiled in Table 1. It should be noted that the time was calculated using a personal computer with a Core i7 Quad Core processor (3.40 GHz) using Ubuntu 15.10 as operating system.

The experiments showed a good performance with data blocks of 30 s, reaching average sensitivity of 89.73%, specificity of 94.77%, $FPRe$ of $6.87h^{-1}$, and accuracy of 92.46%. Nonetheless, increasing the data block size to 70 s significantly improved the performance. The average values of the same measures were raised to: 97.12% of sensitivity, 99.29% of specificity, 98.30% of accuracy, and $0.77\,h^{-1}$ of $FPRe$.

5 Discussion

It should be noted that the database used in this paper is publicly available [19]. As a result, there are other seizure detection methods that process the same information to evaluate their algorithms. We selected 9 methods from 7 studies that analyzed the same database to compare with our approach. Their performances are compiled in Table 2, although some indexes were not found in their evaluation analyses. For instance, some researches emphasized the importance of the sensibility in the evaluation over other indexes and, thus, presented only sensibility values [12,33].

Table 2. Performance comparison with other methods

Study	Year	ACC	SPE	SEN	FPRe	Classifier
[34]	2014	94.80%	95.70%	91.70%	-	Linear SVM
[16]	2015	94.37%	93.84%	93.81%		RBF-SVM
[17]	2015	86.93%	86.26%	87.58%	-	GMM
[33]	2015	-	-	100.00%	0.8	-
[12]	2015	-	-	99.00%	-	-
[4]	2016	96.25%	99.90%	92.60%	0.3	LDA
[4]	2016	89.80%	99.70%	79.90%	3.9	NN
[18]	2016	100.00%	-	-	-	SVM
The proposed method	2016	98.30%	99.20%	97.12%	0.77	Random forest

The current research proposes a different method to process electroencephalographic signals with a high detection accuracy. Comparing with alternative methods, our technique showed the best overall performance. Sensitivity was just behind the technique developed in [33] that was based on a classification

with entropy-related parameters, which obtained a 100% for SEN, but did not present other important evaluation parameters.

Our approach obtained slightly lower values concerning specificity and false positive rate (a difference of 0.70% and 0.44 h^{-1}) compared with one of the methods presented in [4]. However, in general, our method demonstrated to be the best approach when all parameter values are taken into account in combination.

A good system response depends on the combination of the internal algorithms selected. Concerning feature selection, the majority of previously proposed techniques rely, at least partially, on spectral features: STFT, wavelet transform, or discrete Fourier transform [4,5,12,15,17]. Furthermore, dimensionality reduction using PCA, or an improved algorithm based on it, proved to lead to learning models that reach accuracies near 100% [18]. The present research combined both processing types using STFT over biosignals and then PCA on the time-spectrum data. The notable difference with respect to other studies is the use of a singular element typical of amplitude demodulators: An envelope detector to improve the match between waveforms and seizure marks. Regarding the classification method, several authors used SVM [5,15,18], Bayessian classifiers [17], linear discriminators or neural networks [4]. In this work, the random forest method demonstrated a very good predictive power, with the advantage of a quick training time [31].

6 Conclusions

This study describes a patient-specific system for detecting seizures with a great prediction power. The structure of the proposed approach comprises a sequence of several signal processing and machine learning algorithms. The processing step includes the application of PCA after a STFT transformation with a subsequent envelope shape detection stage. After those procedures, the resultant data are regrouped in fixed length blocks of time that are then given as input to a random forest classifier.

The parameters of the proposed system were optimized to maximize the overall performance. Thus, after training and testing with the CHB-MIT database containing 23 subject cases, we could obtain average values of 98.30% for accuracy, 97.12% for sensitivity, 99.29% for specificity, and 0.77h^{-1} as false positive rate. Comparing these indexes with the results of state-of-art alternative approaches, we can conclude that a hardware implementation of our method could lead to a considerable positive impact on epilepsy diagnosis through the automation of seizure detection.

Acknowledgement. This work is supported by grants from Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) and the PAEC agreement between the Organization of American States (OAS) and Universidade Federal de Viçosa (UFV).

References

1. Guo, L., Rivero, D., Dorado, J., Rabuñal, J.R., Pazos, A.: Automatic epileptic seizure detection in EEGs based on line length feature and artificial neural networks. J. Neurosci. Methods **191**(1), 101–109 (2010)
2. Tzallas, A., Tsipouras, M., Fotiadis, D.: Epileptic seizure detection in EEGs using time and frequency analysis. IEEE Trans. Inf. Technol. Biomed. **13**(5), 703–710 (2009)
3. Fatichah, C., Iliyasu, A.M., Abuhasel, K.A., Suciati, N., Al-Qodah, M.A.: Principal component analysis-based neural network with fuzzy membership function for epileptic seizure detection, pp. 186–191. IEEE (2014)
4. Orosco, L., Correa, A.G., Diez, P., Laciar, E.: Patient non-specific algorithm for seizures detection in scalp EEG. Comput. Biol. Med. **71**, 128–134 (2016)
5. Das, A.B., Bhuiyan, M.I.H., Alam, S.M.S.: Classification of EEG signals using normal inverse Gaussian parameters in the dual-tree complex wavelet transform domain for seizure detection. Signal Image Video Process. **10**(2), 259–266 (2016)
6. Sierra-Marcos, A., Scheuer, M.L., Rossetti, A.O.: Seizure detection with automated EEG analysis: a validation study focusing on periodic patterns. Clin. Neurophysiol. **126**(3), 456–462 (2015)
7. Teplan, M.: Fundamentals of EEG measurement. Meas. Sci. Rev. **2**(2), 1–11 (2002)
8. Djemili, R., Bourouba, H., Amara Korba, M.: Application of empirical mode decomposition and artificial neural network for the classification of normal and epileptic EEG signals. Biocybern. Biomed. Eng. **36**(1), 285–291 (2016)
9. Hesse, W., Möller, E., Arnold, M., Schack, B.: The use of time-variant EEG granger causality for inspecting directed interdependencies of neural assemblies. J. Neurosci. Methods **124**(1), 27–44 (2003)
10. Hu, B., Liu, Q., Zhao, Q., Qi, Y., Peng, H.: A real-time electroencephalogram (EEG) based individual identification interface for mobile security in ubiquitous environment, pp. 436–441. IEEE (2011)
11. Page, A., Pramod Tim Oates, S., Mohsenin, T.: An ultra low power feature extraction and classification system for wearable seizure detection, pp. 7111–7114. IEEE (2015)
12. Tsiouris, K.M., Konitsiotis, S., Markoula, S., Koutsouris, D.D., Sakellarios, A.I., Fotiadis, D.I.: An unsupervised methodology for the detection of epileptic seizures in long-term EEG signals, pp. 1–4. IEEE (2015)
13. Orosco, L., Correa, A.G., Laciar, E.: Review: a survey of performance and techniques for automatic epilepsy detection. J. Med. Biol. Eng. **33**(6), 526–537 (2013)
14. Alotaiby, T.N., Alshebeili, S.A., Alshawi, T., Ahmad, I., El-Samie, F.E.A.: EEG seizure detection and prediction algorithms: a survey. EURASIP J. Adv. Signal Process. **2014**(1), 1–21 (2014)
15. Ahmad, M.A., Ayaz, Y., Jamil, M., Omer Gillani, S., Rasheed, M.B., Imran, M., Khan, N.A., Majeed, W., Javaid, N.: Comparative analysis of classifiers for developing an adaptive computer-assisted EEG analysis system for diagnosing epilepsy. Biomed. Res. Int. **2015**, 1–14 (2015)
16. Das, A.B., Pantho, M.J.H., Bhuiyan, M.I.H.: Discrimination of scalp EEG signals in wavelet transform domain and channel selection for the patient-invariant seizure detection, pp. 77–80. IEEE (2015)
17. Gill, A.F., Fatima, S.A., Usman Akram, M., Khawaja, S.G., Awan, S.E.: Analysis of EEG signals for detection of epileptic seizure using hybrid feature set. In: Sulaiman, H.A., Othman, M.A., Abd. Aziz, M.Z.A., Abd Malek, M.F. (eds.) Theory and Applications of Applied Electromagnetics. LNEE, vol. 344, pp. 49–57. Springer, Heidelberg (2015). doi:10.1007/978-3-319-17269-9_6

18. Zhao, W., Qu, J., Chai, Y., Tang, J.: Classification of seizure in EEG signals based on KPCA and SVM. In: Jia, Y., Du, J., Li, H., Zhang, W. (eds.) Proceedings of the 2015 Chinese Intelligent Systems Conference, pp. 201–207. Springer, Heidelberg (2016)
19. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
20. Shoeb, A., Edwards, H., Connolly, J., Bourgeois, B., Treves, S.T., Guttag, J.: Patient-specific seizure onset detection. Epilepsy Behav. **5**(4), 483–498 (2004)
21. Shoeb, A.H., Guttag, J.V.: Application of machine learning to epileptic seizure detection, pp. 975–982 (2010)
22. Natarajan, K., Acharya, R., Alias, F., Tiboleng, T., Puthusserypady, S.K.: Nonlinear analysis of EEG signals at different mental states. BioMed. Eng. OnLine **3**(1), 7 (2004)
23. Yan, A., Zhou, W., Yuan, Q., Yuan, S., Wu, Q., Zhao, X., Wang, J.: Automatic seizure detection using stockwell transform and boosting algorithm for long-term EEG. Epilepsy Behav. **45**, 8–14 (2015)
24. Samiee, K., Kovacs, P., Gabbouj, M.: Epileptic seizure classification of EEG time-series using rational discrete short-time fourier transform. IEEE Trans. Biomed. Eng. **62**(2), 541–552 (2015)
25. Wang, G., Deng, Z., Choi, K.-S.: Detection of epileptic seizures in EEG signals with rule-based interpretation by random forest approach. In: Huang, D.-S., Han, K. (eds.) ICIC 2015. LNCS (LNAI), vol. 9227, pp. 738–744. Springer, Heidelberg (2015). doi:10.1007/978-3-319-22053-6_78
26. Birjandtalab, J., Pouyan, M.B., Nourani, M.: Nonlinear dimension reduction for EEG-based epileptic seizure detection, pp. 595–598. IEEE (2016)
27. Zhang, Z., Smith, R.L.: Modeling financial time series data as moving maxima processes (2001)
28. Fergus, P., Hignett, D., Hussain, A., Al-Jumeily, D., Abdel-Aziz, K.: Automatic epileptic seizure detection using scalp EEG and advanced artificial intelligence techniques, pp. 1–17 (2015)
29. Shiratori, T., Tsubakida, H.A., Ishiyama, H., Ono, Y.: Three-class classification of motor imagery EEG data including "rest state" using filter-bank multi-class common spatial pattern. In: 3rd International Winter Conference on Brain-Computer Interface (BCI), pp. 1–4. IEEE (2015)
30. Patti, C.R., Shahrbabaki, S.S., Dissanayaka, C., Cvetkovic, D.: Application of random forest classifier for automatic sleep spindle detection. In: Biomedical Circuits and Systems Conference (BioCAS), pp. 1–4. IEEE (2015)
31. Czarnecki, M., Gustafsson, N.: Machine learning for detection of epileptic seizures (2015)
32. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newslett. **11**(1), 10–18 (2009)
33. Iqbal, S., Khan, Y.U., Farooq, O.: Nonlinear analysis of EEG for seizure prediction, pp. 1–5. IEEE (2015)
34. Ahmad, M.A., Khan, N.A., Majeed, W.: Computer assisted analysis system of electroencephalogram for diagnosing epilepsy, pp. 3386–3391. IEEE (2014)

A New Approach to Human Activity Recognition Using Machine Learning Techniques

Leandro B. Marinho, A.H. de Souza Junior, and P.P. Rebouças Filho(✉)

Programa de Pós-Graduação em Ciência da Computação, Instituto Federal do Ceará, Maracanaú, CE, Brazil
leandro.marinho@ppgcc.ifce.edu.br, {amauriholanda,pedrosarf}@ifce.edu.br

Abstract. Recognition of human activities aims a wide diversity of applications. However, identifying complicated activities continues a challenging and active research area. In this work, we assess a new approach of feature selection for human activity recognition. For the task, we also compare state-of-the-art classifiers, e.g., Bayes classifier, kNN, MLP, SVM, MLM and MLM-NN. Based on the experiments, the MLM-NN is able to speed up the original MLM while holding equivalent accuracy. MLM and SVM achieved accuracy of more than 99.2% in the original data set and 98.1% using new feature selection method. Results show that the proposed feature selection approach is a promising alternative to activity recognition on smartphones.

Keywords: Human activity recognition · Machine learning · Feature selection

1 Introduction

Recognizing human activities by means of sensors attached on the body has been widely studied [3]. Common activity and functional performance level of a person can be determined by the capability to record and identify distinctive daily activities [21].

In the health care field, initial detection of diseases could be obtained through long term inquiry of human activity [9] or to stimulate people to be physically active and improve their exercise opportunities [24]. As far as physiotherapy, it can help to comprehend if an exercise is been correctly accomplished or even to monitor possible disordered [15].

Security and entertainment are others fields impacted by investigation of human behavior through mobile phone data. Mobile phones were used in police investigations to track suspects and victims [29]. Regarding to entertainment, Badawi and Saddik, in [5], proposed a mobile application that provides users with physical activity recommendations on a daily basis.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_52

Motion capture video systems have become an important research topic in the monitor human activity. Dubois and Charpillet, in [12], proposed a method to detect fall using a system made up of RGB-Depth cameras. In [17], Jalal *et al.* used a depth imaging sensor to capture depth silhouette to recognize the daily activities of elderly people. These methods are obtrusive, need large equipments and could be used in a controlled environment. In addition, they require high memory space to record it likewise large processing time [9]. The use of vision-based method is inadvisable under the situations of occlusions, lighting condition change, complex background, viewpoint change and so forth [34].

An viable alternative to video systems is motion sensors. They are small, have low cost, and are able to record motion signals within wearable and unobtrusive systems. Accelerometers and gyroscopes, the two major motion sensors, were used in research for fall detection [19], analysis of energy expenditure in physical activities [18] and daily activity monitoring [9]. Chen *et al.*, in [7], used information of the motion sensor signals to identify the exercise type and classify if their postures are proper or not. Olivares *et al.*, in [23], proposed an approach of inertial magnitude-based detectors and compares with well known others. Bayat *et al.*, in [6], showed how to identify some human physical activities using a smartphone with accelerometer. In [30], Shoaib *et al.* compared the effect of accelerometer, gyroscope and linear acceleration sensor at both wrist and pocket positions using smartphone and wrist-worn motion sensors.

Essentially, activity recognition requires a math model which enables the identification of interest. In general, models can be created using the knowledge of the specialist about the phenomenon which needs shape or by techniques of machine learning (statistical and neural). A machine learning model, among other benefits, allows us build models based on with little or no prior knowledge about the task of interest. In this regard, examples of pattern should be captured or available [27].

The aim of this paper is to evaluate a new approach of feature selection along with the PCA technique and compare the performance of several machine learning techniques for activity monitoring. The learning algorithms used in the comparison are k-Nearest Neighbors (kNN), Artificial Neural Network Multilayer Perceptron (MLP), Support Vector Machines (SVM), Bayes classifier, Minimal Learning Machine (MLM) and Minimal Learning Machine with Nearest Neighbors (MLM-NN). For the recently proposed Minimal Learning Machine, we evaluate the impact of using different distance metrics. We propose a simple and effective approach of feature selection. Our method significantly reduces the number of features and therefore increases the speed of sensing over the human activity recognition task. With this approach the time step of the acquisition and classification is decreased, then the final application, which consists exclusively these two steps, will become much faster in general.

2 Machine Learning Methods

In this section, the six machine learning techniques used in our comparison are defined.

Minimal Learning Machine (MLM) is a method for supervised learning which assumes a mapping between points in the input and output space. MLM computes this by a simple linear model between distance matrices. The output point is located by the result calculated and thus an estimate for the response is provided Souza *et al.* in [31]. The only hyper-parameter of MLM is number of reference points K. MLM can be extended for classification problems representing the S classes in vector form through the codification binary scheme (1-of-S).

Calculating the output of MLM through the optimization process can be computationally "expensive". Thus, a quick approach to estimate the output step is proposed by Mesquita *et al.* in [20]. Once the distances have been estimated, test sample class is estimated by the classes of k the nearest points of references in the output space $\mathbf{y}$. The class of test sample is chosen by the mode of the k classes of their nearest neighbours. In this paper, MLM with the approach of the nearest neighbors will be referenced by MLM-NN (MLM-Nearest Neighbors).

Multi-layer perceptron (MLP) consists of multiple layers with (nodes) or neurons that interact using weighted connections. Each layer is fully connected to the next one. Between the input and output layers, there can be one or more hidden layers [14,28]. Weights measure the degree of correlation between the activity levels of neurons that they connect. Moreover, a training algorithm needs to be used to adjust the weights [4]. The most popular, the Backpropagation, is composed by two phases [14].

Support Vector Machine (SVM) is a supervised learning model used for classification and regression analysis. SVM is underpinned by statistical learning theory and developed by Vapnik in [33]. This theory establishes a series of principles which must be followed to obtain classifiers with good generalization capacity. The name, Support Vector Machine, emphasizes the importance of the nearest vectors to separation margin in complexity of SVM. Although, SVM have been originally formulated to solve binary problem, there are some SVM formulations to multi-class problem [8]. However, because its complexity, others approachs such as one-versus-one [11], one-versus-all [11], DAGSVM [26] and error correcting code [10] are the ones that are most widely used. The kernel trick is used to generate non-linear versions, once the standard SVM is linear.

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. This approach is based on quantifying the tradeoff between various classification decisions using probability and the costs that accompany such decisions. It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known [13,32].

k-Nearest Neighbour (kNN) is an algorithm of supervised learning, introduced by *Aha et al.* in [1]. The basic idea behind this algorithm is to find the k closest labeled samples to unclassified samples and to use these labels to classify the unknown sample. kNN require little effort in training stage. In contrast, the computing cost for labeling an unknown sample is relatively high. The number of neighbors is the parameter in kNN.

3 Experiments

This section present data set and the feature selection methods.

3.1 Data Set

We use the Human Activity Recognition Using Smartphones data set created by Anguita *et al.* [2] in the experiments. The data is composed by six activities (walking, walking upstairs, walking downstairs, sitting, standing, laying) and 561 features extracted from gyroscope and accelerometer signals present in a smartphone. These activities were performed by 30 volunteers, with each person wearing a smartphone on the waist, within an age bracket of 19–48 years. Table 1 shows the number of samples by class.

Table 1. Amount of elements by class on database.

Number	Class	N^o of elements
1	Walking	1722
2	Walking down stairs	1544
3	Walking up stairs	1406
4	Sitting	1777
5	Standing	1906
6	Lying	1944

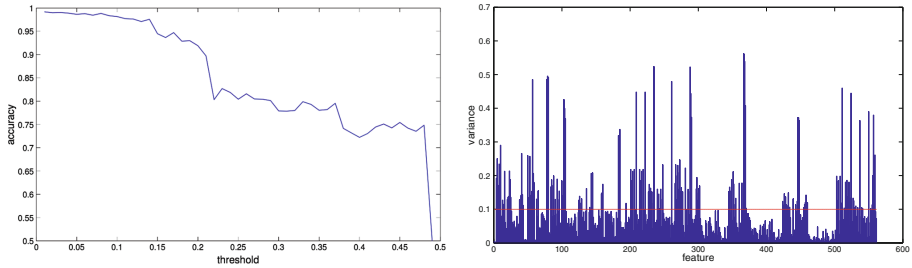
3.2 Feature Selection Methods

In order to reduce the dimensionality of the database, we used the Principal Components Analysis (PCA) [16] to select the features that amounted 98% the information of the eigenvalues. Our approach consists of selecting the features with variance greater than a threshold and after that apply PCA technique, this procedure will be referred to as Var-PCA.

The threshold Var-PCA can be chosen by some method of machine learning. Figure 1(a) shows the trade-off between the threshold values and accuracy of MLM-NN classifier with Euclidean distance, using all training set as K points of references. As can be observed, the accuracy decreases with increasing threshold and, consecutively, with reduced dimensionality. However, there is a range that accuracy is maintained almost constant as the threshold decreases.

Low variance might indicate that the feature is not having an active role in identifying the samples. On the other hand, the higher the threshold, the more aggressive the reduction. Therefore, the decay in the accuracy of the classifiers can be avoid through a suitable choice of threshold value. The optimal threshold can be defined through an optimization step using a classification algorithm. Results show that MLM-NN classifier is a effective method for this purpose.

Figure 1(b) shows the variance for each feature. The red line represents the threshold (0.1) adopted in this paper. Our approach of feature selection considerably reduces the dimensionality. The two data set generated by PCA and Var-PCA have 118 and 39 features, respectively, whereas the original data set has 561 features.



(a) Thresholds (b) Variance for each feature

Fig. 1. (a) Trade-off between the Var-PCA thresholds and accuracy of MLM-NN classifier with Euclidean distance, using all training set as K points of references. (b) Variance for each feature of HAR data set. The red line represents the threshold (0.1) adopted in this paper.

4 Results and Discussion

In this section, we present and discuss the results achieved by machine learning methods on three datasets.

The three data sets have been equally preprocessed. Eighty per cent of samples have been randomly selected to training and the remainder to test, in each of the ten iterations. Training set is then normalized (zero-mean and unit variance) and test sets is also normalized using the same normalization measures from training set. The proportions of the classes are kept balanced in both training and test sets.

The hyper-parameters for the MLM, MLM-NN, MLP and SVM were chosen using 10-fold cross-validation. SVM was trained using the SVM OpenCV toolbox with default settings for the hyper-parameters. The grid search is between 2^{-2} and 2^{11} for each hyper-parameter. Radial basis function and linear function are the kernels used. The strategy one-versus-all is adopted in SVM. MLP is learned using the back-propagation algorithm and a range of hidden units from 1 to 100. Bayes classifier is validated using gaussian (normal) probability distribution. kNN method was configured with 1, 3 and 5 nearest neighbors. MLM and MLM-NN are trained with three distance metrics (Euclidean, Manhattan and Mahalanobis) and ranging the K number of reference points from 5% to 100% (with a step size of 5%) of the training data.

Table 2. Positive predictive value (PPV), sensitivity (Se), Harmonic means (HM) and Accuracy (Acc) for each database, obtained by Bayes, MLP, kNN, SVM, MLM and MLM-NN classifiers.

Original data set					
Classifier	Setup	PPV(%)	Se(%)	HM (%)	Acc(%)
Bayes	**Normal**	**99.04±0.08**	**95.20±0.41**	**98.40±0.14**	**95.20±0.41**
MLP	**561/50/6**	**95.68±2.19**	**78.42±10.97**	**92.81±3.66**	**78.42±10.97**
	561/100/6	92.77±2.79	63.83±13.96	87.94±4.65	63.83±13.96
	561/400/6	92.06±2.45	60.32±12.23	86.77±4.08	60.32±12.23
kNN	1	99.38±0.08	96.91±0.40	98.97±0.13	96.91±0.40
	3	**99.39±0.07**	**96.94±0.34**	**98.98±0.11**	**96.94±0.34**
	5	99.33±0.11	96.65±0.55	98.88±0.18	96.65±0.55
SVM	Linear	99.64±0.04	98.21±0.19	99.40±0.06	98.21±0.19
	RBF	**99.85±0.01**	**99.26±0.06**	**99.75±0.02**	**99.26±0.06**
	Polynomial	99.43±0.13	97.14±0.67	99.05±0.22	97.14±0.67
MLM	**k = 100%, Euclidean**	**99.87±0.04**	**99.41±0.16**	**99.64±0.10**	**99.41±0.16**
MLM-NN	k = 100%, Euclidean	99.86±0.07	99.37±0.33	99.61±0.20	99.37±0.33
MLM-NN	**k = 10%, Euclidean**	**99.60±0.09**	**98.17±0.41**	**98.87±0.25**	**98.17±0.41**
MLM	**k = 10%, Euclidean**	**99.60±0.03**	**98.18±0.14**	**98.88±0.09**	**98.18±0.14**
Data set with PCA					
Classifier	Setup	PPV(%)	Se(%)	HM (%)	Acc(%)
Bayes	**Normal**	**99.46±0.07**	**97.31±0.34**	**99.10±0.11**	**97.31±0.34**
MLP	118/62/6	99.66±0.09	98.30±0.47	99.43±0.16	98.30±0.47
	118/82/6	99.69±0.04	98.43±0.21	99.48±0.07	98.43±0.21
	118/100/6	**99.69±0.06**	**98.45±0.32**	**99.48±0.11**	**98.45±0.32**
kNN	**1**	**99.33±0.08**	**96.65±0.39**	**98.88±0.13**	**96.65±0.39**
	3	99.33±0.08	96.64±0.40	98.88±0.13	96.64±0.40
	5	99.26±0.07	96.28±0.35	98.76±0.12	96.28±0.35
SVM	Linear	99.50±0.07	97.48±0.34	99.16±0.11	97.48±0.34
	RBF	**99.80±0.02**	**99.02±0.11**	**99.67±0.04**	**99.02±0.11**
	Polynomial	98.41±0.32	92.03±1.61	97.34±0.54	92.03±1.61
MLM-NN	**k = 100%, Euclidean**	**99.82±0.03**	**99.19±0.14**	**99.50±0.09**	**99.19±0.14**
MLM	k = 100%, Euclidean	99.81±0.04	99.15±0.17	99.48±0.10	99.15±0.17
MLM-NN	**k = 10%, Manhattan**	**99.21±0.09**	**96.35±0.41**	**97.74±0.26**	**96.35±0.41**
MLM	**k = 10%, Manhattan**	**99.23±0.10**	**96.45±0.47**	**97.79±0.30**	**96.45±0.47**
Data set with Var-PCA					
Classifier	Setup	PPV(%)	Se(%)	HM (%)	Acc(%)
Bayes	**Normal**	**98.94±0.08**	**94.72±0.38**	**98.24±0.13**	**94.72±0.38**
MLP	39/15/6	99.07±0.11	95.35±0.53	98.45±0.18	95.35±0.53
	39/22/6	99.23±0.08	96.14±0.42	98.71±0.14	96.14±0.42
	39/30/6	**99.37±0.11**	**96.86±0.53**	**98.95±0.18**	**96.86±0.53**
kNN	1	98.73±0.10	93.66±0.52	97.89±0.17	93.66±0.52
	3	**98.89±0.06**	**94.45±0.31**	**98.15±0.10**	**94.45±0.31**
	5	98.86±0.05	94.32±0.27	98.11±0.09	94.32±0.27
SVM	Linear	99.05±0.10	95.27±0.51	98.42±0.17	95.27±0.51
	RBF	**99.48±0.06**	**97.41±0.28**	**99.14±0.09**	**97.41±0.28**
	Polynomial	96.69±0.45	83.46±2.23	94.49±0.74	83.46±2.23
MLM-NN	**k = 100%, Mahalanobis**	**99.65±0.07**	**98.41±0.31**	**99.02±0.20**	**98.41±0.31**
MLM-NN	k = 100%, Euclidean	99.61±0.09	98.19±0.41	98.88±0.26	98.19±0.41
MLM-NN	**k = 10%, Euclidean**	**99.24±0.08**	**96.49±0.35**	**97.81±0.23**	**96.49±0.35**
MLM	**k = 10%, Euclidean**	**99.25±0.05**	**96.54±0.26**	**97.85±0.16**	**96.54±0.26**

Table 2 show the average and standard deviation values to accuracy (Acc), sensitivity (Se), positive predictive value (PPV) and harmonic means (HM)[1] obtained for the classifiers based on Bayes, MLP, kNN, MLM, MLM-NN and SVM, using the three case of the database. The results were obtained on an iMac 2.5 GHz processor Core i5 with 4 GB RAM.

The best results achieved by each classifier by three cases of databases, shown in Table 2 are presented in Table 3 detailing the training and testing times, so we high-light the MLM-NN, MLP and SVM methods.

As is shown in the Table 2, the MLM-NN speeded up the original MLM while holding equivalent accuracy. The highest values obtained was 99.41% with the MLM using 100% of training data for reference points and 99.26% with the SVM using RBF kernel.

Table 3. Accuracy (Acc), training time and testing time by classifiers, for each database, obtained by Bayes, MLP, kNN, SVM, MLM and MLM-NN and their best configuration.

Original Data set			
Classifier	Acc(%)	Training Time (s)	Test Time (μs)
MLM (Euclidean, $k = 100\%$)	99.41±0.16	62.648±0.979	71433.148±151.227
SVM-RBF	99.26±0.06	11833.625±94.858	1320.43±53.4
MLM (Euclidean, $k = 10\%$)	98.18±0.14	6.412±0.081	16091.952±272.653
MLM-NN (Euclidean, $k = 10\%$)	98.17±0.41	6.209±0.094	466.563±10.248
kNN ($k = 3$)	96.94±0.34	0.008±0.004	4868.98±61.97
Bayes Normal	95.20±0.41	4319.694±251.635	1705.15±152.88
MLP 561/50/6	78.42±10.97	189.612±34.437	2551.7803±0.207
Data set with PCA			
Classifier	Acc(%)	Training Time (s)	Test Time (μs)
MLM-NN (Euclidean, $k = 100\%$)	99.19±0.14	35.007±0.452	3430.557±48.228
SVM-RBF	99.02±0.11	1194.688±1.744	374.09±15.9
MLP 118/100/6	98.45±0.32	229.043±1.724	639.28±0.01
Bayes Normal	97.31±0.34	0.249±0.012	385.41±53.5
kNN ($k = 1$)	96.65±0.39	0.007±0.018	1167.86±9.56
MLM (Manhattan, $k = 10\%$)	96.45±0.47	6.904±0.707	14523.608±526.131
MLM-NN (Manhattan, $k = 10\%$)	96.35±0.41	8.887±0.899	233.649±89.994
Data set with Var-PCA			
Classifier	Acc(%)	Training Time (s)	Test Time (μs)
MLM-NN (Mahalanobis, $k = 100\%$)	98.41±0.31	129.1372±51.5090	33441.8900±130.9996
SVM-RBF	97.41±0.28	494.0407±13.3079	126.97±16.78
MLP 39/30/6	96.86±0.53	7.369±2.456	38.2746±0.011
MLM (Euclidean, $k = 10\%$)	96.54±0.26	3.8034±0.7849	9332.1170±1552.0267
MLM-NN (Euclidean, $k = 10\%$)	96.49±0.35	3.3613±0.4066	99.5988±27.777
Bayes Normal	94.72±0.38	0.0187±0.011	165.27±67.61
kNN ($k = 3$)	94.45±0.31	0.00054±0.000036	766.04±162.2

[1] The Harmonic Means (HM) between sensitivity and specificity is computed by $HM = 2 \cdot \frac{Se \cdot Sp}{Se+Sp}$.

Table 2 (Data set with PCA) shows the results obtained by the methods using the PCA technique. Once more, the MLM-NN spent less time to process a sample in test while holding the accuracy in relation the original MLM. Moreover, the MLP achieved an average increase of 10%. The highest values obtained was 99.19% with the MLM-NN using 100% of training data for reference points and 99.02% with SVM using RBF kernel. As can be seen in the Table 3, this technique decreased the training and test times while remained the methods accuracies.

We can analyze the results achieved in the data set with the Var-PCA technique in the Table 2 (Data set with Var-PCA). The main advantage of this technique is the decrease of the training and test times. However, there was a discrete decline in accuracy in all methods. The highest values obtained was 98.41% with the MLM using Mahalanobis distance, 98.19% with MLM-NN using Euclidean distance, both with 100% of training data for reference points and 97.41% with SVM using RBF kernel.

The fastest method to train was the kNN classifier and to test was the MLM-NN and MLP. MLM and SVM had the highest accuracy among the evaluated methods in comparison to MC-SVM and MC-HF-SVM proposed by [2]. The MLM-NN is able to speed up the original MLM while holding equivalent accuracy. Concerning the feature selection method, we verify that the proposed method is a good approach for embedded applications. Table 2 (Original Data set) summarizes the results obtained by each method considering the settings which reached accuracy greater than 60% in the data set with 561 attributes. All MLMs and MLM-NNs obtained accuracy above 98%, but we highlight the two higher. The MLM-NN which took less time to process a sample test was compared with the MLM using the same setting.

5 Conclusion

In this paper we provided an evaluation of consolidated machine learning techniques focusing human activity recognition tasks. In addition, we proposed a new approach of feature selection along with the PCA technique.

The objectives were reached, since the proposed feature selection approach makes the final application faster while maintaining the accuracy of classifiers. This characteristic is achieved due the feature selection significantly reduces the dimensionality of problem. Our approach reduces the time in acquisition and classification step, then the final application, which consists exclusively these two steps, become much faster in general. Furthermore, MLM-NN presented promising results, especially considering the increased speed in the test case.

As suggestions for a future study there are others classifiers such as the Optimum-Path Forest (OPF) [22,25]. Another suggestion is the use other distance metrics in MLM and MLM-NN.

Acknowledgments. The first author acknowledge the sponsorship from the Cearense Foundation for the Support of Scientific and Technological Development (FUNCAP) and the National Council for Research and Development (CNPq) by providing financial support.

References

1. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Mach. Learn. **6**(1), 37–66 (1991)
2. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In: Bravo, J., Hervás, R., Rodríguez, M. (eds.) IWAAL 2012. LNCS, vol. 7657, pp. 216–223. Springer, Heidelberg (2012). doi:10.1007/978-3-642-35395-6_30
3. Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., Amirat, Y.: Physical human activity recognition using wearable sensors. Sensors **15**(12), 29858 (2015)
4. de Azevedo, F.M., Brasil, L.M., de Oliveira, R.C.L.: Redes Neurais com aplições em Controle e em Sistemas Especialistas. Visual Books (2000)
5. Badawi, H., Saddik, A.E.: Towards a context-aware biofeedback activity recommendation mobile application for healthy lifestyle. Procedia Comput. Sci. **21**, 382–389 (2013)
6. Bayat, A., Pomplun, M., Tran, D.A.: A study on human activity recognition using accelerometer data from smartphones. Procedia Comput. Sci. **34**, 450–457 (2014)
7. Chen, K.H., Chen, P.C., Liu, K.C., Chan, C.T.: Wearable sensor-based rehabilitation exercise assessment for knee osteoarthritis. Sensors **15**(2), 4193 (2015)
8. Crammer, K., Singer, Y., Cristianini, N., Shawe-taylor, J., Williamson, B.: On the algorithmic implementation of multiclass kernel-based vector machines. J. Mach. Learn. Res. **2**, 265–292 (2001)
9. Czabke, A., Marsch, S., Lueth, T.C.: Accelerometer based real-time activity analysis on a microcontroller. In: 5th International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth 2011, Dublin, Ireland, 23–26, May 2011 (2011)
10. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. J. Artif. Int. Res. **2**(1), 263–286 (1995)
11. Duan, K.-B., Keerthi, S.S.: Which is the best multiclass SVM method? An empirical study. In: Oza, N.C., Polikar, R., Kittler, J., Roli, F. (eds.) MCS 2005. LNCS, vol. 3541, pp. 278–285. Springer, Heidelberg (2005). doi:10.1007/11494683_28
12. Dubois, A., Charpillet, F.: Human activities recognition with RGB-depth camera using HMM. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4666–4669, July 2013
13. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley-Interscience, Hoboken (2000)
14. Haykin, S.O.: Neural Networks and Learning Machines. Pearson Prentice Hall, Upper Saddle River (2008)
15. Ho, E.S., Chan, J.C., Chan, D.C., Shum, H.P., Cheung, Y.M., Yuen, P.C.: Improving posture classification accuracy for depth sensor-based human activity monitoring in smart environments. Comput. Vis. Image Underst. **148**, 97–110 (2016)
16. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**, 417 (1933)
17. Jalal, A., Kamal, S., Kim, D.: A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. Sensors **14**(7), 11735 (2014)
18. Lee, M.W., Khan, A.M., Kim, T.S.: A single tri-axial accelerometer-based real-time personal life log system capable of human activity recognition and exercise information generation. Pers. Ubiquit. Comput. **15**(8), 887–898 (2011)

19. Li, Q., Stankovic, J.A., Hanson, M.A., Barth, A.T., Lach, J., Zhou, G.: Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information. In: Proceedings of the 2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks, pp. 138–143. IEEE Computer Society, Washington, DC (2009)
20. Mesquita, D.P.P., Gomes, J.P.P., Junior, A.H.S.: Ensemble of minimal learning machines for pattern classification. In: Rojas, I., Joya, G., Catala, A. (eds.) IWANN 2015. LNCS, vol. 9095, pp. 142–152. Springer, Heidelberg (2015). doi:10.1007/978-3-319-19222-2_12
21. Mohammadi, E., Khoshab, H., Kazemnejad, A.: Activities of daily living for patients with chronic heart failure: a partnership care model evaluation. Appl. Nurs. Res. **30**, 261–267 (2016)
22. Nunes, T.M., Coelho, A.L., Lima, C.A., Papa, J.P., de Albuquerque, V.H.C.: EEG signal classification for epilepsy diagnosis via optimum path forest a systematic assessment. Neurocomputing **136**, 103–123 (2014)
23. Olivares, A., Ramrez, J., Grriz, J.M., Olivares, G., Damas, M.: Detection of (in)activity periods in human body motion using inertial sensors: a comparative study. Sensors **12**(5), 5791 (2012)
24. Osmani, V., Balasubramaniam, S., Botvich, D.: Human activity recognition in pervasive health-care: supporting efficient remote collaboration. J. Netw. Comput. Appl. **31**(4), 628–655 (2008)
25. Papa, J.P., Falco, A.X., de Albuquerque, V.H.C., Tavares, J.M.R.: Efficient supervised optimum-path forest classification for large datasets. Pattern Recogn. **45**(1), 512–520 (2012)
26. Platt, J.C., Cristianini, N., Shawe-Taylor, J.: Large margin dags for multiclass classification. In: Advances in Neural Information Processing Systems, pp. 547–553. MIT Press (2000)
27. del Rosario, M.B., Redmond, S.J., Lovell, N.H.: Tracking the evolution of smartphone sensing for monitoring human movement. Sensors **15**(8), 18901 (2015)
28. Ruck, D.W., Rogers, S.K., Kabrisky, M., Oxley, M.E., Suter, B.W.: The multilayer perceptron as an approximation to a bayes optimal discriminant function. Trans. Neur. Netw. **1**, 296–298 (1990)
29. Shimonski, R., Zenir, J., Bishop, A.: Chapter 4 - mobile phone tracking. In: Bishop, R.S.Z. (ed.) Cyber Reconnaissance, Surveillance and Defense, pp. 113–143. Syngress, Boston (2015)
30. Shoaib, M., Bosch, S., Incel, O.D., Scholten, H., Havinga, P.J.M.: Complex human activity recognition using smartphone and wrist-worn motion sensors. Sensors **16**(4), 426 (2016)
31. de Souza Júnior, A.H., Corona, F., Barreto, G.A., Miche, Y., Lendasse, A.: Minimal learning machine: a novel supervised distance-based approach for regression and classification. Neurocomputing **164**, 34–44 (2015)
32. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Academic Press, Cambridge (2008)
33. Vapnik, V.N.: Statistical Learning Theory. Wiley, Hoboken (1998)
34. Wang, Y., Jiang, X., Cao, R., Wang, X.: Robust indoor human activity recognition using wireless signals. Sensors **15**(7), 17195 (2015)

Lung Segmentation in Chest Computerized Tomography Images Using the Border Following Algorithm

Murillo B. Rodrigues, Leandro B. Marinho, Raul Victor M. Nóbrega, João Wellington M. Souza, and Pedro Pedrosa Rebouças Filho(✉)

Programa de Pós-Graduação em Ciência da Computação, Instituto Federal do Ceará, Maracanaú, CE, Brazil
{murillobarata,raulmedeiros}@lapisco.ifce.edu.br, leandro.marinho@ppgcc.ifce.edu.br, {wellmendes,pedrosarf}@ifce.edu.br

Abstract. This paper proposes a new method of lung segmentation in chest Computerized Tomography (CT) images called Follower of Lung Contour (FLC). This method works as follows: firstly, the image pixels are classified as pulmonary or not through an Artificial Neural Network (ANN) Multilayer Perceptron (MLP) based on pulmonary radiologic densities. After this, the lung detection is made based on achieved through the Border Following Algorithm together with predetermined rules that consider the detected objects area and positioning on the image. The proposed method validation is performed considering as Gold Standard a manual segmentation realized by a pulmonologist at Walter Cantídio Hospital of Federal University of Ceará. Moreover, 30 chest CT images were used, in which 10 are from patients diagnosed with Fibrosis, 10 are from patients with Chronic Obstructive Pulmonary Disease (COPD) and 10 are from healthy patients. The FLC results are compared with six other segmentation methods results using the Gold Standard as reference. Thus, the FLC algorithm shows good results with an average accuracy of 98% and average harmonic means of 98%. Furthermore, it can be concluded that this method may be part of a system to aid in medical diagnosis on Pulmonology.

Keywords: Chest computerized tomography · Border following algorithm · Multilayer perceptron · Lung segmentation · Pulmonary densities analysis

1 Introduction

World Health Organization (WHO) estimates that there are 300 million people suffering from asthma and causes about of 250 thousand deaths per year worldwide [1]. Recent studies reveal that Chronic Obstructive Pulmonary Diseases

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_53

(COPD) are present in individuals aged 20 to 60 years old, although it is more frequent among adults over 50 years. In this sense, WHO estimates that COPD is the leading cause of death worldwide [9].

Computational systems are being developed in order to match the medical knowledge to provide aid for the diagnosis of diseases and monitoring of treatments. The use of Digital Image Processing techniques applied on the medical image analysis stands out among the promising areas for this purpose [2].

The development of tools that provide aid to diagnosis of medical specialist is important to prevent, to detect and to accompany diseases on pneumology area, specifically in the chest Computerized Tomography (CT) images. Moreover, lung segmentation techniques on CT images can improve the analysis speed and efficacy, since the manual lung segmentation is a subjective process because involves a large inter-individual variability of pulmonologists [13].

This paper focuses its efforts on, from chest CT images, perform and analyze an automatic segmentation of the lungs, which are a critical stage in lungs diagnosis due to previously mentioned problems. This developed algorithm is called Follower of Lung Contour (FLC).

2 Materials and Methods

FLC aims to reduce lung segmentation errors performed on chest CT images. For this, it uses the pixels classification approach based on the analysis of lung density proposed by Rebouças Filho et al. [11]. Such approach uses an Artificial Neural Network (ANN) Multilayer Perceptron (MLP) in conjunction with the border following algorithm, that is responsible for the detection of contours. Therefore, FLC is not just a new technic but also a new approach for the use of Pulmonary Densities Analysis (PDA) proposed by Rebouças Filho et al. [11].

This section presents the image acquisition process used in this paper, the proposed method operation and an comparative analysis among FLC and other segmentations technics. These methods are: SISDEP [15], THRmod and THRmult with automatic initialization [3], VFC [7,8], Crisp [13] and Adaptive Crisp [10,11,14].

SISDEP is a system for detecting and quantifying of pulmonary emphysema and was developed from digital image processing techniques, gathering tools that are capable of segmenting the lung parenchyma automatically and accurately, including lung section area measurements on high-resolution CT images [4].

THRmod, THRmult, VFC, Crisp and Adaptive Crisp are Active Contour Methods, also referred as Snakes. A snake consists in tracing a initial curve around or inside an object of interest, which deforms as some forces acting on it and moving to the edges of the object [13]. Each point of the curve has an energy that can be defined by the sum of an intern energy with an external energy, that have an specific value for each point [13].

In this paper it was used 30 chest CT images. They are from patients with 3 types of diagnoses: healthy, with Fibrosis or with COPD. These 30 samples were divided in three groups, each one with 10 images separated according to the previously mentioned diagnostics. The specifications of these images are: dimensions

of 512 × 512 pixels and 16 bits of resolution. Another important feature is that this images are stored in the DICOM (Digital Imaging and Communications in Medicine) format.

The images were obtained in partnership with Walter Cantídio Universitary Hospital from Ceará Federal University (UFC). They were assigned by a previous study [5,17] that was evaluated and approved by the Ethics Committee in Research of UFC - COMEPE (Protocol 35/06) and incorporated to the requirements of the Resolution 196/96 of National Health Council regarding research on humans.

Figure 1 presents the proposed approach steps for lung segmentation with the original CT image in Fig. 1(a) and the pulmonary regions in Fig. 1(d). The proposed approach is divided in two steps. The first step uses a MLP network to label the pixels and the second step uses the border following algorithm to identify the lungs through their contours. These two steps are represented by Fig. 1(b) and (c), respectively.

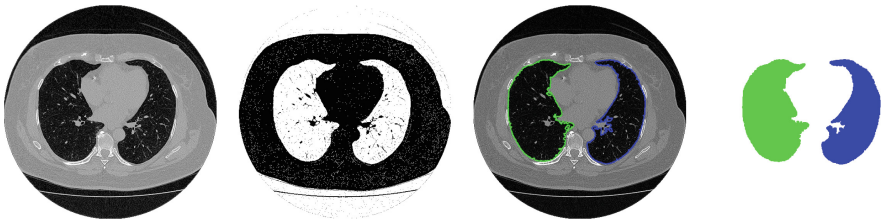


Fig. 1. Lungs segmentation steps of the proposed method: (a) Original image of a chest CT exam; (b) Result of the neural network MLP using the percentage of lung ranges as input; (c) Lung contours detection by the border follow algorithm; (d) Segmentation result by the FLC; left lung is highlighted in green and right lung is highlighted in blue.

2.1 First Step from the Proposed Approach: Classification of Regions

The first step of the proposed approach is based on Rebouças Filho et al. [11] work and makes a density ranges analysis on the lungs. These ranges are: hyper-aerated, normally aerated, little aerated, non aerated and bone.

The first contribution from this paper is a new application of this lung density analysis aiming to improve the process of lung segmentation on chest CT images.

In this paper, the classification of pixels is used to find the lungs. This is realized considering a matrix of lung densities D, in which (x, y) are the coordinates whose 9×9 neighborhood is analyzed. Thereby, the function that determines the number of points with densities present in each class u_i is [11]:

$$f(x, y, u_i) = \sum_{l=-4}^{4} \sum_{m=-4}^{4} R(x - l, y - m), \tag{1}$$

where,

$$R(x,y) = \begin{cases} 1, lim_{inf}(u_i) \leq D(x,y) < lim_{sup}(u_i), \\ 0, \text{otherwise}, \end{cases} \quad (2)$$

where $lim_{inf}(u_i)$ and $lim_{sup}(u_i)$ are the inferior limits and upper limits of the density range in Hounsfield Unit (HU) for each class u_i: u_0 - hyperaerated (-1000 a -950 HU), u_1 - normally aerated (-950 a -500 HU), u_2 - little aerated (-500 a -100 HU), u_3 - non aerated (-100 a 100 HU), u_4 - bone (600 a 2000 HU) and u_5 - not classified areas (densities that do not fit into the others classes).

From these data is possible calculate the percentage P_i from each class i by [12]:

$$P_i = \frac{f(x,y,u_i)}{\sum_{j=0}^{4} f(x,y,u_i)}. \quad (3)$$

The proposed approach uses an ANN MLP to classify the pixels. The inputs of the MLP are six percentages of classes u_i obtained by PDA using the Eq. 1. This MLP has an only output that indicates whether the found pixel on chest CT images are (or not) part the pulmonary region.

In the proposed approach, only one hidden layer is used in the MLP topology, which is responsible for solving nonlinear problems. In this sense, to determine the number of neurons in the hidden layer the HN uses the solution described by [18]:

$$HN = \frac{2 \times max\,(M_1, M_2)}{3}, \quad (4)$$

where M_1 and M_2 are the number of neurons in the input and output layers, respectively.

The result of the Eq. 4 is 4 where M_1 is the 6 u_i classes percentages and M_2 is the only output. This output indicates if the border is pulmonary or not. Thus, this MLP has a topology 6/4/1, in other words, 6 neurons in the input layer, 4 neurons in the hidden layer and 1 neuron in the output layer. The specified topology is evaluated by the cross-validation method [11].

The dataset used in training is extracted from chest CT images of patients diagnosed as healthy, with fibrosis or with COPD, in the apex, hilum and base positions [11].

For the training step, the data are manually selected, aiming to achieve the greater possible representation of the lung structures. On average, 150 pixels are added per image to indicate if the pixels are pulmonary or not. In this set, it can be found borders of bones, vessels, airways, air and others non lung structures. 500 images from different exams, diseases and cut levels are used in this step [11].

Each input set data is obtained from a pixel $P(x_p, y_p)$ where x_p and y_p are the pixel coordinates used in the function $f(x_p, y_p, s_i)$ of the Eq. 1 to obtain all u_i percentages used as MLP input.

The MLP Neural Network output, previously trained, in its execution step is given by:

$$f_{mlp}(\mathbf{u}) = \begin{cases} 1, \text{ when the pixel is pulmonary,} \\ 0, \text{ otherwise.} \end{cases} \quad (5)$$

An example of Eq. 5 is presented in the Fig. 1(b), in which the pulmonary region has output "1" and the non pulmonary region has output "0". Noteworthy that the regions are colored as white and black, respectively. In this way, a new PDA application, proposed by Rebouças Filho et al. [11], that is used to class the pixels as lungs or not.

2.2 Second Step of Proposed Approach: Identification of the Lung Contours

The Border Following algorithm is applied after determine the pixels labels. Suzuki and Abe [16] proposes the border following algorithm and Gonzalez and Woods [6] explains as follows:

1. Let the starting point, b_0, be the uppermost, leftmost point in the image that is labeled 1. Denote by c_0 the west neighbour of b_0. Clearly, c_0 always is a background point. Examine the 8-neighbour of b_0, starting at c_0 and proceeding in a clockwise direction. Let b_1 denote the first neighbour encountered whose values is 1, and let c_1 be the (background) point immediately preceding b_1 in the sequence. Store the locations of b_0 and b_1 for use in Step 5.
2. Let $b = b_1$ and $c = c_1$.
3. Let the 8-neighbour of b, starting at c and proceeding in a clockwise direction, be denoted by $n_1, n_2, \cdots, n_8$. Find the first n_k labeled 1.
4. Let $b = n_k$ and $c = n_k - 1$.
5. Repeat Steps 3 and 4 until $b = b_0$ and the next boundary point found is b_1. The sequence of b points found when the algorithm stops constitutes the set of ordered boundary points.

The result of segmentation obtained by FLC can be seen in Fig. 1(d).

After the border following algorithm has been applied in CT binary images, several contours are found, including lung outlines. Thus, some rules are applied to distinguish lung contours among other detected contours.

Rebouças Filho et al. [13] concludes that, after several tests, a chest CT image can be split into upper (U), bottom (B), left (L) and right (R) regions, as can be seen in Fig. 2.

Considering this division in chest CT image, the centroid of all detected contours are analyzed and assigned to the right or left lung. For a contour be classified as "Left Lung", its centroid should be located on the left side of the image. Similarly, to classify a contour as "Right lung", its central coordinate must be located on the right side of the image. If more than one contour has its centroid region located on the left or right region, the contour with the largest area is regarded as lung, whereas the lungs are the main objects in right and left regions.

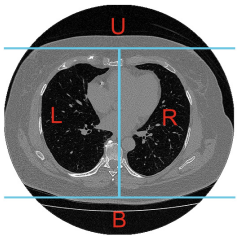


Fig. 2. Chest CT image divided in four regions: Upper (U), Left (L), Right (R) and Bottom (B).

The boundaries of each region used in this study were the same used for Rebouças Filho et al. [13], which are: 0–256 for the left region and 256–512 for the right region on the horizontal axis. In the vertical axis, either right and left regions use 180–360 as boundaries.

After the lung contours have been detected, all pixels that are outside of these contours are labeled as non-pulmonary and remaining labeled as lungs.

3 Results and Discussion

This section presents the results obtained for each method when compared to the Gold Standard. Furthermore, the results are discussed and the proposed method is evaluated.

3.1 Experimental Results

In this step, the proposed method is compared with SISDEP, VFC, THRmult, THRmod, Crisp and Adaptive Crisp using a Gold Standard, which is a result of manual lung segmentation achieved by a pulmonologist at Walter Cantídio Hospital of Federal University of Ceará. These analyses are realized based on the values of accuracy (ACC) and harmonic means (HM), given by Eqs. 6 and 7, respectively.

$$ACC = (TP + TN)/(TP + TN + FP + FN), \tag{6}$$

$$HM = (2 \cdot TPR \cdot SPC)/(TPR + SPC), \tag{7}$$

where $TPR = TP/(TP + FN)$ and $SPC = TN/(FP + TN)$. TP, TN, FP and FN are the respective values of true positive, true negative, false positive and false negative, which are obtained based in a comparative analysis with the ground truth obtained by a specialist. SPC is the value of specificity and TPR is the value of sensitivity.

Values of true positive are obtained when a method and the gold standard classify a pixel as pulmonary and a true negative is obtained when both classify a pixel as not-pulmonary. Moreover, a false positive is when a method labels a pixel as pulmonary and the gold standard indicates that the same pixel is not

pulmonary. Furthermore, the inverse case is called a false negative that is when a method classifies a pixel as not-pulmonary and the gold standard indicates that this pixel is pulmonary.

The values of ACC represent the capacity of a method in correctly solve a problem. The sensitivity of a method is the success percentage in classify the pulmonary region. In the other hand, the specificity is the success percentage in classify the non-pulmonary region. Furthermore, the harmonic means represent how balanced specificity and sensitivity are.

Some of the methods and gold standard results can be seen at Fig. 3.

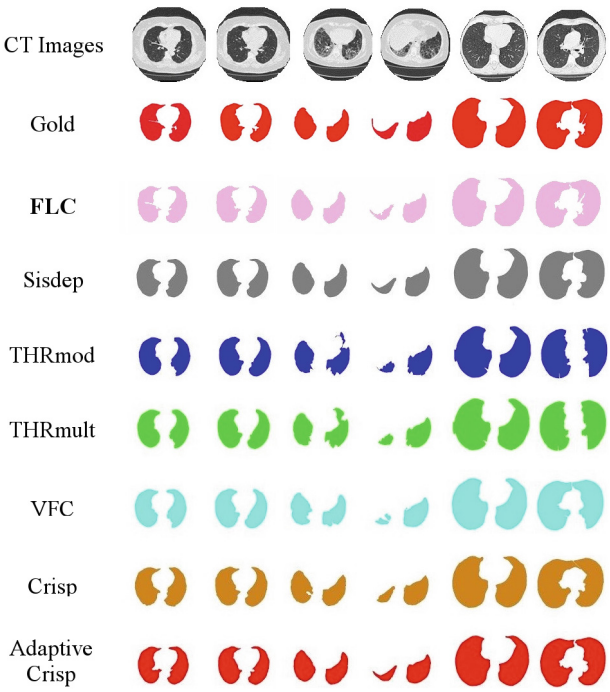


Fig. 3. CT images of the original chest and application results of their segmentation methods on them. From left to right on CT images, there are two healthy lungs images, two lung fibrosis images and two lung images with COPD. Below these CT images, there are the lungs segmentation results by each of the methods listed on the left. Furthermore, it is also shown the Gold Standard in red.

Table 1 contains the values of ACC and HM of FLC, Adaptive Crisp, Crisp, SISDEP, THRmod, THRmult and VFC for each one of the 30 CT images in the following order: COPD (1 to 10), Healthy (11 to 20) and Fibrosis (21 to 30). Moreover, the values of avarage and standard deviation for each measure in each method is also shown at Table 1. In order to improve the analysis of this table, the highest values are highlighted in bold.

The method that stands out in terms of ACC is the proposed method FLC and it is followed by Adaptive Crisp and SISDEP. The FLC is also the best at

Table 1. Table with statistical measures values of each method in each one of 30 images.

Image number	FLC		ACACM		Crisp		Sisdep		THRmod		THRmult		VFC	
	ACC (%)	HM (%)	ACC (%)	HM (%)	ACC (%)	HM (%)	ACC (%)	HM (%)	ACC (%)	HM (%)	ACC (%)	HM (%)	ACC (%)	HM (%)
1	98.06	98.28	98.21	98.32	96.98	97.47	98.27	98.39	95.41	95.56	92.51	90.92	96.10	96.43
2	97.89	98.20	98.00	98.24	96.09	96.64	98.06	98.31	91.03	89.79	92.69	91.77	91.96	91.25
3	99.33	99.63	99.41	99.53	98.79	99.34	99.44	99.31	98.91	99.41	98.85	99.38	98.86	99.38
4	98.60	99.00	98.70	99.00	96.51	95.50	98.78	99.03	97.80	98.01	94.21	86.52	97.75	97.70
5	98.66	99.06	98.74	99.03	96.94	96.92	98.32	98.87	95.97	94.46	96.75	95.26	96.67	96.88
6	99.41	99.47	99.38	99.23	98.37	99.10	99.41	99.11	99.02	99.45	99.02	99.45	98.97	99.37
7	98.61	98.23	98.47	97.95	97.99	97.83	98.33	97.27	95.69	93.19	93.13	86.67	97.99	97.07
8	98.05	98.64	98.28	98.78	95.71	95.12	97.16	95.93	96.50	96.53	91.92	85.18	97.01	97.46
9	98.03	98.49	98.15	98.57	96.71	97.45	98.29	98.67	96.47	97.20	87.37	80.16	96.86	97.55
10	98.84	98.47	98.50	98.07	97.80	97.66	98.36	98.07	93.83	91.50	92.36	88.06	96.59	95.98
11	99.47	99.45	99.56	99.60	98.50	99.04	99.00	99.39	98.85	99.35	98.94	99.32	98.84	99.35
12	99.40	99.49	99.44	99.48	98.31	99.00	99.15	99.46	98.79	99.28	98.80	99.29	97.97	98.81
13	99.32	99.55	99.41	99.52	98.53	99.12	99.33	99.57	98.61	99.18	98.65	99.16	98.59	99.17
14	98.94	99.10	98.89	99.10	97.50	98.35	98.58	99.12	97.98	98.80	98.00	98.81	97.97	98.79
15	99.31	98.03	99.05	97.73	98.95	99.03	99.22	98.45	97.36	93.62	98.78	97.95	99.20	99.23
16	99.11	97.58	97.41	94.25	97.80	95.45	97.41	94.63	97.03	94.23	97.06	94.25	97.79	96.57
17	96.94	93.05	97.56	95.39	98.00	97.21	97.59	95.80	97.46	94.88	96.04	90.26	97.61	96.19
18	97.23	94.85	97.17	94.70	97.44	96.70	97.24	95.40	97.17	94.61	97.01	94.06	97.29	95.91
19	99.01	99.27	97.60	96.49	97.44	98.04	97.64	97.48	96.95	95.14	96.91	94.11	96.88	95.21
20	99.08	99.29	99.03	99.29	97.54	97.90	97.83	97.35	97.78	97.75	97.29	95.48	97.73	97.65
21	99.45	99.59	99.54	99.31	99.14	99.54	99.65	99.30	99.26	99.57	97.78	98.81	99.29	99.55
22	99.48	99.57	99.47	99.25	99.07	99.48	99.52	98.63	99.23	99.45	99.23	99.40	99.15	99.35
23	99.35	99.53	99.34	99.38	98.96	99.41	99.32	97.95	98.83	98.12	98.74	98.07	98.89	98.00
24	99.23	99.41	99.33	99.36	98.79	99.14	99.26	99.39	98.51	96.53	98.28	95.14	98.55	96.51
25	98.88	99.31	99.21	99.26	98.50	98.17	98.88	98.90	98.25	95.64	97.38	91.43	97.01	88.80
26	99.09	99.42	99.25	99.30	98.77	99.00	99.31	99.09	98.61	97.32	96.79	87.64	98.15	94.54
27	99.04	99.31	99.13	99.11	98.49	98.88	99.12	99.11	98.37	96.95	95.69	83.07	98.67	97.84
28	98.51	98.84	98.53	98.47	97.89	96.05	98.57	98.89	97.04	92.85	95.29	81.28	96.96	91.10
29	98.69	99.12	98.75	98.68	98.12	97.83	98.79	99.11	97.39	93.42	96.57	88.77	97.80	94.72
30	98.82	99.27	99.14	99.22	98.29	97.78	99.10	99.36	96.89	93.07	94.92	80.44	97.95	95.31
Average	**98.8**	**98.69**	98.76	98.46	97.94	97.94	98.64	98.32	97.37	96.17	96.24	92.34	97.71	96.73
SD	**0.65**	**1.42**	0.69	1.42	0.90	1.29	0.73	1.32	1.75	2.70	2.79	6.31	1.39	2.64

HM rate with 98.69%. The proposed method stands out for being practically constant in all the cases examined (COPD, Healthy and Fibrosis) and achieving results near of specialists methods.

The best method in the segmentation of COPD lungs in accuracy terms was Adaptive Crisp with 98.59% and the second was the FLC with 98.55%. In the healthy lungs, the method that stands out was the FLC, with accuracy of 98.78%. For the case of Fibrosis lungs, the proposed method was third with 99.05% of ACC and SISDEP was first with 99.15%. Thus, it is noteworthy that where FLC is not the best, its accuracy values are still high.

A comparison considering the average and the standard deviation of each rate for each method is performed for a more general analysis. Therefore, it is

also noteworthy that the proposed approach is very effective, since it obtained the highest average of ACC and HM and it is followed for Adaptive Crisp. In terms of standard deviation, Crisp and Adaptive Crisp distinguished themselves and FLC had the lowest for accuracy.

SISDEP stands out for obtain good accuracy rates in some specific images as is shown at Table 1. However, FLC and Adaptive Crisp perform a more general analysis.

According to the metrics used it can be concluded that the proposed approach owns the higher average of ACC and HM indicates that FLC have a general good rate of success and a good balance between affirm what is lung region or not, respectively. Thus, it can conclude that FLC has a good precision.

4 Conclusion

FLC presents a new approach for the Pulmonary Densities Analysis developed by Rebouças Filho et al. [11]. Thus, comparative analysis of the results, it can be noticed that Adaptive Crisp and FLC are very effectives methods and both have similar results, but the proposed approach has a greater simplicity when compared to Rebouças Filho et al. [11] approach, which performs calculations of internal and external energies for the lung contours detection.

FLC becomes efficient due the results of pixels classification through the PDA with the border following algorithm. PDA can classify the pixels more accurately because it considers the pixel neighborhood in analysis and not only the lung radiodensity of this pixel. Therefore, noise pixels can be correctly classified.

Finally, it can conclude that FLC is an efficient approach for lung segmentation in chest CT images, because it presents good results when compared with others effective methods, besides propose a new use for PDA [11].

Future work will apply the proposed method in a larger database with a gold standard respective, statistical validation in the results and three-dimensional display of the pulmonary region are considered.

Acknowledgments. The third author acknowledge the sponsorship from the Cearense Foundation for the Support of Scientific and Technological Development (FUNCAP) and the National Council for Research and Development (CNPq) by providing financial support.

References

1. Campos, H., Lemos, A.: Asthma and COPD in view of the pulmonologist. Braz. J. Pulmonol. **35**(4), 301–309 (2009)
2. Cavalcante, T.S.: Technique for automatic segmentation of airway in computed tomography images of the chest. Master's thesis, Department of Computer Engineering, Federal University of Ceará, Fortaleza, Ceará, Brazil (2010)
3. Felix, J., Cortez, P., Cavalcante, T., Alexandria, A., Holanda, M.: Autoin: method of automatic initialization of active contours applied to lungs in CT images. IEEE Latin Am. Trans. (Revista IEEE Am. Lat.) **10**(4), 1954–1960 (2012)

4. Felix, J.H.S., Cortez, P.C., Costa, R.C.S., Fortaleza, S.C.B., Pereira, E.D.B., Holanda, M.A.: Avaliação computacional de enfisema pulmonar em tc: comparação entre um sistema desenvolvido localmente e um sistema de uso livre. J. Bras. de Pneumol. **35**(9), 868–876 (2009)
5. Fortaleza, S.C.B.: Efeitos da Administração de Pressão Positiva Contínua em Vias Aéreas de Modo não Invasivo Sobre a Aeração do Parênquima Pulmonar em Pacientes com Doença Pulmonar Obstrutiva Crônica. Master's thesis, Department of Clinical Medicine, Federal University of Ceará, Fortaleza, CE (2006)
6. Gonzalez, R.C., Woods, R.: Digital Image Processing, 3a edn. Pearson Prentice Hall, New Jersey (2010)
7. Li, B., Acton, S.: Vector field convolution for image segmentation using snakes. In: IEEE International Conference on Image Processing, pp. 1637–1640. IEEE Image Processing, Atlanta, EUA, Outubro 2006
8. Li, B., Acton, S.: Active contour external force using vector field convolution for image segmentation. IEEE Trans. Image Process. **16**, 2096–2106 (2007)
9. de Marco, R., Accordini, S., Cerveri, I., Corsico, A., Sunyer, J., Neukirch, F., Kunzly, N., Leynaert, B., Janson, C., Gislason, T., Vermeire, P., Svanes, C., Anto, J., Burney, P.: An international survey of chronic obstrutive pulmonary disease in young adults according to gold stages. Thorax **59**, 120–125 (2004)
10. Ramalho, G.L.B., Reboucas Filho, P.P., Medeiros, F.N.S.D., Cortez, P.C.: Lung disease detection using feature extraction and extreme learning machine. Revista Brasileira de Engenharia Biomed
11. Rebouças Filho, P.P., Cortez, P.C., Felix, J.H.S., Cavalcante, T.D.S., Holanda, M.A.: Adaptive 2D crisp active contour model applied to lung segmentation in CT images of the thorax of healthy volunteers and patients with pulmonary emphysema. Rev. Bras. de Engenharia Bioméd. **29**(4), 363–376 (2013)
12. Rebouças Filho, P.P.: MCA crisp: nova técnica de segmentação dos pulmões em imagens de Tomografia Computadorizada baseada no Metodo de Contorno Ativo. Master's thesis, Universidade Federal do Ceará, Fortaleza, CE, Março 2010
13. Rebouças Filho, P.P., Cortez, P.C., Holanda, M.A.: Active contour models crisp: new technique for segmentation the lungs in CT images. Revista Brasileira de Engenharia Biomed. **27**, 259–272 (2011)
14. Rebouças Filho, P.P., Cortez, P.C., da Silva Barros, A.C., de Albuquerque, V.H.C.: Novel adaptive balloon active contour method based on internal force for image segmentation - a systematic evaluation on synthetic and real images. Expert Syst. Appl. **41**(17), 7707–7721 (2014)
15. Silva Felix, J.H., Cortez, P.C., Filho, P.P.R., Alexandria, A.R., Costa, R.C.S., Holanda, M.A.: Identification and quantification of pulmonary emphysema through pseudocolors. In: Gelbukh, A., Morales, E.F. (eds.) MICAI 2008. LNCS (LNAI), vol. 5317, pp. 957–964. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88636-5_90
16. Suzuki, S., Abe, K.: Topological structural analysis of digitized binary images by border following. Comput. Vis. Graph. Image Process. **30**(1), 32–46 (1985)
17. Holanda, M.A., Fortaleza, S.C.B., Alves-de-Almeida, M., Winkeler, G.F.P., Reis, R.C., Felix, J.H.S., Lima, J.W.O., Pereira, E.D.B.: Continuous positive airway pressure effects on regional lung aeration in patients with COPD: a high-resolution CT scan study. Chest (American College of Chest Physicians) **138**(2), 305–314 (2010)
18. Yin, X., Liu, C., Han, Z.: Feature combination using boosting. Pattern Recogn. Lett. **25**(14), 2195–2205 (2005)

Learner's Profile Hierarchization in an Interoperable Education System

Leila Ghorbel[1(✉)], Corinne Amel Zayani[1], Ikram Amous[1], and Florence Sèdes[2]

[1] MIRACL-ISIMS Sfax University, Tunis Road Km 10, 3021 Sfax, Tunisia
leila.ghorbel@gmail.com, {corinne.zayani,ikram.amous}@isecs.rnu.tn
[2] Department of Computer Science, Paul Sabatier University, IRIT Laboratory, Toulouse, France
sedes@irit.fr

Abstract. In recent years, several education systems have been developed. Consequently, each learner can have different profiles which each one is related to a system. Each profile can be completed and enriched by the data coming from the other profiles in order to return results reflecting the learner's need. The profile enrichment requires the establishment of an interoperable system which (i) resolves the problem of learner's profile heterogeneity based on a matching process and (ii) integrates the data in the different profiles based on a data fusion process. The data fusion approaches mainly aim at resolving the conflicts occurring in the data values. They are based on non organized profiles which may produce inconsistent results. The profile organization is done either by using the machine learning techniques or the notion of temperature. In this paper, we propose a new data fusion approach to improve the conflict resolution by organized profiles. Each profile is organized by respectively merging a clustering algorithm and the temperature and by taking into account the data semantic relationship.

Keywords: Learner's profile · Hierarchization · Data fusion · Semantic similarity · Temperature · Clustering

1 Introduction

With the emergence of the educational systems, such as the learning management systems, the learner can have several profiles situated in different systems. Each profile can have incomplete learner's data (interests, preferences, learning styles, etc.). Hence, its data can complement other profiles in order to return to the learner's results that correspond to his need. In order to complete the profiles with data, an interoperable system should be implemented. This system aims, first, at resolving the large number of the profile differences in syntax, structure and semantics by using a matching process. The result of this latter is a set of

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_54

mapping rules that expresses the similarity between a learner's profile element and another element of a further learner's profile. Second, it aims at integrating the data in the different profiles based on the mapping rule set and the data fusion process. This process should resolve the conflict occurring in the data values. Thus, a learner's profile can provide conflicting data (erroneous, out-of-date, incomplete, duplicate ...), which leads to return incorrect data in a query result.

The conflict resolution is still a critical problem in the interoperable systems in different fields (social, education, etc.). It is significant, especially, with the spread of conflicts and irrelevant data over the different profiles. In literature, we find several strategies for conflict resolution [2]. The first is a conflict-ignoring strategy which is not aware of conflicts, performs no resolution, and thus may produce inconsistent results. The second is conflict-avoiding strategy which is aware of conflicts, but does not resolve each conflict. Finally, conflict-resolving strategies provide the means for fusion decisions for each conflict. The decision can be made through two main methods [8]: (i) the majority voting or averaging and (ii) the reliability of sources (profile data provider) and truth discovery. These methods are based on non structured and non organized profiles.

In this context, we propose a novel approach to handle conflicts in the data fusion process related to a learner's profile interoperability system [4]. This approach detects the reliable sources and the true data values from each one. This detection is based on the organized learner's profiles. The profile organization takes into account the concept of temperature [10,12], the k-means algorithm [7] and the semantic similarity measures [16,22].

The rest of this paper is organized as follows. In Sect. 2, we provide a state of the art about the data fusion approaches in the learners' profile interoperability systems. Then, in Sect. 3, we describe our proposed approach. In Sect. 4, we describe the hierarchization process included in the approach. In Sect. 5, we describe the evaluation results. In the last section, conclusion and future studies are disclosed.

2 State of the Art

In the user's profile interoperability systems, we distinguish two data integration ways. In the first, the data collected from the different profiles are not merged (without fusion) in the profile that requires to be completed, but they are used only when needed [17]. In the second way, the collected data are merged (with fusion) into the learner's profiles [11,20]. This approach requires the data integration and the conflict resolution operations. The conflicts can occur in the data or in the data values collected from the learner's profile of the different systems.

The conflict resolution provides the means for fusion decisions for each conflict. The decision is made through two methods which are the majority voting or averaging and the source reliability and truth discovery [8]. The majority voting method consists in merging in the profile the data values with the highest number of occurrences in the other profiles. The biggest shortcoming of this

method is that it assumes all the sources providing the profile data are equally reliable. For this reason, the second method emerged in order to estimate the source reliability degrees and infer true data values. The sources that provide true data values will be assigned higher reliability, and the data supported by reliable sources will be regarded as truth [8]. Several learner's profile interoperability studies adopted the second method. They consider that a source is reliable only if it is frequently used [11] or manually notified [13,20] by the learner. Thus, all the data in the profile of this source are considered as true which may produce inconsistent results. In fact, some data can be out-of-date, irrelevant and duplicated due to the lack of structuration and organization of the profile data.

In literature, we find several studies, in different fields, which organize the user's (learner's) profile in order to reduce its overload. In fact, the user's profile evolves over time and its content can vary between relevant and irrelevant. In this context, we distinguish two main categories of works which vary according to the way of keeping the user's interest in the profile.

The first category uses the concept of temperature in order to organize and keep in the user's profile some interests relative to a specific period of time. In Manzat et al. [10], temperature corresponds to the popularity of the user's interests detected from the consumed resources over a period of time. The authors calculated the weight for each interest. The authors in [12] improve the concept of temperature proposed in [10] by adding the freshness and social similarity. The second category represents three groups of methods that use the machine learning techniques [21] which can be supervised, semi-supervised or unsupervised.

The first group of methods consists in classifying the new interests to be added. These methods are based on the K Nearest Neighbors (KNN) [1,15]. They provide the user's profile with a content that consists only of semantically similar interests. As a consequence, the profile has only very close interests. The second group of methods consists in classifying the already existing interests in order to remove the irrelevant ones. For example, the authors in [24] proposed a new version of the semi-supervised Co-Training algorithm in order to remove the irrelevant interests from the user's profile. These two first methods use the supervised or semi-supervised learning technique depending on the manual labeling in order to generate a prediction model. The third group of methods appears in order to resolve the problem of the manual labeling defined in the two first groups. These methods use the unsupervised learning technique. They consist in dividing the user's interests into clusters and separate the ones containing irrelevant ambiguous and redundant interests [7,23]. For example, the authors in [7] proposed a new method to measure the interest similarity based on the common co-occurrence group. After that, they used the unsupervised K-Means clustering algorithm to aggregate the interests.

From the analyzed studies, it should be noted that the source reliability and truth discovery conflict resolution methods can produce consistent results for the learner compared by the majority voting or averaging method. To our knowledge, the works related to this method are based on non structured and

organized user's profiles which can foster the merge of false data. In addition, most of the studies organize the user's profile by keeping some interests relative to a specific period of time or by deleting the out-of-date interests based on a machine learning algorithm. These studies can provide a user's profile with relevant data (interests) in terms of ambiguity and redundancy. However, the user's data can be semantically dependent from one source to another regardless of time.

For this reason, we propose a novel approach for data fusion with conflict resolution which consists in: (i) organizing the learner's profile data based on the notion of temperature and the machine learning algorithm and (ii) detecting the reliable sources and the true data values.

3 The Data Fusion Approach

Our approach is integrated in our proposed architecture for the learner's profile interoperability [4]. This architecture allows the exchange of the learner's profile data between educational systems in order to: (i) get a uniform access to the learners' profiles which are represented by different standards, such as PAPI and IMS LIP [14] based on a "Mechanism for Mapping Generation" and (ii) integrate the exchanged data with fusion in the target learner's profile based on a "Mechanism of Data Fusion".

Our architecture is composed of four layers: client, adaptation, interoperability and sources. In this paper, we detail the extent of the interoperability layer by

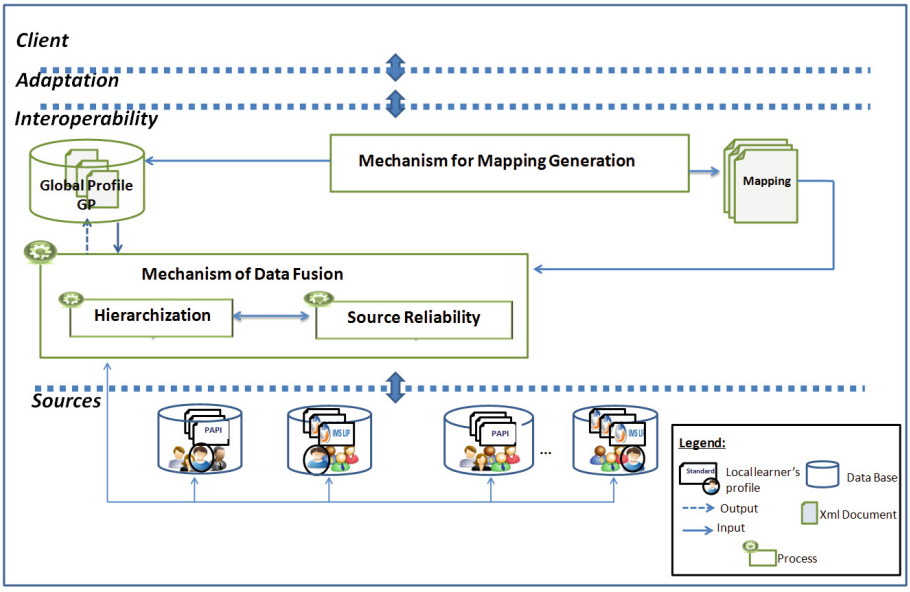


Fig. 1. Interoperability layer extension: mechanism of data fusion

the "Mechanism of Data Fusion". This latter, as shown in Fig. 1, takes as input the generated mapping rules that ease the fusion of data between heterogeneous learner's profile schemas (the "Global Profile" GP and the local profiles).

The "Mechanism of Data Fusion" resolve the generated data value conflicts based on two dependent processes. The first process, which is called "Hierarchization", consists in organizing and structuring each learner's profile by hierarchy. The second process, which is called "Source Reliability", consists in detecting the sources that provide the true data values with the highest reliability weights.

These processes are built on the basis of the hypothesis that a source (the system that is the profile data provider) is reliable for a learner only if his profile (in this source) contains his interest true values. In fact, in a learner's profile, we can find two interest types: (i) local interests which are detected by the system or provided explicitly by the learner and (ii) copied interests which are provided from other profiles of the same learner. The latter ones may be considered as false interests because a false value can be spread through copying, which makes the reliability weight decreases. Moreover, a copied interest may be considered true: (i) with the increase of its temperature values which are the freshness and the score or (ii) if there are other interests that are semantically close to this interest and have better temperature values. For this reason, the reliability process should retrieve the false interests in each source in order to calculate its reliability weight. The retrieval step takes a lot of time in browsing all the interests in each source. Therefore, we propose to organize the learner's interest in each source by hierarchy based on the interest temperature values, the K-means clustering algorithm and the semantic similarity measures. Due to the space limitation, we detail, in the next sections, only the hierarchization process and its evaluation results.

4 The Hierarchization Process

This process consists of three steps. The first step divides the user's interests into clusters based on the semantic similarity measures. We measure the similarity between two interests by combining the Information Content (IC) and Path similarity measures. Therefore, the similarity degree between two interests is calculated using the average of the similarity values between all the IC (res [16], lin [9] and jcn [5]) measure values and all the Path (lch [6], wup [22] and path [18]) measure values. Then, in the second step, each cluster is divided into SubClusters based on the temperature values which are the interest freshness "F" and the scores "Sc". The values of "F" and "Sc" are respectively related to the interest date and frequency. The final step enables to represent the clusters with their respective SubClusters in a tree. The root represents all the interests. The number of node the tree first level is equal to the number of clusters generated in the first step. Each cluster is represented by a sub-tree. The number of levels, in each sub-tree is relative to the number of SubClusters. In the first level, each node is related to a SubCluster which has the highest "F" and "Sc" values. The levels in each sub-tree are related by the most semantically related interests. These steps are summarized in Algorithm 1.

Algorithm 1. Hiearchization:H_Kmeans

Input: Interest[] : String; kc,ksc, nbIteration :integer ; distanceType: String
Output: *hierarchy* : *tree*

```
1: Begin
2: distanceType ← "semantic"
3: Cluster[] ← Kmeans(Interest, kc, nbIteration, distanceType)
4: for i ← 1 to kc do
5:    distanceType ← "temporal"
6:    Cluster[i].SubCluster[] ← Kmeans (Cluster[i].InterestC, ksc, nbIteration, distanceType)
7:    reorder(Cluster[i].SubCluster[])
8:    Hierarchy.level[1].interests ← Cluster[i].SubCluster[1].InterestSC
9:    for j ← 2 to ksc do
10:      for h ← 1 to sizeOf(Cluster[i].SubCluster[j-1].InterestSC) do
11:         distanceType ← "semantic"
12:         AffectInterestsTo(Hierarchy.level[j-1].interests[h], selectInterest(distance
           (Hierarchy.level[j-1].interests[h], Cluster[i].SubCluster[j].InterestSC, distanceType))
13:      end for
14:   end for
15: end for
16: End
```

This algorithm helps classify the interest set into clusters with "Kmeans" based on the semantic similarity measures in order to calculate the interest semantic distance. Then, it classifies each resulted cluster into SubClusters based on "F" and "Sc" values in order to calculate the interest temperature distance. Afterwards, it constructs the interest hierarchy (the tree): in each level, it affects the SubClusters relative to each one of the clusters. The relation between the levels (parents and descendants) is detected based on the semantic distance. Our proposed algorithm takes as input the Interest set "Interest[]", the number of clusters "kc", the number of iterations "nbItetration", and the distance type which can be "semantic" or "temperature". The output of this algorithm is tree that contains the Interest hierarchy constituted by sub-hierarchies (sub-trees). First, the algorithm constructs the "Kc" clusters (lines 2–3) based on the algorithm "Kmeans" with the semantic distance type (Cf. Algorithm 2). Second, for each cluster (lines 4–6), it constructs "ksc" SubClusters with the temperature distance type (line 5). Third, based on these SubClusters (relative to each cluster), it reorders, for each cluster, its SubClusters beginning with the ones that have the highest interest freshness and score values (line 7). Finally, it generates the tree hierarchy by adding the interests relative to each cluster into a sub-hierarchy as follow: the first level is composed of the interests of the first SubCluster (line 8), the other levels are composed of interests relative to the other SubClusters and each interest is affected to its direct parent based on a comparison with the semantic distance type (lines 9–13).

The "Kmeans" algorithm generates, after nbIterations, "k"clusters containing interests that are semantically or temporally very close. First, it randomly selects, from the interest set, "k" initial centroids relative to the clusters (line 2). Second, in each iteration, it affects interests to each cluster that are very close to its centroid based on the distance type (line 5). Then, it selects the new centroid of each cluster (line 6).

Algorithm 2. Kmeans

Input: Interest[] :String; k, nbIteration :integer; distanceType : String
Output: Cluster[]
1: **Begin**
2: $Centroid[] \leftarrow$ SelectInitialKInterset(Interest[],K)
3: **for** $i \leftarrow 1$ **to** $nbIteration$ **do**
4: **for** $j \leftarrow 1$ **to** k **do**
5: $Cluster[j].InterestC[] \leftarrow select(Centroid[j], distance(Centroid[j], Interest,$ distanceType))
6: $Centroid[] \leftarrow$selectNewCentroid$(Cluster[j].InterestC, Centroid)$
7: **end for**
8: **end for**
9: **End**

5 Experimental Results

Our experiment was conducted on real learners who are subscribed in the MOODLE and the Position Platform educational system. The learner's profiles in MOODLE and the Position Platform contain the learner's interests which are explicitly provided by the learners themselves or implicitly detected based on the visited and learned course or lesson domain. Each interest is described by its date, wight and its type.

Therefore, we applied our proposed approach in order to organize each user's interests. As previously described, our approach is based on the Hierarchization H_Kmeans algorithm in order to divide in the two first steps: (1) the user's interests into Kc clusters based on the semantic distance and (2) the interests relative to each generated cluster into Ksc SubClusters based on the temporal distance. For this reason, we choose the Silhouette coefficient [19] and the Dunn Index [3] metrics to evaluate the clustering result: a larger value means better clustering result. After the clustering and SubClustering step, our algorithm creates the learner's interest hierarchy which is composed of sub-hierarchies. We evaluated the latter ones the basis of the human judgment.

The evaluation consists of two phases. The evaluation of the first phase aims at measuring the average of the silhouette and the Dunn Index values for each learner based on the simple K-means (Euclidean distance) and our proposed algorithm (semantic and temperature distance). Figure 2A shows a clear improvement in the silhouette values for all the learners. In fact the average increased from "54%" to "75%". Moreover, the Dunn index values showed a clear improvement (cf. Fig. 2B). We recorded an average of "0.97" for the simple K-Means and "1.09" for our proposed algorithm.

In the second phase, we evaluate the created hierarchy by our algorithm for each learner on the basis of human judgment. Figure 3 shows the evaluation result related to each learner. There are three levels of hierarchy quality: high, medium and low. We notice the percentage for each level, with respect to the total number of the created sub-hierarchy for each learner. For example, for the learner "L1", there are "60%" of sub-hierarchies with high quality, "20%" with medium level and "20%" with low level. Based on these evaluation values, we

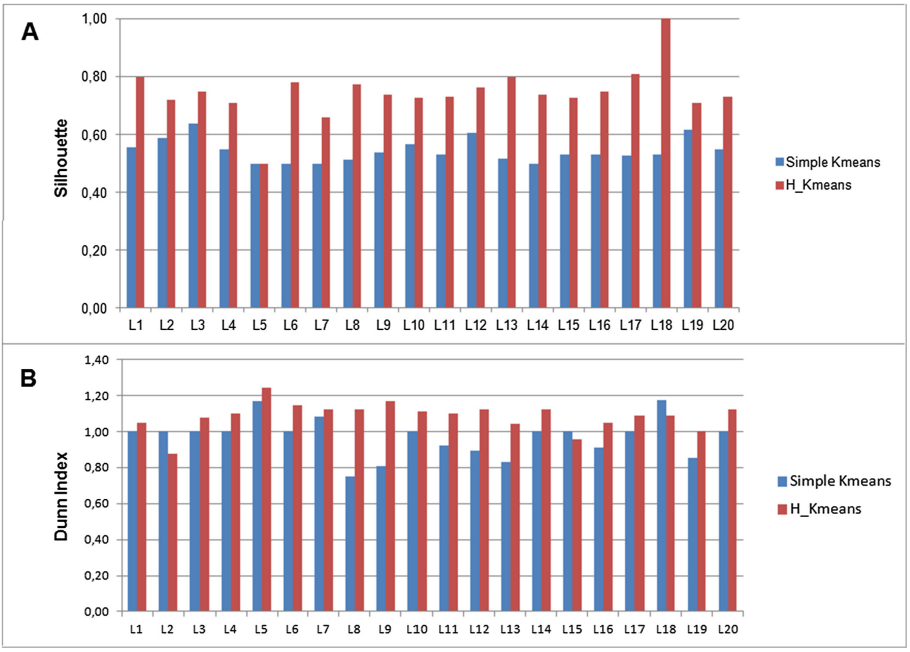


Fig. 2. The silhouette coefficient (A) and the Dunn Index (B) comparison values with the simple k-means and our proposed algorithm

notice the percentage average value, for each level, relative to all the learners: "54%" of the total learners' interest hierarchies have high quality, "18%" have medium quality and "28%" have low quality.

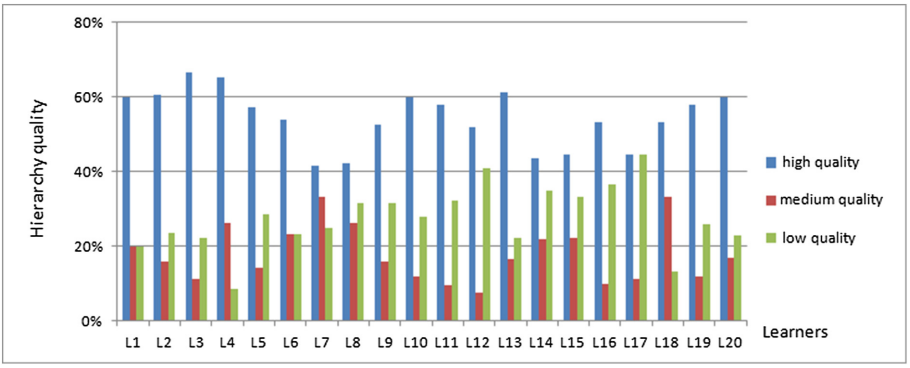


Fig. 3. The learners' interest hierarchy quality

6 Conclusion

In this paper, we proposed the "Mechanism for Data Fusion" which is integrated in our previously proposed architecture [4] for the learner's profile interoperability. This mechanism is constituted of two processes: the Hierarchization and Source Reliability.

We experimented the first process through the MOODLE and Position Platform system. We have calculated the silhouette and the dunn index values in order to evaluate the clustering results. Then, we evaluated the generated hierarchies based on the human judgment. Results show the effectiveness and the performance of our proposition.

In future studies, we will evaluate the "Source Reliability" process based on the generated hierarchies. Then, we will evaluate the overall mechanism in our architecture.

Acknowledgments. We would like to acknowledge with much appreciation the crucial role of the 'PHC Utique' program of the French Ministry of Foreign Affairs and the French Ministry of higher education and research and the Tunisian Ministry of higher education and Scientific Research in the CMCU project number 14g1417 for their financial support.

References

1. Chunyan, L.: User profile for personalized web search. In: Fuzzy Systems and Knowledge Discovery FSKD, pp. 1847–1850 (2011)
2. Dong, X.L., Naumann, F.: Data fusion: resolving data conflicts for integration. Proc. VLDB Endowment **2**(2), 1654–1655 (2009)
3. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. J. Cybern. **4**, 95–104 (1974)
4. Ghorbel, L., Zayani, C.A., Amous, I.: A novel architecture for learner's profiles interoperability. In: ICIS, pp. 405–410 (2015)
5. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of 10th International Conference on Research in Computational Linguistics, ROCLING 1997 (1997)
6. Leacock, C., Chodorow, M.: Combining local context and wordnet sense similarity for word sense identication. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, pp. 265–283. MIT Press, Cambridge (1998)
7. Li, H., Hu, X., Lin, Y., He, W., Pan, J.: A social tag clustering method based on common co-occurrence group similarity. Frontiers IT EE **17**(2), 122–134 (2016)
8. Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., Han, J.: A survey on truth discovery. SIGKDD Explor. Newsl. **17**(2), 1–16 (2016)
9. Lin, D.: Extracting collocations from text corpora. Workshop on Computational Terminology, pp. 57–63 (1998)
10. Manzat, A., Grigoras, R., Sèdes, F.: Towards a user-aware enrichment of multimedia metadata. Workshop on Semantic Multimedia Database Technologies (SMDT), pp. 30–41 (2010)
11. Martinez, M.L., González-Mendoza, M., Valle, I.D.D.: Enrichment of learner profile with ubiquitous user model interoperability. Computacion y Sistemas **18**, 359–374 (2014)

12. Mezghani, M., Zayani, C.A., Amous, I., Péninou, A., Sèdes, F.: Dynamic enrichment of social users' interests. In: RCIS, pp. 1–11 (2014)
13. Musa, D.L., De Oliveira, J.P.M.: Sharing learner information through a web services-based learning architecture. J. Web Eng. **4**(3), 263–278 (2005)
14. Paramythis, A., Loidl-Reisinger, S.: Adaptive learning environments and e-learning standards. Second Eur. Conf. E-Learn. **1**, 369–379 (2003)
15. Pazzani, M.J., Billsus, D.: Content-based recommendation systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web. LNCS, vol. 4321, pp. 325–341. Springer, Heidelberg (2007). doi:10.1007/978-3-540-72079-9_10
16. Resnik, R.: Using information content to evaluate semantic similarity. In: 14th International Joint Conference on Artificial Intelligence, pp. 448–453 (1995)
17. Santos, O.C., Boticario, J.G.: Practical guidelines for designing and evaluating educationally oriented recommendations. Comput. Educ. **81**, 354–374 (2015)
18. Sebti, A., Barfroush, A.A.: A new word sense similarity measure in WordNet. In: IEEE IMCSIT, pp. 369–373 (2008)
19. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining (2006)
20. Walsh, E., O'Connor, A., Wade, V.: The fuse domain-aware approach to user model interoperability. A comparative study. In: 2013 IEEE 14th International Conference on Information Reuse and Integration (IRI), pp. 554–561. IEEE (2013)
21. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques Edition, Techniques edn. Morgan Kaufmann, Burlington (2005)
22. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: 32nd Annual Meeting of the Association for Computational Linguistics, pp. 133–138 (1994)
23. Xu, G.D., Zong, Y., Jin, P.: KIPTC: a kernel information propagation tag clustering algorithm. J. Intell. Inform. Syst. **45**(1), 95–112 (2015)
24. Zghal, R.R., Ghorbel, L., Zayani, A., Amous, I.: An adaptive method for user profile learning. In: Proceedings of 17th East-European Conference on Advances in Databases and Information Systems, Genoa, Italy, pp. 126–134, September 2013

ACO-PSO Optimization for Solving TSP Problem with GPU Acceleration

Olfa Bali[1(✉)], Walid Elloumi[1], Ajith Abraham[2], and Adel M. Alimi[1]

[1] REGIM-Laboratory: Research Groups on Intelligent Machines, National Engineering School of Sfax (ENIS), University of Sfax, BP 1173, 3038 Sfax, Tunisia
bali.olfa@gmail.com, {walid.elloumi, adel.alimi}@ieee.org
[2] Intelligence Research Labs (MIR Labs), P.O. Box 2259, Auburn, WA 98071-2259, USA
ajith.abraham@ieee.org

Abstract. In this paper, we present a novel approach named "ACO-PSO-TSP-GPU" to run PSO and ACO on Graphical Processing Units (GPUs) and applied to TSP (Parallel-PSO&ACO-A-TSP). Both algorithms are implemented on GPUs. Well-known benchmark problems for many heuristic and meta heuristic algorithms presented by Travelling Salesman Problem (TSP) are known as NP hard complex problems.TSP was investigated using classical approaches as well as intelligent techniques employing Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO). Parallel computing is well suited to the execution of nature and bio-inspired algorithms due to the rapidity of parallel implementation. Results show better performance optimization when using parallelism compared to results using sequential CPU implementation.

Keywords: PSO · ACO · TSP · GPU · Optimization · Parallelism

1 Introduction

In the field of Engineering, the optimum solution of a problem is defined by using optimality criteria.

Swarm intelligence techniques are bio-inspired methods, where group behavior is used to solve a problem based on the individualities of its members.

Particle swarm optimization (PSO), is a branch of Swarm intelligence used for solving many engineering optimization problems. Among the stochastic global optimization techniques initially designed for non-linear continuous function optimization, Swarm Intelligence algorithms offer a number of attractive features, global search capability and easy implementation. Since 1995, Kennedy and Eberhart developed a meta-heuristic population based on global optimization called PSO [1], presented in Fig. 1.

PSO suffer from premature convergence for small population size but can be improved by increasing the population. Fortunately, the PSO is very easy to be

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_55

parallelized since the particles do not depend on each other while moving in the search space. Many approaches simulate multiple particles at a time or propose multiple swarm versions of the PSO [2]. The task of calculation can be heavy and the speed of the course PSO can be seen as slow as the operation of the PSO algorithm requires a number of iterations and a stop condition; which are formed in a sequential manner on the processor.

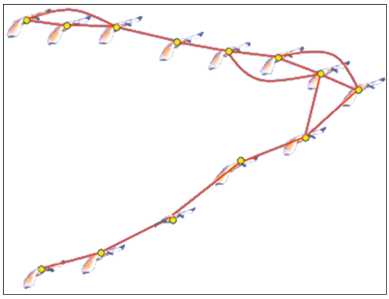


Fig. 1. Bird flocks

Ant systems [3] are inspired from the real behavior of real ants and are employed for combinatorial optimization problems. The basics of ant systems are founded on the theory of self-organizing systems [4] and the notion of stigmergy is presented in Fig. 2.

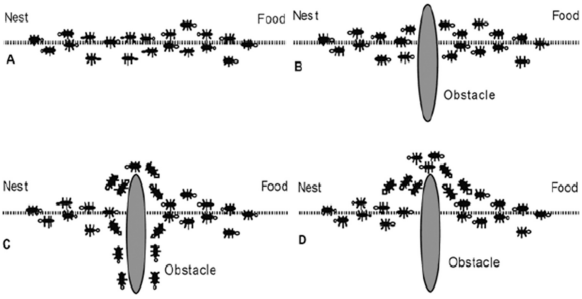


Fig. 2. The collective behavior changes

In [5], the authors have given an overview on the state of the art of the theoretical analysis of Ant Colony Optimization (ACO) algorithm. On a second stage, PSO is coupled with ACO for combinatory optimization.

Elloumi et al. [6] have presented an improved ACO algorithm supervised by PSO to solve continuous optimization problems. PSO algorithms are used to resolve continuous optimization problems while ACO algorithms are used for the discrete ones.

In [7], the authors have studied the multi-objective Particle Swarm Optimization (MOPSO) and found that more the number of the swarm increases more the accuracy of object achievement is increased.

In [8], the authors have proposed an approach that consists in combining fuzzy logic with ant colony Optimization (FACO) and fuzzy particle swarm optimization (FPSO) for solving TSP optimization problems.

In [9], the authors have proposed an optimization technique using multi-objective PSO (MOPSO) and FACO. This technique consists in combining these two methods. The objective of this combination is to reduce computation time and getting the shortest path.

In [10], the authors have presented an improved hybrid method (PSO–ACO) using the TSP benchmarks to validate our results.

During the last years, the trend was to use and improve graphics processing units (GPUs) as aco-processor. Designed mainly for graphics and game industry, GPUs have attracted many researchers due to its arithmetic computation power [11].

The paper is organized as follows. In Sect. 2, we define TSP problem. In Sect. 3, we present the basics of NVIDIA GPU based computing. Our system is illustrated in Sect. 4. Implementations of two algorithms and a number of experiments are done on four benchmarks with details of experiments and the reported results are exhibited in Sect. 5. Finally, we conclude our paper in Sect. 6.

2 The TSP Problem

TSP, the traveling salesman problem, which may be defined as follows: at first we initialize (n) cities that must be visited. Initially we start from a city chosen randomly and then returns to the starting city. The objective is to determine the overall distance and visit every city just once with respect to fixed start/end locations [12].

An illustration of this problem with 5 cities is given in Fig. 3; it shows two possible solutions, one in red and the other in green color. The two routes don't have the same length. A travelling salesman will choose the shortest path to reduce the cost of the travel. However, the TSP is said NP-complete. In fact, for n cities the number of possible route is equal to $(n-1)!/2$.

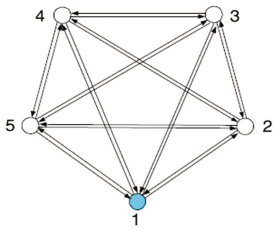


Fig. 3. TSP possible solutions for a simplified cities representation, here the number of cities is limited to 5

3 NVIDIA GPU Architecture

GPU processors are supported by the image processing and 3D data, as well as display. We distinct two main types of memory: local memory and global memory. This distinction allows memory spaces clearly separate memory areas read-only (constant, texture), local to a thread (local), read/write (global), short latency and deterministic (shared). The latency of each type of memory that can be estimated precisely statically. This parallel architecture, composed of a large number of calculations units, heavily exploit modes of SIMD (Single Instruction Multiple Data)/MIMD (Multiple Instruction Multiple Data) [13], allowing the simultaneous execution of many parts code. Parallelism can be achieved by using the concept of threads, i.e. lightweight processes that can run in parallel. When running a program, several thread groups will be spears in parallel to perform operations on a large data set.

The CUDA language for Compute Unified Device Architecture is available to the public since 2007. CUDA is a programming language similar to C/C++ to exploit the capabilities of GPU and material resources, particularly in regard to memory management and organization of treatment.

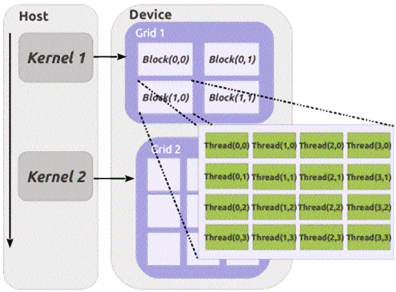


Fig. 4. Decomposition of blocks of grid data, where each block contains a number of elements executed in parallel. The term defines the Host CPU and Device term corresponds to the GPU. The size used for the definition of the grid is two-dimensional block-level and threads.

The principle of treatment of a problem on a highly parallel architecture is to break the problem into smaller problems that can be solved in parallel. Thus the partition of a wide array of data is performed by its decomposition into multiple blocks. Each block is run independently in parallel and the elements of each block are executed cooperatively in parallel. Figure 4 shows the decomposition of a set of data in a grid of 2×2 blocks are decomposed themselves in 4×4 elements.

All threads contained in a grid are sent to execution by a kernel, which can be defined as a simple function or program. To manage a large number of concurrent threads that can cooperate with each other, the architecture of graphics cards introduced threads of cooperation sets, also known as blocks of threads or thread blocks in CUDA terminology.

4 GPU-PSO and ACO-A-TSP

In this section, we will study our approach having two essential parts. The first part explain GPU PSO-A to TSP while the second part explain GPU ACO-A to TSP.

PSO and ACO optimization require high CPU computation resources. That's explain the necessity to use GPU accelerated system.

4.1 GPU PSO

The Fig. 5 shows roughly the Graphical Process Unit - Particle Swarm Optimization Applied to Travelling Salesman Problem.

Our goal is to cover all cities (designated nodes) once (if the particle passes through the city i to j it does not cross the town in the other direction, from j to i). Finally, the particle returns to the starting city, so we get a cycle.

The "**gpuArray**()" function allows copying data from the memory of the CPU to the GPU memory brings us to manipulate the table on the GPU memory.

First, the global best (P_{ig}) and local best (P_{il}) are elected, then we update the positions and velocities of the particles. These particles are assigned for N data and N threads.

We had to repeat these steps until reaching the maximum number of iterations; it is assigned to each node. This allows us to obtain an archive, according to the latter; we can make a comparison between the different obtained paths. We choose the best way in terms of its execution time. Finally, we return the GPU data to the CPU through the control "**gather**()".

4.2 GPU ACO

The Fig. 6 illustrates the diagram of the Graphical Process Unit - Ant Colony Optimization Applied to Travelling Salesman Problem.

The operation of the classic ACO is based on parameters that are often defined by the user of the algorithm. Thus, found settings that are appropriate for a problem are not suitable for other problems, forcing the user to perform numerous tests to define the parameters.

After coping data to GPU memory, the algorithm starts by assigning a city for each ant until all cites are affected to ants. When ants return to the starting city the amount of pheromone is updated in a cycle. Then, we Assign each node a thread and repeat these steps until reaching the maximum number of iteration. If we arrive at stopping criterion the concept of pheromone placing procedure guides the building procedure to each thread. The solutions to the intermediate partial problems are seen that we display the best lap, otherwise we return in step 2.

The dynamic memory structure that is inspired by the movement of the ant k for each iteration of the algorithm, see Fig. 6.

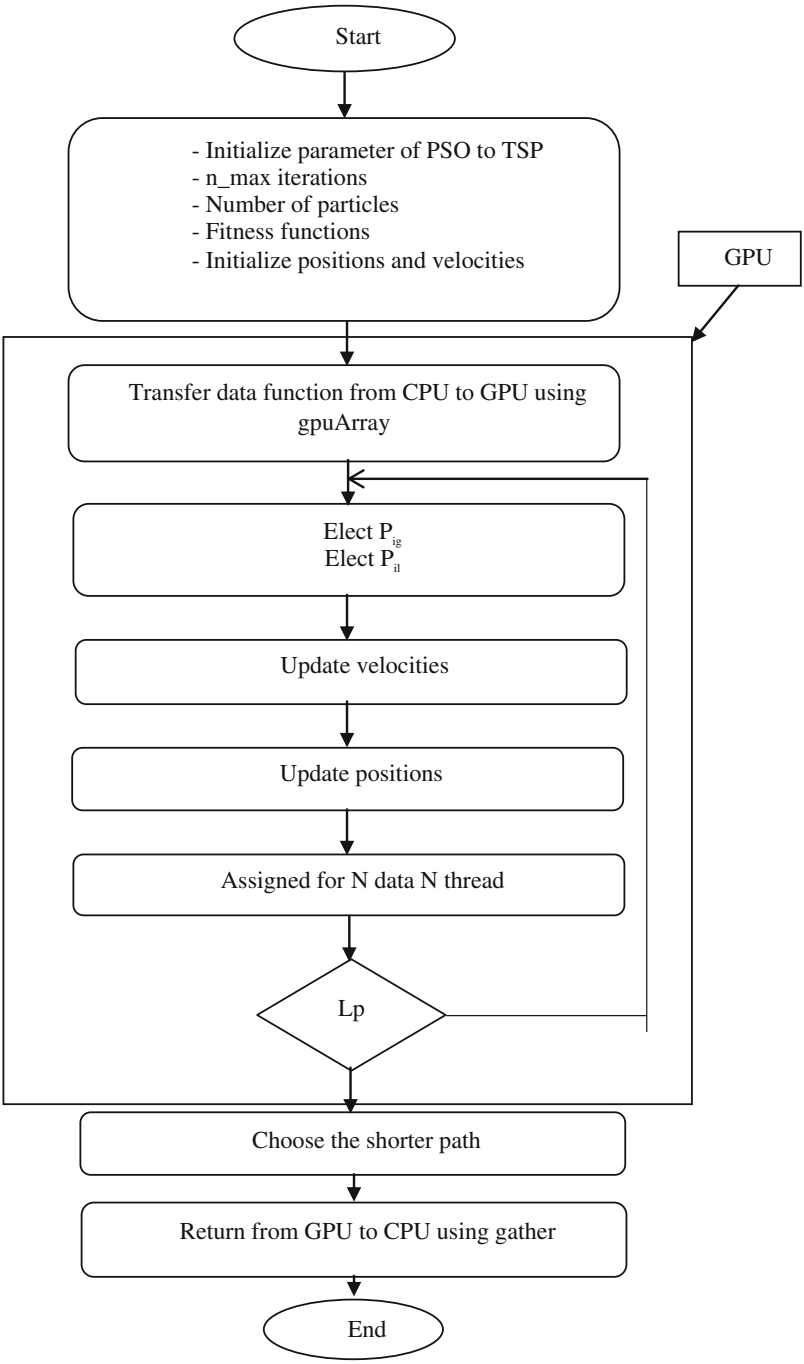


Fig. 5. The process of GPU-PSO-A-TSP

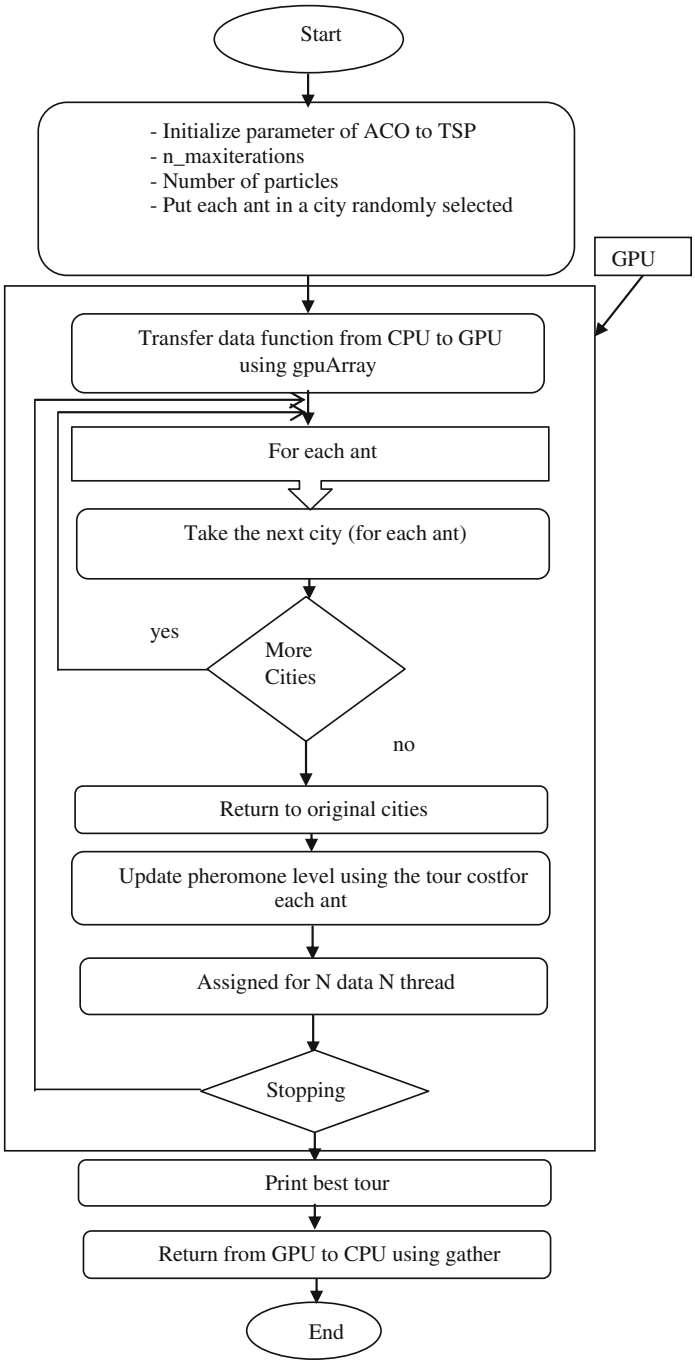


Fig. 6. The process of GPU-ACO-A-TSP

5 Experimental Results

We are based on well known benchmarks to validate the developed algorithms [10]. We have chosen in this paper from the benchmarks four TSP problems which consist in finding the optimal path to travel between graph cities of 22, 29, 30 and 48 nodes. In our approach, we begin by resetting the parameters feature of PSO applied to the TSP; to say the number of nodes contained in a graph, a weight of every particle and the coefficients of acceleration. The maximum number of iterations in this case was taken as 1000 iterations. The number of used particles depends on the graph.

For example, for a graph with 22 cities, the possible number of particles of the population to use in this first test is 22; the second test is 70 for the third test is 100.

We had 1000 iteration because we have tests by using 2000 and 3000 iterations.

Our method consists in using 1000 iterations because the number and the time of the cycle are proportional among iterations to avoid the wasting.

Our approaches GPU Optimization Swarm of particles and Ant Colony Optimization applied to the TSP (GPU-PSO ACO-A-TSP) is coded in Matlab 2014b and executed on a processor Intel ® Core T i7-4700MQ (6 MB cache memory, 3.40 GHz) PC with memory of 12 GB and NVIDIA GeForce 480 M and Windows 7.

There are many parameters used for our approach. The size of the population, which we are three times going to increase, is the number of knots of the social and cognitive probability, having c1 and c2, defined as c1 = c2 = 2. The mass of inertia w is taken as 0,9 and the maximum of the speed live taken as 100 and the dimension of the space as 10. Every cycle of TSP is executed during five replications and 1000 iterations. Both has and ß control the relative importance of pheromone trail and the distance between cities TSP where has $\alpha = 1.5$, ß = 2 Refers to the speed of pheromone evaporation $\rho = 0{,}7$. Each test TSP is performed for 5 replications iterations and 1000.

5.1 TSP Solved by PSO and GPU PSO

In Table 1, N refers to the number of nodes, CPU-T.PSO refers to the best time for PSO (per seconds) and CPU-L.PSO refers to CPU-PSO the length for PSO. GPU used the same indices.

In the same table, we have tried to represent the different numbers of nodes, after that, we have attempted to increase the population of PSO keeping the same number of nodes. Thus, It was found from number 22 to 48 nodes, that the route of the shortest path decreases when the number of PSO population increases. Therefore, when the execution time increases, the number of (CPU-GPU) PSO population increases too.

5.2 TSP Solved by ACO and GPU ACO

In Table 2, N refers to the number of nodes, T.ACO refers to the best time for ACO (per seconds), L.ACO is the best Length for ACO.

Table 1. PSO and GPU PSO for TSP

		CPU		GPU	
N	Size of population of PSO	T.PSO (s)	L.PSO (km)	T.PSO (s)	L.PSO (km)
22	22	0.1406	90.6884	**0.0848**	90.6884
	70	0.4556	90.4220	**0.2791**	90.4220
	100	0.5707	89.4898	**0.3950**	89.4898
29	29	0.4181	1.1761e+004	**0.2346**	1.1761e+004
	70	0.9632	1.0900e+004	**0.5831**	1.0900e+004
	100	1.2177	1.0472e+004	**0.7763**	1.0472e+004
30	30	0.3577	584.0341	**0.2736**	584.0341
	70	0.7334	562.0160	**0.5969**	562.0160
	100	1.1059	545.7844	**0.8535**	545.7844
48	48	3.0961	4.5973e+004	**0.6366**	4.5973e+004
	70	4.3249	4.4654e+004	**0.9269**	4.4654e+004
	100	4.8480	4.1158e+004	**1.4414**	4.1158e+004

Table 2. ACO and GPU ACO for TSP

		CPU		GPU	
N	Size of population of ACO	T.ACO (s)	L.ACO (km)	T.ACO (s)	L.ACO (km)
22	22	0.2438	77.8000	**0.0930**	77.8000
	70	0.4969	77.1834	**0.2964**	77.1834
	100	0.6866	76.1212	**0.4223**	76.1212
29	29	0.4190	1.1621e+004	**0.2602**	1.1621e+004
	70	1.0881	1.0530e+004	**0.6132**	1.0530e+004
	100	1.3713	1.0432e+004	**0.9036**	1.0432e+004
30	30	0.3724	537.9874	**0.2933**	537.9874
	70	0.7652	495.5985	**0.6699**	495.5985
	100	1.4225	491.7651	**1.0160**	491.7651
48	48	4.0005	4.2086e+004	**1,1433**	4.2086e+004
	70	4.5047	4.1420e+004	**1,6658**	4.1420e+004
	100	6.8139	4.0585e+004	**2.3992**	4.0585e+004

In the second table we have tried to represent the different numbers of nodes. After that we have tried to increase the number of people of ACO keeping the same number of nodes. In this Table it was found that the route of the shortest path decreases when the number of ACO population increases. When the execution time increases the number of population increases too. The execution time depends on the complexity of the TSP as well.

Now comparing the results of the two tables, we notice that the results of the shortest path of (CPU-GPU) ACO are better compared to the (CPU-GPU) PSO, but the best performance is that of (CPU-GPU) PSO compared to that of (CPU-GPU) ACO time.

Compared to previous works ACO-A-TCP and PSO-A-TCP [10], our algorithms GPU-ACO-A-TSP and GPU-PSO-A-TSP implemented on GPU have the advantage of reducing the computational time for solving TSP problems.

6 Conclusion

In this paper we have given two approaches the GPU-PSO–A-TSP and GPU-ACO–A-TSP. We have used PSO and ACO, meta-heuristics optimization algorithms, for resolving TSP problem. We have also used the parallel GPU programming model to reduce the PSO and ACO algorithms computational time.

According to the results of the two tables, we notice that the results of the shortest path of (CPU-GPU) ACO are better compared to the (CPU-GPU) PSO, but the best performance is that of (CPU-GPU) PSO compared to that of (CPU-GPU) ACO time.

For this reason, we have achieved hybridization between PSO and ACO using the GPU based on [14].

Acknowledgments. The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program.

References

1. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948 (1995)
2. Hendtlass, T.: WoSP: a multi-optima particle swarm algorithm. In: The IEEE Congress on Evolutionary Computation, vol. 1, pp. 727–734 (2005)
3. Dorigo, M., Maniezzo, V., Colorni, A.: The ant system: optimization by a colony of cooperating agents. IEEE Trans. Syst. Man Cybern. B Cybern. **26**(2), 29–41 (1996)
4. Garnier, S., Gautrais, J., Theraulaz, G.: The biological principles of swarm intelligence. Swarm Intell. (2007)
5. Elloumi, W., Alimi, A.M.: Combinatory optimization of ACO and PSO. In: International Conference on Metaheuristique and Nature Inspired Computing, pp. 1–8, October 2008
6. Elloumi, W., Rokbani, N., Alimi, A.M.: Ant supervised by PSO. In: International Symposium on Computational Intelligence and Intelligent Informatics, pp. 161–166, October 2009
7. Elloumi, W., Alimi, A.M.: A more efficient MOPSO for optimization. In: The Eight ACS/IEEE International Conference on Computer Systems and Applications, AICCSA, pp. 1–7, May 2010
8. Elloumi, W., Baklouti, N., Abraham, A., Alimi, A.M.: Hybridization of fuzzy PSO and fuzzy ACO applied to TSP. In: 13th International Conference on Hybrid Intelligent Systems (HIS), pp. 106–111, December 2013
9. Elloumi, W., Baklouti, N., Abraham, A., Alimi, A.M.: The multi-objective hybridization of particle swarm optimization and fuzzy ant colony optimization. J. Intell. Fuzzy Syst., 515–525 (2014). http://dx.doi.org/10.3233/IFS-131020

10. Elloumi, W., El Abed, H., Abraham, A., Alimi, A.M.: A comparative study of the improvement of performance using a PSO modified by ACO applied to TSP. J. Appl. Soft Comput. **25**, 234–241 (2014)
11. Kirk, D.B., Hwu, W.W.: Programming Massively Parallel Processors: A Hands-on Approach. Morgan Kaufmann, San Francisco (2010)
12. Gavish, B., Graves, S.C.: The travelling salesman problem and related problems (1978)
13. Flynn, M.J.: Some computer organizations and their effectiveness. IEEE Trans. Comput. **21**, 948–960 (1972)
14. Bali, O., Elloumi, W., Abraham, A., Alimi, A.M.: GPU PSO and ACO applied to TSP for vehicle security tracking. J. Inf. Assur. Secur. **11**, 369–384 (2016)

A Parallel Adaptive PSO Algorithm with Non-iterative Electrostatic Repulsion and Social Dynamic Neighborhood

Daniel Soto[1(✉)] and Wilson Soto[2]

[1] Department of Computer Science, Antioquia University, Medellín, Colombia
anderson.soto@udea.edu.co
[2] School of Exact Sciences and Engineering,
Sergio Arboleda University, Bogotá, Colombia
wilson.soto@usa.edu.co

Abstract. This paper presents a variation on the DEPSO-2S algorithm, called Parallel Adaptive PSO (PA2PSO). The goals of the PA2PSO algorithm are to find the value closest the global minimum of the function evaluated improving the location as well as the interaction of the particles by means of two important characteristics: non-iterative electrostatic repulsion and social dynamic neighborhood, and to reduce the response time with a parallel implementation. The PA2PSO achieves in most cases positive results in solving benchmark test functions (unimodal and multimodal functions) compared with nine outstanding PSO algorithms.

Keywords: Bio-inspired algorithms · Metaheuristics · Particle swarm optimization

1 Introduction

The particle swarm optimization (PSO) is an evolutionary algorithm with particle population that explores a search space in order to find an optimal solution of a problem. This algorithm was proposed in 1995 [4] and it is inspired in the social behavior of the animals living in swarms such as fishes or birds.

The particle swarm is composed by simples agents called particles, each particle located into a search space represents a possible solution for a given problem. Each particle has three properties: a coordinates vector, a velocity vector and the memory capacity. The trajectory of the particle is influenced by three components: the inertial component, the cognitive component and the social component.

Formally, the particle swarm optimization is expressed through the mathematical model:

$$v_i^{t+1} = wv_i^t + c_1 r_1 (x_i^{lbest} - x_i^t) + c_2 r_2 (x^{gbest} - x_i^t) \quad (1)$$

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_56

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{2}$$

Where, the variable w is the inertial component. The inertial component was introduced by [10]. Besides, the c_1 is the cognitive component, c_2 is the social component and the variables r_1 and r_2 are random variables between 0 and 1.

The values *lbest* and *gbest* are obtained depending of the optimization objective, if the objective is minimization the equations are:

$$lbest_i(t+1) = \begin{cases} lbest(t) & \text{if } f(x_i(t+1)) \geq lbest_i(t) \\ x_i(t+1) & \text{otherwise} \end{cases} \tag{3}$$

$$gbest(t+1) = argmin_{lbest_i} f(lbest(t+1)) \text{ where } 1 \leq i \leq N \tag{4}$$

W.L.O.G the equations are similar for maximization.

The applications of the PSO algorithm are several and diverse, for example, in [8] shows in-depth a state of art on the applications of the PSO algorithm, such as, communication network, combinatorial optimization, financial, entertainment, prediction and robotics, among others.

A considerable amount of literature has been published on improvements of PSO algorithm, for example, Constriction Coefficient Particle Swarm, FIPS algorithm (Fully Informed Particle Swarm) and Tribe-PSO [1]. These studies are oriented to the behavior and location of the particles and the variation of the topology.

There are two kind of recent important works that were evaluated with several benchmark functions for global optimization: non-swarm algorithms and PSO variants. Some non-swarm algorithms: Real Genetic Algorithm (RGA) and Gravitational Search Algorithm (GSA) [9]; Teaching Learning Based Optimization (TLBO) algorithm [14]. Some PSO variants: Multi-Swarm PSO (PSO-2S), Multi-Swarm PSO with Differential Evolution (DEPSO-2S), Efficient Population Utilization Strategy PSO (EPUS-PSO), Comprehensive Learning PSO (CLPSO), among others [1] and A2PSO [11].

Several parallel PSO algorithms have been recently proposed in [7], the authors present the Cooperative Parallel PSO (CPPSO) algorithm with four different implementations: OpenMP, MPI, MPI+OpenMP and CUDA. The results of this study indicate that the OpenMP implementation achieved the better results. In [12], the authors studied four parallel and distributed PSO algorithms: Particle Swarm Evolver (PSE), Repulsive PSE (RPSE), Multi-swarm PSO (MPSO) and Multi-swarm Repulsive (MRPSO), the idea of these algorithms is that exchanging individuals at fixed time rates between the different swarms. However, the two above works have a number limited of objective functions used as benchmarks.

The DEPSO-2S is a powerful algorithm for global optimization introduced by El Dor *et al.* [2]. The DEPSO-2S avoids the premature convergence in local optima and improves the efficiency of PSO, however the iterative calculation for locating the particles in the space wastes a lot of time. The performance of DEPSO-2S algorithm highly depends of a fixed social network topology, specially designed as regular, trying gather the individuals in a same zone of the space.

Another weakness of the DEPSO-2S algorithm is that the evolution is easily influenced for the particles with the fitness values opposite due to that each particle is affected for the neighbor with better fitness.

The algorithm proposed in this work (PA2PSO) is based in the DEPSO-2S algorithm with a parallel implementation (OpenMP) and two improvements: (*i*) the location of the particles in the search space through the concept of electrostatic repulsion and (*ii*) the interaction of the particles through the concept of social dynamic neighborhood. The improvements have the advantages of deleting the iterative calculation and building a dynamic topology.

Our approach is compared in three frameworks [1,9,14], where are included basic PSO, GSA, RGA, PSO-2S, DEPSO-2S, EPUS-PSO, CLPSO, TLBO and A2PSO, with different parameters in order to asses the results in some benchmarking test functions.

The structure of the paper is: the Sect. 2 explains the proposed algorithm. The experimental results are reported in Sect. 3 and for last the conclusions in Sect. 4.

2 The Proposed Algorithm

The proposed algorithm called PA2PSO has a parallel implementation and is based in DEPSO-2S [2,3]. The algorithm PA2PSO has two important characteristics. The characteristics attempt to improve the particles location in the search space. The first characteristic improves the covering of the swarm in the search space with regard to an initialization random. The second characteristic is the influence vector for each particle based in the concept of social dynamic neighborhood.

2.1 Initial Location

The initial location is based in the concept the electrostatic repulsion [1]. The electrostatic repulsion consists in assign to each particle a repulsion area inside which cannot locate another particle.

The PSO algorithm with electrostatic repulsion randomly places the particles inside the search space and iteratively corrects the positions. The graphical representation of the methods is shown in the Fig. 1.

The PA2PSO algorithm uses the electrostatic repulsion eliminating the iterative stage with the calculate of the volume, radio and interaction area of the repulsion of each particle.

The hypervolume $\mathcal{V}$ of electrostatic repulsion for each particle is calculated by the Eq. (5).

$$\mathcal{V} = \frac{L_1 * L_2 * \ldots * L_d}{2 * N} \tag{5}$$

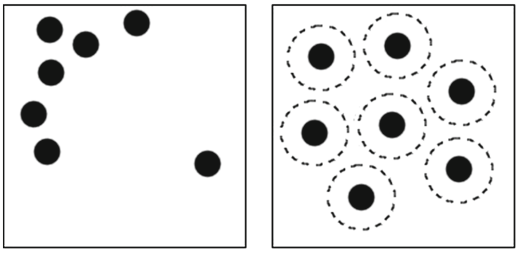


Fig. 1. Initialization stage for 7 particles in a space 2D. Initialization random (*left*) and initialization random with electrostatic repulsion (*right*).

Where, N is the population, d are the dimensions of the search space and L_i represents the length of the search space on the dimension i.

With the hypervolume $\mathcal{V}$ calculated we can determine the repulsion radio $\mathcal{R}$ which is utilized in the particles localization. The Eq. (6) calculates the repulsion radio with the hypervolume of a hypersphere due to be a scheme n-dimensional.

$$\mathcal{R} = \left(\frac{\mathcal{V} * \Gamma((d/2)+1)}{(\pi)^{d/2}} \right)^{1/d} \tag{6}$$

For avoiding that a particle is in a repulsion area of another particle, the particle must satisfies the Eq. (7).

$$\sqrt{(x_1^{'} - x_1^{''})^2 + (x_2^{'} - x_2^{''})^2 + \ldots + (x_d^{'} - x_d^{''})^2} > 2 \cdot \mathcal{R} \tag{7}$$

2.2 Influence Vector

An influence vector is a vector based in the concept of the social neighborhood. The social neighborhood is a kind of neighborhood generally without fixed topology, adaptive and static or dynamic behavior. The social neighborhood does not have geographic or distance rules because a particle simply can have neighbors defined in anywhere of the search space [1].

The influence vector in the PA2PSO algorithm is built with the Eq. (8).

$$z_i^{rand(N)} = 0.2; |z_i^n \neq 0| = 5, \forall n \tag{8}$$

The values of the influence vector are iteratively update under the concept of symmetric weakly influences [6]. The Eq. (9) calculates the updating rule and depends of the nth–neighbor selected, concept-based in the dynamic neighborhood [15].

$$z_i^n(t+1) = \begin{cases} z_i^n(t) + \dfrac{x_i^{lbest} - x_n^{lbest}}{\sum_{j=0}^{N}(x_i^{lbest} - x_j^{lbest})} & \text{if } 0.2 \leq z_i^n(t) < 1.0 \\ 1 & \text{if } z_i^n(t) \geq 1.0 \\ 0 & \text{if } z_i^n(t) < 0.2 \end{cases} \tag{9}$$

Recent works have determined that inside a birds swarm each individual is in average influenced by 6 or 7 nearest neighbors [5]. The Eq. (10) verifies the above condition, if the condition is false then a random position is selected with value equal to zero for changing it in the vector. The process is iterative until to obtain the predefined number of individuals for creating the social neighborhood.

$$z_i^{rand(N)}(t+1) = 0.2, \text{ if } z_i^{rand(N)}(t) = 0; |z_i^n(t+1) \neq 0| = 5, \forall n \tag{10}$$

After of update the values of the influence vector, for each particle i are calculated the location and velocity with the Eqs. (11) and (12).

$$v_i^{t+1} = wv_i^t + r_1c_1(x_i^{lbest} - x_i^t) + r_2c_2\left(\sum_{n=0}^{N} z_i^n(t)(x_n^{lbest} - x_i^t)\right) \tag{11}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{12}$$

2.3 Parallel Implementation

The principal advantage of parallel computing in intelligent optimization especially in the PSO algorithms is the exploration and the exploitation of the search space in less time, because the population can be divided into distinct subpopulations as independent units performing in parallel.

The PA2PSO algorithm uses the parallel island model. The island model reduces the convergence probability in the local optimum, transforming it to a cooperative scheme between subpopulations [13].

In this model, the population is divided into several subpopulations and the different processors run the same algorithm on each subpopulation. Eventually, after a certain iteration number, the processors commute to a stage of intercommunication where individuals are interchange.

The algorithm chooses the better solution of the processors for each iteration. In the end of the iterations, the algorithm has the best solution of the execution.

The full PA2PSO algorithm is shown in Algorithm 1, where for each iteration there is a parallel implementation (see lines 6–14).

Algorithm 1. PA2PSO Algorithm

Require: N ▷ Number of individuals or particles
Ensure: *gbest* ▷ Global optimum

1: Locate N particles ▷ Eqs. 6, 7, 8
2: Evaluate the fitness of N particles
3: Determine the local optimum of N particles (*lbest*) ▷ Eq. 3
4: Update the *lbest* value ▷ Eq. 3
5: Calculate the neighborhood values ▷ Eqs. 9, 10
6: **while** number of iterations **do**
7: **if** iteration is a multiple of iteration interchange **then**
8: Interchage individuals
9: **end if**
10: Move the particles ▷ Eqs. 11, 12
11: Evaluate the fitness N particles
12: Update the *lbest* value ▷ Eq. 3
13: Calculate the neighborhood values ▷ Eqs. 9, 10
14: **end while**
15: Update *gbest* value ▷ Eq. 4
16: **return** *gbest*

3 Results

The proposed PA2PSO algorithm was implemented in language C++ using the MPFR library for the arithmetic computes with decimal multiple precision. Our algorithm presents a parallel implementation using OpenMP. OpenMP is an API for multi-platform shared-memory parallel programming in C/C++ and Fortran.

All the experiments were executed on a PC with Quad-Core Intel Xeon processor 2.4 GHz and 4 GB memory.

Commonly, a set of benchmark test functions is used for evaluating the PSO algorithms (show Table 1).

We present experimental results for our algorithms compared against other algorithms in three frameworks [1,9,14]. Each one contains different classes of algorithms. The first framework (Framework A) includes the basic PSO algorithm (original parameters setting: $\omega \in [0.2, 0.9]$, $c_1 = 2$ and $c_2 = 2$), RGA algorithm and GSA algorithm. The dimension size is 30 for the functions except for the function F7 that is 2. The parameters for the algorithms were: 1000 iterations and 50 individuals. Table 2 shows the results obtained from the Framework A.

The second framework (Framework B) includes the algorithms with better results in [1]: DEPSO-2S, PSO-2S, EPUS-PSO and CLPSO. This framework has two tests. In the test 1 the dimension size is 10 for the functions except for the function F11 that is 2. The parameters for the algorithms were: 40000 iterations and 320 individuals. We decrease the parameters for the PA2PSO and A2PSO

Table 1. The test functions for evaluating PSO algorithms.

Function	Function name	Modality	Reference
F1	Sphere function	Uni	[1,14]
F2	Quadric function	Uni	[1]
F3	Schwefel 2.22	Uni	[14]
F4	Rosenbrocks	Uni	[1,14]
F5	Step	Uni	[14]
F6	Schwefel	Multi	[14]
F7	6 Hump CamelBack	Multi	[14]
F8	Griewank's function	Multi	[1,14]
F9	Ackley's function	Multi	[1,14]
F10	Rastrigin's function	Multi	[1,14]
F11	Tripod's function	Multi	[1]
F12	Quartic	Uni	[14]
F13	Penalized	Multi	[14]
F14	Penalized 2	Multi	[14]
F15	Branin	Multi	[14]
F16	GoldStein–Price	Multi	[14]
F17	Schwefel 1.2	Uni	[14]

Table 2. The results Framework A. The letter b means the best results and letter w means the worst results.

Name		RGA	PSO	GSA	A2PSO	PA2PSO
F1	b	21.87	1.2×10^{-3}	7.1×10^{-11}	6.40×10^{-13}	$\mathbf{8.13 \times 10^{-30}}$
	w	23.45	5.0×10^{-2}	2.1×10^{-10}	6.08×10^{-12}	$\mathbf{1.26 \times 10^{-15}}$
F2	b	5.6×10^{3}	2.2×10^{3}	0.15×10^{3}	8.81	$\mathbf{9.80 \times 10^{-4}}$
	w	5.6×10^{3}	4.1×10^{3}	0.16×10^{3}	**36.95**	2.40×10^{4}
F3	b	1.07	1.9×10^{-3}	4.03×10^{-5}	4.82×10^{-7}	$\mathbf{5.11 \times 10^{-16}}$
	w	1.13	2.0	$\mathbf{4.07 \times 10^{-5}}$	4.93×10^{-4}	0.0011095
F4	b	1.0×10^{3}	1.7×10^{3}	25.16	21.61	**16.408**
	w	1.1×10^{3}	3.6×10^{4}	25.18	**23.49**	105.74
F5	b	24.01	1.0×10^{-3}	7.7×10^{-11}	3.89×10^{-13}	$\mathbf{6.17 \times 10^{-28}}$
	w	24.55	0.02	2.6×10^{-10}	9.84×10^{-13}	$\mathbf{1.26 \times 10^{-16}}$
F6	b	$\mathbf{-1.2 \times 10^{4}}$	-9.8×10^{3}	-2.8×10^{3}	-4.98×10^{3}	-8645
	w	$\mathbf{-1.2 \times 10^{4}}$	-9.8×10^{3}	-1.1×10^{3}	-4.39×10^{3}	-4119.7
F7	b	**−1.0313**	**−1.0316**	**−1.0316**	**−1.0316**	**−1.0316**
	w	**−1.0316**	**−1.0316**	**−1.0316**	**−1.0316**	**−1.0316**
F8	b	1.14	0.0081	0.04	**0**	**0**
	w	1.16	0.055	0.29	4.17×10^{-2}	$\mathbf{2.25 \times 10^{-2}}$

algorithms: 5000 iterations and 50 individuals. In the test 2 the functions have the dimension size 30. The parameters for the algorithms were: 150000 iterations and 400 individuals. We decrease the parameters for the PA2PSO and A2PSO algorithms: 50000 iterations and 150 individuals. Table 3 shows the results obtained from the Framework B test 1 and Table 4 shows the results obtained from Framework B test 2. The original parameters setting of the algorithms of the framework were: $\omega \approx 0.72$, $c_1 \approx 1.19$ and $c_2 \approx 1.19$.

Table 3. The results Framework B test 1. The letter b means the best results and letter w means the worst results. The character "–" means not evaluated.

Name		DEPSO-2S	PSO-2S	EPUS-PSO	CLPSO	A2PSO	PA2PSO
F1	b	1.3×10^{24}	1.05×10^{-86}	5.55×10^{-153}	4.02×10^{-21}	7.20×10^{-247}	$\mathbf{2.08 \times 10^{-264}}$
	w	3.33×10^{-2}	3.10×10^{-86}	2.44×10^{-134}	6.49×10^{-21}	$\mathbf{8.40 \times 10^{-243}}$	1.92×10^{-22}
F2	b	6.18×10^{-28}	6.18×10^{-28}	9.32×10^{-12}	5.04×10^{2}	1.23×10^{-36}	$\mathbf{4.30 \times 10^{-101}}$
	w	$\mathbf{2.13 \times 10^{-27}}$	$\mathbf{2.13 \times 10^{-27}}$	4.16×10^{-11}	1.99×10^{2}	4.62×10^{-27}	1.55×10^{-16}
F8	b	–	5.3×10^{-5}	–	–	**0**	**0**
	w	–	$\mathbf{2.93 \times 10^{-3}}$	–	–	1.47×10^{-2}	4.02×10^{-3}
F9	b	4.79×10^{-13}	4.12×10^{-15}	2.72×10^{-15}	1.78×10^{-11}	$\mathbf{-1.06 \times 10^{-14}}$	$\mathbf{-1.06 \times 10^{-14}}$
	w	2.53×10^{-2}	1.21×10^{-15}	1.5×10^{-15}	1.63×10^{-11}	1.77×10^{-14}	$\mathbf{-1.06 \times 10^{-14}}$
F10	b	**0**	**0**	**0**	2.35×10^{-9}	**0**	**0**
	w	**0**	**0**	**0**	2.72×10^{-9}	8.9×10^{-1}	6.6997
F11	b	–	1.79×10^{-5}	–	–	1.20×10^{-49}	**0**
	w	–	$\mathbf{2.45 \times 10^{-3}}$	–	–	2.13×10^{-1}	1.0001

The third framework (Framework C) has the TLBO algorithm. The dimension size is 30 for the functions. The parameters for the algorithms were in (100, 666, 2000) iterations and (25, 50, 75) individuals depending of the function. Table 5 shows the specific parameters for each function and the results obtained from the Framework C.

In the three frameworks, the parameters setting of the PA2PSO and A2PSO algorithms were random values in a range: $\omega \in [0.5, 0.8]$, $c_1 \in [0.2, 1.6]$ and $c_2 \in [0.5, 2.1]$.

We use two measures for evaluating the performance of our algorithm, standard error and confidence interval. The standard error (SE) is calculated with the Eq. (13).

$$SE = \frac{s}{\sqrt{n}} \tag{13}$$

where, s is the standard deviation or a measure of how varied the fitness values are among our replicates and n is the number of replicates we ran. For our statistical analysis the value n is 30.

Table 4. The results Framework B test 2. The letter b means the best results and letter w means the worst results.

Name		DEPSO-2S	PSO-2S	EPUS-PSO	CLPSO	A2PSO	PA2PSO
F1	b	1.25×10^{-25}	3.02×10^{-119}	8.50×10^{-263}	1.34×10^{-25}	5.05×10^{-379}	$\mathbf{0}$
	w	3.24×10^{-24}	1.56×10^{-118}	1.02×10^{-262}	1.71×10^{-25}	$\mathbf{5.18 \times 10^{-371}}$	8.17×10^{-32}
F2	b	3.88×10^{-12}	6.18×10^{-11}	197	2.41×10^{4}	1.27×10^{-13}	$\mathbf{8.0457 \times 10^{-189}}$
	w	2.50×10^{-11}	2.63×10^{-11}	4.69	6.78×10^{3}	$\mathbf{5.29 \times 10^{-13}}$	3.94×10^{-3}
F9	b	6.48×10^{-14}	1.63×10^{-1}	3.91×10^{-15}	9.90×10^{-14}	4.61×10^{-14}	$\mathbf{0}$
	w	3.33×10^{-14}	3.85×10^{-1}	1.07×10^{-15}	3.80×10^{-14}	2.16×10^{-13}	$\mathbf{0}$
F10	b	3.98×10^{-14}	1	$\mathbf{0}$	1.87×10^{-9}	10.9	6.96
	w	2.33×10^{-14}	11.8	$\mathbf{0}$	5.34×10^{-9}	14.9	55.4

Table 5. The results Framework C. The letter b means the best results and letter w means the worst results.

Name	Iter.	N	TLBO		A2PSO		PA2PSO	
			b	w	b	w	b	w
F1	2000	25	2.09×10^{-5}	7.27×10^{-4}	$\mathbf{1.61 \times 10^{-37}}$	$\mathbf{8.16 \times 10^{-37}}$	6.18×10^{-27}	2.09×10^{-13}
F3	2000	25	3.53×10^{-5}	9.58×10^{-4}	1.12×10^{-5}	$\mathbf{4.44 \times 10^{-4}}$	$\mathbf{7.79 \times 10^{-19}}$	5.70×10^{-3}
F4	666	75	**13.88407**	**19.08793**	21.42206	27.77500	23.664	93.483
F5	2000	25	1.81×10^{-5}	8.06×10^{-4}	4.04×10^{-6}	7.85×10^{-4}	$\mathbf{0}$	$\mathbf{2.07 \times 10^{-12}}$
F6	666	75	**−12569.49**	**−12158.04**	−4297.20	−3868.12	−7333.6	−4377.7
F7	2000	25	**−1.031628**	**−1.031628**	**−1.031628**	**−1.031628**	**−1.031628**	**−1.031628**
F8	2000	25	8.93×10^{-5}	$\mathbf{5.27 \times 10^{-4}}$	3.55×10^{-15}	8.42×10^{-4}	$\mathbf{0}$	5.89×10^{-2}
F9	2000	25	1.89×10^{-4}	6.21×10^{-4}	1.99×10^{-8}	$\mathbf{6.22 \times 10^{-6}}$	$\mathbf{-1.07 \times 10^{-14}}$	1.778
F10	2000	25	$\mathbf{1.78 \times 10^{-4}}$	$\mathbf{8.67 \times 10^{-4}}$	33.21	42.78	21.889	153.55
F12	1000	50	0.001253	0.014182	2.20×10^{-7}	8.93×10^{-7}	$\mathbf{1.21 \times 10^{-47}}$	$\mathbf{7.69 \times 10^{-18}}$
F13	666	75	2.67×10^{-4}	8.27×10^{-4}	1.96×10^{-8}	$\mathbf{2.00 \times 10^{-6}}$	$\mathbf{7.56 \times 10^{-22}}$	3.03×10^{-5}
F14	666	75	$\mathbf{2.37 \times 10^{-8}}$	$\mathbf{6.77 \times 10^{-4}}$	4.13×10^{-1}	5.08×10^{-1}	0.1	0.1229
F15	2000	25	**0.3978**	0.3984	**0.3978**	0.4023	**0.3978**	**0.3978**
F16	2000	25	3	3	**2.9999**	**2.9999**	**2.9999**	3
F17	2000	25	3.87×10^{-5}	8.17×10^{-4}	6.51×10^{-6}	8.48×10^{-4}	$\mathbf{2.94 \times 10^{-34}}$	$\mathbf{1.47 \times 10^{-24}}$

The confidence interval (CI) is a way of estimating the mean of an unknown distribution from a set of data drawn from this distribution. We can compute an estimated 95% confidence interval by using the Eq. (14).

$$CI = \bar{x} \pm (1.96 * SE) \tag{14}$$

where, $\bar{x}$ is the sample mean.

The Table 6 shows the statistical analysis for each framework and its algorithms. Besides in the table appears the average time in seconds.

Table 6. The statistical analysis for each experimental framework.

Fram.	Name	Mean	SE	CI−	CI+	Time (sec.)
A	F1	1.03×10^{-16}	5.47×10^{-17}	-4.20×10^{-18}	2.10×10^{-16}	12.961
A	F2	807.77	809	-7.62×10^{2}	2.38×10^{3}	14.468
A	F3	6.02×10^{-5}	3.94×10^{-5}	-1.70×10^{-5}	1.37×10^{-4}	13.649
A	F4	21.638	2.9033	15.95	27.33	14.427
A	F5	7.02×10^{-18}	4.29×10^{-18}	-1.40×10^{-18}	1.54×10^{-17}	13.195
A	F6	−5416.6	190.5	−5788.98	−5042.22	22.819
A	F7	−1.0316	9.79×10^{-8}	−1.0316	−1.0316	1.105
A	F8	0.0045312	0.0011938	0.002191	0.006871	17.603
B_1	F1	6.42×10^{-24}	6.40×10^{-24}	-6.12×10^{-24}	1.90×10^{-23}	20.792
B_1	F2	5.94×10^{-18}	5.21×10^{-18}	-4.27×10^{-18}	1.61×10^{-17}	21.593
B_1	F8	0.0025367	0.0040265	0.0174751	0.0332589	31.016
B_1	F9	-1.07×10^{-14}	0	-1.07×10^{-14}	-1.07×10^{-14}	24.951
B_1	F10	3.1563	0.36701	2.43696	3.87564	36.235
B_1	F11	0.68296	0.0083719	0.518871	0.847049	5.498
B_2	F1	0	0	0	0	1812.526
B_2	F2	0.0001313	0.0001313	−0.000126	0.0003887	1988.459
B_2	F9	0	0	0	0	1954.219
B_2	F10	17.5666	1.35892	12.853324	22.280008	2921.701
C	F1	7.28×10^{-15}	6.97×10^{-15}	-6.73×10^{-15}	2.09×10^{-14}	13.174
C	F3	4.02×10^{-4}	2.33×10^{-4}	-5.39×10^{-5}	8.59×10^{-4}	13.894
C	F4	26.788	2.3045	22.2711	31.3048	14.406
C	F5	8.01×10^{-14}	6.92×10^{-14}	-5.55×10^{-14}	2.16×10^{-13}	13.37
C	F6	−5423.7	151.46	−5720.56	−5126.83	22.699
C	F7	−1.0316	2.04×10^{-7}	−1.0316	−1.0316	13.333
C	F8	0.0089802	0.0027117	0.0036653	0.0142951	16.452
C	F9	0.23276	0.099975	0.036809	0.428711	16.827
C	F10	65.954	6.1327	53.933908	77.974092	22.929
C	F12	2.49×10^{-19}	2.58×10^{-19}	-2.11×10^{-19}	7.99×10^{-19}	12.699
C	F13	1.01×10^{-6}	1.01×10^{-6}	9.70×10^{-7}	2.99×10^{-6}	34.885
C	F14	0.10077	7.63×10^{-4}	0.09927	0.10227	31.768
C	F15	0.3979	4.47×10^{-6}	0.39789	0.39791	13.185
C	F16	3	1.15×10^{-15}	3	3	13.908
C	F17	6.08×10^{-26}	4.91×10^{-26}	-3.55×10^{-26}	1.57×10^{-25}	15.243

4 Conclusions

This work proposed a PSO algorithm. The proposed approach includes two improvements, non-iterative electrostatic repulsion and social dynamic neighborhood with parallel implementation. The experimental results were promising, these shown that our approach gets better accuracy of some solutions in an extensive benchmarking of unimodal and multimodal functions. Even with regard to the algorithms used in the experimental frameworks our approach achieves positive results with less iterations and fewer individuals, although the current study has only examined a set of functions and specific values for the parameters. The improvements of PA2PSO were meaningful in the distribution of the particles in the search space for each iteration and its performance was competitive.

The PA2PSO algorithm does not aspire to be the best, by contrast we want that it will be part of the state of art of the PSO algorithms.

References

1. El Dor, A.: Perfectionnement des algorithmes d'optimisation pas essaim particulaire. Applications en segmentation d'images et en électronique. Ph.D. thesis, Université Paris-Est (2012)
2. El Dor, A., Clerc, M., Siarry, P.: Hybridization of differential evolution and particle swarm optimization in a new algorithm: DEPSO-2S. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) EC/SIDE -2012. LNCS, vol. 7269, pp. 57–65. Springer, Heidelberg (2012). doi:10.1007/978-3-642-29353-5_7
3. El Dor, A., Lepagnot, J., Nakib, A., Siarry, P.: PSO-2S optimization algorithm for brain MRI segmentation. In: Pan, J.S., Krömer, P., Snášel, V. (eds.) Genetic and Evolutionary Computing. AISC, vol. 238. Springer, Cham, Heidelberg (2014). doi:10.1007/978-3-319-01796-9_2
4. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE International Conference on Neural Networks, pp. 1942–1948 (1995)
5. Ballerini, M., Cabibbo, N., Candelier, R., et al.: Interaction ruling animal collective behavior depends on topological rather than metric distance: evidence from a field study. In: Proceedings of National Academy of Sciences, pp. 1232–1237 (2008)
6. Martin, S., Girard, A.: Continuous-time consensus under persistent connectivity and slow divergence of reciprocal interaction weights. SIAM J. Control Optim. **51**(3), 2568–2584 (2013)
7. Nedjah, N., Calazan, R.D.M., de Macedo Mourelle, L., Wang, C.: Parallel implementations of the cooperative particle swarm optimization on many-core and multicore architectures. Int. J. Parallel Prog. **44**(6), 1173–1199 (2016)
8. Poli, R.: Analysis of the publications on the applications of particle swarm optimisation. J. Artif. Evol. Appl. 1–10 (2008)
9. Rashedi, E., Nezamabadi-Pour, H., Saryazdi, S.: GSA: a gravitational search algorithm. Inf. Sci. **179**(13), 2232–2248 (2009). Special Section on High Order Fuzzy Sets
10. Shi, Y., Eberhart, R.: Empirical study of particle swarm optimization. In: IEEE International Conference on Evolutionary Computation, pp. 1945–1950 (1999)

11. Soto, D., Soto, W.: Adaptative particle swarm optimization algorithm with non-iterative electrostatic repulsion and social neighborhood. Actas de ingeniería **1**, 55–60 (2015)
12. Vanneschi, L., Codecasa, D., Mauri, G.: An empirical study of parallel and distributed particle swarm optimization. In: de Vega, F.F., Pérez, J.I.H., Lanchares, J. (eds.) Parallel Architectures and Bioinspired Algorithms. SCI, vol. 415, pp. 125–150. Springer, Heidelberg (2012). doi:10.1007/978-3-642-28789-3_6
13. de Vega, F.F., Pérez, J.I.H., Lanchares, J.: Parallel Architectures and Bioinspired Algorithms. Springer Publishing Company, Incorporated, Berlin (2014)
14. Venkata, R., Patel, V.: Comparative performance of an elitist teaching-learning-based optimization algorithm for solving unconstrained optimization problems. Int. J. Ind. Eng. Comput. **4**(1), 29–50 (2013)
15. Wang, C., Liu, Y., Zhao, Y.: Application of dynamic neighborhood small population particle swarm optimization for reconfiguration of shipboard power system. Eng. Appl. Artif. Intell. **26**(4), 1255–1262 (2013)

Fishing Monitor System Data: A Naïve Bayes Approach

Joao C. Ferreira[1,3(✉)], Serge Lage[2], Iola Pinto[2,4], and Nuno Antunes[5]

[1] Business Research Unit (BRU-IUL),
Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
joao.carlos.ferreira@iscte.pt
[2] ISEL - Instututo Superior de Engenharia de Lisboa, Lisbon, Portugal
sergelage5@hotmail.com, ipinto@deetc.isel.ipl.pt
[3] Centro Algoritmi, Universidade Minho, Guimaraes, Portugal
[4] Centro de Matemática e Aplicações, FCT/UNL, Lisbon, Portugal
[5] Xsealance - Sea Technologies SA, Leira, Portugal
nuno.antunes@xsealence.pt

Abstract. This paper discusses the results of an applied research on the fishing activity based on a monitor system developed by a company and fishing reports produced at the end of each fishing activity. Due to economic interests combined with fishing limitations there is a natural tendency for wrong reporting. We apply Data Mining (DM) methodologies to find fishing patterns. These DM techniques in SQL tool allow to deal with the high volume of this data set and determine the major factors that influence fishing activity.

Keywords: Fishing activity · Data mining · Naïve Bayes · Patterns

1 Introduction

Monitor fishing activity is necessary to control natural resources, as well as to access the fishing data produced at the end of each fishing boat trip (captured weight per type of fish). Combining these data sources, it is possible to control fishing hunting by year periods, location and weight capture. In European Union (EU) there are limit rules and the larger fish companies try to overtake this rules in order to maximize their profit. In south EU the fishing activity is intensive and Portugal has one of the biggest fishing areas in EU. With the aim of controlling this fishing activity, the Portuguese regulator created rules to oblige fishing boats to use monitor system and to report fishing capture types and the respective weights. This mechanism has led to intensive databases, big data, where the application of data mining methodologies is needed for the process of knowledge extraction. These approaches to access information database has been attracting a significant amount of interest, from both research and industry. There is an increasing need for tools to assist humans in extracting useful information from the fast growing volumes of digital data.

This paper discusses the results of an applied research on the fishing domain that used the data automatically produced from the fishing Portuguese fleet monitor system, together with the trip fishing reports. These data, around 85,000 boat trips with one

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_57



Fig. 1. Representation of the boat trips monitored in 2015

million of registers, was merged and analyzed using a data mining (DM) process. All the performed trips are illustrated in Fig. 1, corresponding to 193 998 km.

This paper is organized in five sections: Sect. 1 presents the motivation and the goals of the present research; Sect. 2 presents Monicap system; Sect. 3 presents the working methodology and describes the four main processes involved; Sect. 4 describes the data analyses and representation process; Finally, Sect. 5 presents the main conclusions.

2 Monicap System

The vessel monitoring system (VMS) is a satellite-based monitoring system, which at regular intervals, provides data to the fisheries authorities concerning the location, course and speed of the vessels. Nowadays, VMS is a standard tool used for fisheries monitoring and control worldwide, although it was the EU which led the way, becoming the first part of the world to introduce compulsory VMS, tracking for all the larger boats in its fleet. The EU legislation requires that all coastal EU countries should set up systems that are compatible with each other, so that, countries can share data and the Commission can monitor compliance with the rules. EU funding is available for Member States to acquire state-of-the-art equipment and to train their staff to use it. In Portugal was developed a solution - the Monicap bluebox - installed on board, records a set of data (position, date/time, integrity, status records) by using an integrated GPS + SBAS (EGNOS) system, guaranteeing precise position information, see Fig. 2.

Powered by the electrical system on board, it also has a smart backup battery circuit enabling it to run for several days without external power. This application allows the user to view the map any vessel, provided that they are equipped with AIS or MONICAP devices. The application also grants the user additional information about the vessels, such as, the routes performed and the captured fish (as long as there is a MONICAP device onboard). This project is an addition to the existent proprietary system developed by XSealence, the Integrated Control Center. In the year 2015 the system produced 1.048.576 registers based on 85.024 boat trips. In Fig. 3 is presented a

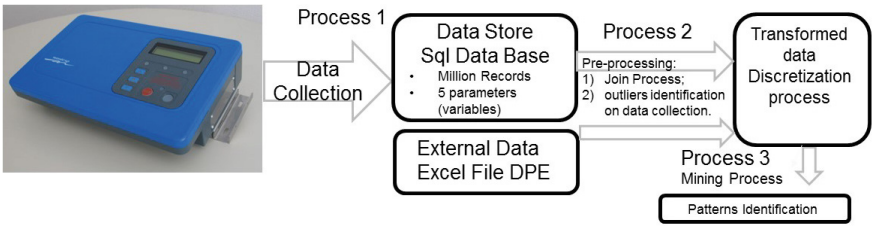


Fig. 2. Monicap System and the associated developed processes

small sample, with the following variables: (1) NAME: the name of the boat. The system was installed in 310 different boats; (2) DATETIME: Date and time in format (DD-MM-AAAA HH:MM); (3) GPS coordinates of Latitude and Longitude; (4) COURSE: represents movement direction and ranging from 0 (north) to 360. Average measurements of 138.4 with deviation of 116.55; (5) SPEED: boat speed in knots. Data ranging from 0 to 23, with mean of 1.13, and standard deviation of 2.7. These data were created automatically in opposition to the manual report produced by the boat responsible that is performed at the end of each trip.

NAME	DATETIME	LATITUD	LONGITUD	COUR	SPE
SN:000000201	26-11-2015 16:37	387.375.333	-93.042.333	1	0,8
SN:000000201	26-11-2015 16:38	387.375.667	-93.042.667	77	1,2
SN:000000201	26-11-2015 16:39	387.377.167	-93.046	197	1

Fig. 3. Data sample of Monicap registers

3 Work Methodology

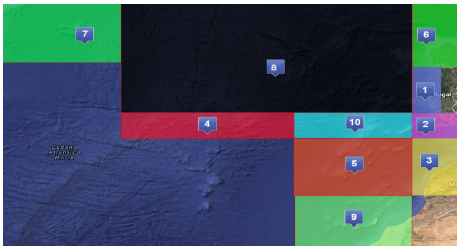
The work methodology of the research presented in this paper involves three main processes, see Fig. 2:

Process-1: Data was acquired by MONICAP system with 8 fields in a total of 1.048.576 registers based on 85.024 trips in 2015. These data were combined with manual fishing reports, which enabled the identification of outliers. Inconsistent data were also removed in an outlier's implementation approach, [1]. This process to discard outliers allowed a 20% reduction in data.

Process-2: The data pre-processing mechanism involved the identification of data in use, and was followed by data manipulation techniques towards a set of pre-defined classes, created by the discretization data process. This approach was important for the application of Naive Bayes (NB) methodology, supported by the Microsoft platform (SQL Server 2008R2). In literature, NB technique showed a better performance with discrete data [1, 5–7]. For this reason, was performed a discretization process based on the following criteria:

- Heuristics were applied based on some pre-defined criteria. One example is the time variable, where were defined 12 classes corresponding to the months of the year.
- Equal area clustering based on the method of equal areas, [2]. The main idea was to divide the data population in subclasses with approximately the same number of events. An example of this was the weight classes, where 9 classes were created, using SQL function. C_1– Weight from 0 to 12 kg with 6979 events; C_2– Weight from 13 to 29 kg with 6497 events; C_3– Weight from 30 to 49 kg with 5760 events; C_4– Weight from 50 to 64 kg with 6728 events; C_5– Weight from 65 to 99 kg with 5576 events; C_6– Weight from 100 to 150 kg with 7083 events; C_7– Weight from 151 to 273 kg with 7017 events; C_8– Weight from 274 to 499 kg with 6297 events; C_9– Weight from 500 to 1049 kg with 7018 events;

K-Means algorithm, [3], was used to create k groups of the GPS location data, and was implemented in SQL Server Data Analysis, with k = 10. The resulting output is illustrated in Fig. 4. This information was used to identify the main fishing areas.



	Name	Latitude	Longitude
Cluster 1	Costa Continental	38.9352 - 42.0752	>= -11.1321
Cluster 2	Costa Sul	36.8902 - 38.9352	>= -11.1321
Cluster 3	Entrada Mediterrano	31.1081 – 36.8902	>= -11.1321
Cluster 4	Açores	36.8902 - 38.9352	-36.6042 - -20.6570
Cluster 5	Madeira	31.1081 – 36.8902	-20.6570 - -11.1321
Cluster 6	Costa Espanha Norte	>=42.0752	>= -11.1321
Cluster 7	Atlântico NO	>=42.0752	< -36.6042
Cluster 8	Atlântico	>=38.9352	-36.6042 - -11.1321
Cluster 9	Canarias	< 31.1081	-20.6570 - -11.1321
Cluster 10	Norte Madeira	36.8902 - 38.9352	-20.6570 - -11.1321

Fig. 4. Cluster Geographic position into 10 groups

Process-3: Data analysis and representation process (detailed in Sect. 4) led to the identification of the most important variables, statistics and patterns.

This system was developed based on Microsoft SQL Server 2008R2, with SSIS (SQL Server Integration Services), SSAS (Microsoft SQL Server Analysis Services) and SEMMA (Sample, Explore, Modify, Model and Assess) [4, 5].

4 Data Analyses and Representation Process

Mainly, the patterns identification process was based on Naïve Bayes algorithm. They have been used 4 variables: TP_x – type of fish taking into account the 10 species most fished, $x = 1, \ldots 10$. P_y – fish weight, with $y = 1, 2, \ldots 9$ (9 classes defined for this variable). L_z – location where $z = 1, \ldots, 10$, represents the cluster geographic position and M_w – date, where w indicates the month number. It was intended to use NB algorithm to estimate the probabilities associated with different fish species. These conditional probabilities are, TP_X, given the characteristics, P_y, L_z and M_w, where the indexes y, z and w represents the several classes of each variable, respectively. As previously stated, for the implementation of this algorithm was used the SQL Server 2012, the process identified in [1] and adapted to this problem by

$$P(TP_x|X) = [P(TP_x) \cdot P(X|TP_x)]/P(X) \quad (1)$$

In Eq. 1 X structure includes a set of known characteristics P_y, L_z and M_w. The conditional independence assumption was used, which means that each characteristic in X is conditionally independent of any other and also independent of any subset of features in X, given TP_x, which leads to the Eq. 2:

$$P(X|TP_x) = P(P_y|\,TP_x) \cdot P(L_z|\,TP_x) \cdot P(M_w|\,TP_x) \quad (2)$$

Replacing in Eq. 1 the result stated by Eq. 2, and according to Bayes rule is obtained

$$P(TP_x|X) = [P(TP_x) \cdot P(P_y|\,TP_x) \cdot P(L_z|\,TP_x) \cdot P(M_w|\,TP_x)]/P(X) \quad (3)$$

Moreover, given the attributes in X, the term $P(X)$ in the denominator of Eq. 3 represents a scaling factor, not depending on the various fish species, TP_X, $x = 1, \ldots, 10$. Therefore, the model used can be expressed by:

$$P(TP_x|X) \propto P(TP_x) \cdot P(P_y|\,TP_x) \cdot P(L_z|\,TP_x) \cdot P(M_w|\,TP_x) \quad (4)$$

For example, we described the data regarding sardine in Fig. 5a, b, c and d. We found that sardine was fished mainly at positions 1 and 2 (Portuguese cost and south cost), from February (small fishing) to November (small fishing). It was caught in values belonging to higher classes (Classes 10, 11 and 9). This procedure can be used for any of the fish species. Figure 6 shows an overview of all probabilities, taking into account the defined variable (position, month and weight). Patterns can be found easily based on colors. Concerning position, differences are evident in the distribution of the data correspondent to different species. For example, COD and BET input data are divided only into two classes of the variable position, and GHL and SWO only have one class correspondent to the position. There are also visible differences in the distribution of data by classes of variable weight and date.

It is possible to resume major findings regarding weight, position and date. For weight and analyzing Table 1 leads to conclude that each one of the fish species has its own pattern. Taking into account the main species and the variable weight we can conclude that: (1) Black sword fish (BSF) was more fish in the higher weight class P_8, P_9 and P_{10}, with no record of lower fisheries to the P_5 and no record in P_{11} class; (2) Fish Blue Shark (BSH) was more fish in the higher weight classes, from P_7 to P_{10}; (3) Atlantic cod (COD) was more fish in higher classes to the P_5; (4) HKE (hake) was more fish in the lower classes to P_7; (5) mackerel (HOM) was more fish in classes between P_4 and P_{10};(6) freckle (MAC) was more fish in classes below P_7 class; (7) Mackerel (MAS) was more fish the higher weight classes P_{10} and P_{11}; (8) Sardine (PIL) and Redfish – Loons (RED) were more fish to the higher weight classes P_9; (9) Blue whiting (WHB) was more fish in classes between $P_2 - P_{10}$.

Major findings for the species of fish versus position : (1) BSF was more particularly fish in cluster P_2 (*costa sul*), P_5 (*Madeira*) and P_1 (*costa continental*), with small registers at positions P_0, P_4 e P_9; (2) BSH was fish mainly in P_1, with small registers at

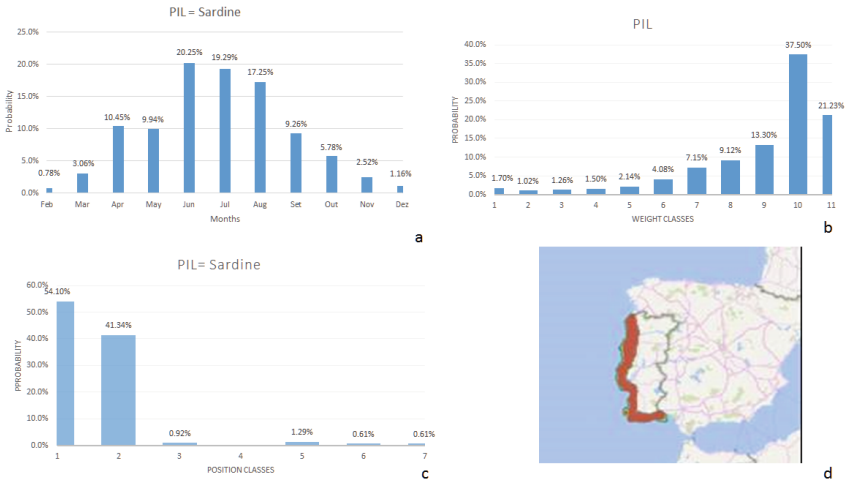


Fig. 5. In a, b and c is presented the results of the conditional probabilities correspondent to the variables date, weight and geographical position, respectively. In d is shown a graphical representation of sardine fishing locations (Portuguese coast).

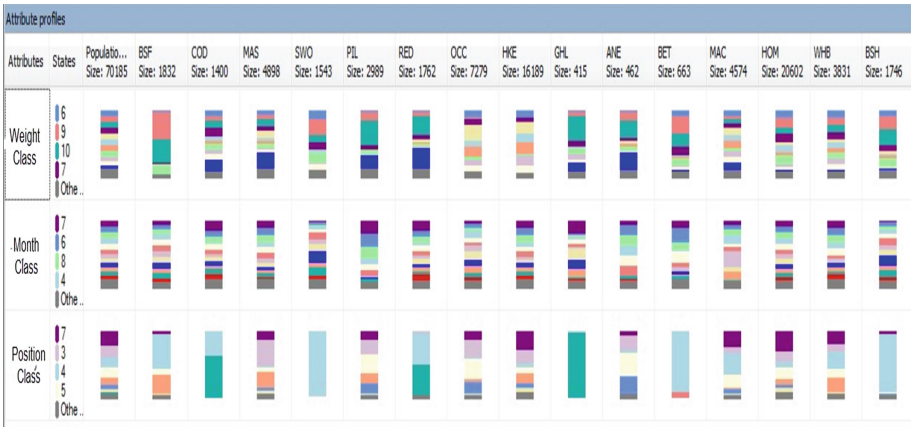


Fig. 6. NB probabilities for classes of weight, date and position

P_0, P_2, P_3, P_4, P_5, P_7 e P_{10}; (3) COD was only fished at P_7 (*Atlântico Norte*); (4) HKE was more particularly fish in class P_1 and P_2, with small registers at P_3 and P_7; (5) HOM was fished mainly in cluster P_1 and P_2, with small registers at P_3 and P_7; (6) MAC was fished mainly in class P_1 and P_7, with small registers at P_2 and P_3; (7) MAS was fished mainly in class P_2 and P_1, with small registers at P_3, P_5 and P_7; (8) PIL was fished mainly in class P_1 and P_2, with small registers at P_3, P_5, P_6 e P_7; (9) RED was fished mainly in class P_7, with small registers at P_1, P_2, P_3, P_5 and P_9; (10) WHB was fished mainly in class P_1, P_2 and P_7, with small registers at P_3. Table 2

Table 1. NB Conditional Probabilities of the weight classes, given the fish specie

Attributes	States	Population (All)	BSF	COD	MAS	SWO	PIL	RED	OCC	HKE	GHL	ANE	BET	MAC	HOM	WHB	BSH
Size		70185	1832	1400	4898	1543	2989	1762	7279	16189	415	462	663	4574	20602	3831	1746
Weight Class	1	6979	0,3%	0,2%	12,1%	0,2%	1,7%	4,7%	**23,7%**	**17,1%**	0,0%	4,3%	0,2%	11,1%	4,6%	4,4%	5,8%
Weight Class	2	6497	0,2%	5,3%	6,3%	0,6%	1,0%	3,9%	**15,5%**	**17,8%**	1,9%	4,8%	2,4%	11,6%	5,0%	8,0%	4,5%
Weight Class	3	5760	0,4%	7,9%	4,2%	4,1%	1,2%	2,7%	9,8%	15,3%	8,7%	1,9%	4,1%	9,9%	5,6%	9,8%	2,1%
Weight Class	4	6728	0,4%	6,8%	5,8%	5,7%	1,5%	1,5%	10,3%	13,8%	3,4%	1,7%	3,9%	10,5%	10,9%	9,2%	3,8%
Weight Class	5	5576	0,6%	10,4%	6,1%	9,0%	2,1%	2,0%	9,3%	12,5%	5,1%	2,6%	5,3%	9,1%	6,3%	9,4%	2,5%
Weight Class	6	7083	2,4%	9,2%	7,4%	13,5%	4,0%	3,3%	11,2%	10,8%	3,6%	3,9%	10,1%	9,6%	11,9%	11,7%	8,8%
Weight Class	7	7017	3,9%	13,4%	7,1%	12,4%	7,1%	7,2%	10,6%	7,9%	8,7%	5,6%	14,2%	8,2%	13,0%	11,7%	10,4%
Weight Class	8	6297	15,7%	9,1%	6,4%	**16,8%**	9,1%	6,4%	6,2%	3,7%	9,6%	8,4%	11,2%	6,3%	13,8%	9,1%	14,1%
Weight Class	9	7018	**40,7%**	7,1%	8,2%	**25,4%**	**13,3%**	8,3%	3,0%	1,0%	7,7%	13,0%	**25,9%**	5,7%	**15,4%**	10,9%	**19,5%**
Weight Class	10	7018	**35,2%**	11,8%	11,0%	12,0%	**37,6%**	**29,4%**	0,5%	0,1%	**37,3%**	**26,2%**	**20,1%**	7,3%	10,4%	12,6%	**24,5%**
Weight Class	11	4212	0,3%	19,0%	**25,3%**	0,3%	**21,2%**	**30,6%**	0,0%	0,0%	14,0%	**27,5%**	2,7%	10,8%	3,1%	3,3%	4,0%

Table 2. NB Conditional Probabilities of the position classes, given the fish specie

Attributes	States	Population (All)	BSF	COD	MAS	SWO	PIL	RED	OCC	HKE	GHL	ANE	BET	MAC	HOM	WHB	BSH
Size		70185	1832	1400	4898	1543	2989	1762	7279	16189	415	462	663	4574	20602	3831	1746
Position Class	1	3977	4,0%	0,0%	4,4%	0,1%	3,4%	1,7%	2,4%	8,2%	0,0%	1,3%	0,2%	3,3%	8,4%	3,7%	1,8%
Position Class	2	8366	**28,3%**	0,0%	**24,3%**	0,2%	15,5%	0,7%	5,8%	**16,8%**	0,0%	1,7%	0,0%	2,7%	10,0%	21,7%	1,0%
Position Class	3	12275	0,0%	0,0%	**39,3%**	0,0%	**22,3%**	2,8%	**26,5%**	**20,0%**	0,0%	**21,6%**	0,0%	9,6%	**16,9%**	12,0%	0,2%
Position Class	4	11942	**51,2%**	**38,1%**	3,0%	**96,3%**	1,2%	**47,5%**	2,5%	8,4%	4,1%	5,2%	**89,6%**	**32,7%**	8,7%	**26,5%**	**85,3%**
Position Class	5	10822	9,4% %	0,0%	7,1%	0,1%	**27,7%**	0,0%	**30,5%**	10,4%	0,0%	**35,1%**	0,0%	**20,4%**	**18,7%**	14,6%	3,4%
Position Class	6	5070	0,4°%	0,0%	5,9%	0,1%	16,0%	0,0%	18,7%	6,4%	0,0%	**26,4%**	0,0%	7,6%	6,4%	1,7%	2,7%
Position Class	7	15185	5,6% %	0,0%	14,0%	0,8%	12,9%	0,7%	13,3%	**29,6%**	0,0%	8,4%	0,2%	**23,5%**	**30,5%**	**19,4%**	4,5%
Position Class	9	112	1,0%	0,0%	0,0%	1,4%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	9,8%	0,0%	0,0%	0,0%	0,1%
Position Class	10	2085	0,0%	**61,9%**	0,0%	0,0%	0,0%	**46,6%**	0,0%	0,0%	**95,9%**	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Position Class	out	351	0,1%	0,0%	2,1%	0,9%	1,0%	0,0%	0,4%	0,1%	0,0%	0,2%	0,3%	0,3%	0,5%	0,4%	1,0%

Table 3. NB Conditional Probabilities of the time period classes versus fish specie

Attributes	States	Population (All)	BSF	COD	MAS	SWO	PIL	RED	OCC	HKE	GHL	ANE	BET	MAC	HOM	WHB	BSH
Size		70185	1832	1400	4898	1543	2989	1762	7279	16189	415	462	663	4574	20602	3831	1746
Month Class	1	5523	9,2%	7,5%	6,5%	11,9%	0,3%	10,4%	14,8%	5,4%	**19,0%**	1,1%	2,1%	8,3%	8,8%	3,8%	9,9%
Month Class	2	5160	7,4%	5,6%	6,2%	8,3%	0,8%	6,4%	10,9%	6,8%	14,7%	1,9%	3,8%	13,8%	6,6%	7,3%	7,4%
Month Class	3	5642	9,2%	5,6%	8,3%	8,5%	3,0%	5,4%	10,4%	5,7%	2,7%	1,9%	8,0%	**24,9%**	7,1%	4,8%	7,5%
Month Class	4	6614	9,3%	3,1%	8,2%	4,9%	10,4%	4,4%	10,7%	8,5%	4,6%	14,9%	4,1%	14,0%	10,2%	10,4%	6,8%
Month Class	5	5976	8,7%	9,9%	9,0%	4,8%	10,0%	9,5%	7,4%	8,2%	7,5%	15,2%	**18,9%**	4,7%	8,7%	12,4%	6,6%
Month Class	6	7009	7,4%	10,6%	11,3%	4,9%	**20,3%**	10,2%	6,6%	10,7%	2,7%	14,9%	**22,0%**	7,3%	9,7%	10,8%	7,4%
Month Class	7	7368	6,6%	16,1%	12,0%	2,3%	**19,2%**	15,2%	6,1%	11,6%	**21,0%**	7,4%	12,1%	8,3%	10,3%	12,3%	4,1%
Month Class	8	6799	7,6%	9,8%	11,4%	3,9%	17,1%	10,0%	4,8%	11,0%	2,9%	**16,0%**	10,7%	8,7%	9,4%	12,7%	5,8%
Month Class	9	5863	9,6%	7,9%	10,5%	11,1%	9,3%	6,9%	5,8%	9,2%	5,1%	**16,9%**	6,5%	7,0%	8,2%	6,4%	11,1%
Month Class	10	5086	10,1%	10,0%	8,0%	14,2%	5,9%	5,8%	7,1%	8,5%	2,2%	5,6%	4,7%	2,2%	6,9%	4,8%	12,2%
Month Class	11	5379	9,5%	6,9%	5,6%	19,1%	2,5%	5,6%	8,9%	8,4%	13,7%	1,7%	5,1%	0,7%	8,2%	6,7%	**16,0%**
Month Class	12	3766	5,5%	6,9%	2,9%	6,1%	1,2%	10,1%	6,3%	6,1%	4,1%	2,4%	2,1%	0,1%	6,0%	7,6%	5,3%

shows these findings. After the class 10 one class has been defined as "out", in order to show positions not defined in the ten identified clusters.

Regarding year period, it was found: (1) BSF and HOM were caught identically throughout the months of the year; (2) BSH was more fish in September, October and November; (3) COD was more fish in May, June and July; (4) HKE was more fish in June, July and August; (5) MAC was more fish in February, March and April, with a sharp drop in October, November and December; (6) MAS was more fish from March to October; (7) PIL was more fish in June, July and August, with a marked fall from November–March; (8) RED was more fish in July; (9) WHB was more fish from April–August, see Table 3.

5 Conclusions

In this work patterns were identified in fishing activities. Analysis and identification of well-defined standards was carried out for the main species of fish. We point as major findings: (1) fishing patterns towards fraud detection, on manual fishing reports created by boat fish Capitan. Records that showed a deviation from the standards can be selected for authority inspection and then reduce the tendency for fraud, taking into account that these standards are based on previous fishing data. (2) k-means was used to identify 10 fishing zones. This number 10 was defined by Portuguese fishing authority. The identification of these areas may be important for authorities in their control processes. Based on this work it is possible, in a next step (future work), to identify the fishing activity, Dredges, Trawls, Hooks and Lines, Traps, Nets and Seines. This can be achieved with similar work taking into account registers of the boot speed, direction and position.

References

1. MacLennan, J., Crivat, B., Tang, Z.: Data Mining with Microsoft SQL Server 2008–2009. Wiley Publishing, Indiana (2009). ISBN 978-0-470-27774-4
2. Lachev, T.: Applied Microsoft Power BI: Bring Your Data to Life. Prologika (2015). ISBN13: 9780976635369
3. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 881–892 (2002)
4. Winston, W.L.: Microsoft Excel 2013 Data Analysis and Business Modeling. O' Reilly Media, inc. (2014). ISBN: 978-0-7356-6913-0
5. Azevedo, A., Santos, M.F.: KDD, SEMMA and CRISP-DM: a parallel overview. In: Proceedings of the IADIS European Conference on Data Mining 2008, pp. 182–185 (2008)
6. Ng, A.Y., Jordan, M.I.: On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: Conference on Neural Information Processing Systems (2002)
7. Webb, G.I., Boughton, J., Wang, Z.: Not so Naive Bayes: aggregating one-dependence estimators. Mach. Learn. **58**(1), 5–24 (2005). Springer

Detection of Behavioral Patterns for Increasing Attentiveness Level

Dalila Durães[1(✉)], Sérgio Gonçalves[2], Davide Carneiro[3,4], Javier Bajo[1], and Paulo Novais[3]

[1] Department of Artificial Intelligence, Technical University of Madrid, Madrid, Spain
d.alves@alumnos.upm.es, jbajo@fi.upm.es
[2] Informatics Department, University of Vigo, Ourense, Spain
sgoncalves@alumnos.uvigo.es
[3] Algoritmi Center, Minho University, Braga, Portugal
{dcarneiro,pjon}@di.uminho.pt
[4] CIICESI, ESTGF, Polytechnic Institute of Porto, Felgueiras, Portugal

Abstract. In the current world, performance is one of the most important issues concerning work and competition. Performance is strongly connected with learning and when it comes to acquiring new knowledge, attention is one the most important mechanisms as the level of the learner's attention affects learning results. When students are doing learning activities using new technologies, it is extremely important that the teacher has some feedback from the students' work in order to detect potential learning problems at an early stage. The goal of this research is to propose a system that measures the level of attentiveness in real scenarios, and detects patterns of behavior associated to different attention levels among different students. This system measures attention and uses this information for training a decision support system that shows the level of attention of a group of students in real time.

Keywords: Ambient intelligence · Learning activities · Innovative scenarios · Attentiveness

1 Introduction

Education, training, skills development, and learning are processes that are continuously performed since we are born. These characteristics and the ability to learn and teach, allow us to grow as a person. Besides, learning is inevitably linked to the History of Humanity, to its construction as a social being capable of adapting to new situations. The learning process is described as the act of acquiring knowledge, behaviors, abilities, standards or preferences and the study of learning has been closely linked to the development of psychology as a science [1].

Many contemporary educators argue the value of a constructivist approach to teaching. Advocates argue that constructivist learning better equips learners to successfully master new and novel situations [2]. It emphasizes problem solving over the mere

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_58

accumulation of facts. One of the central arguments for the use of Web-based resources in the classroom is that it gives learners access to information resources in ways that allow them to search for relevant data, synthesize that information, and draw their own conclusions.

For these reasons, the school must create systems that are capable of involving the students and capture their attention. Advances in computers and wireless technologies have also had an impact on the educational setting, thus generating a new approach for Technology-Enhanced Learning (TEL). The rapid development of these technologies combined with access to content in a wide variety of settings, allows learners to experience new learning situations beyond the school's walls [3].

The concept of adaptively refers to the idea that different learning contexts should be adaptable to the particular setting in which the learner is situated [3].

Being a cognitive process, attention is strongly connected with learning [4]. When it comes to acquiring new knowledge, attention can be considered one of the most important mechanisms [5]. The level of the learner's attention affects learning results. The lack of attention can define the success of a student. In learning activities, attention is also very important to perform these tasks in an efficient and adequate way.

There is no universally accepted definition of attention because there is a diversity of disciplines that study it. In the past, only psychologists studied attention; however attention is currently highly important for fields such as philosophy, chemistry, anatomy, and even computational science [6]. With the existence of these multiple sets of disciplines that study attention, its definition diverges depending on the field of study. For example, we can differentiate the concept of attention in human beings and machines. In humans, attention is processed in the brain while in machines there is a processor unit with a certain memory capacity that will process data. As with the brain, computers should analyze more and more data, but unlike the brain they do not or do rarely, "pay attention" [7] to the data.

Attention is the cognitive process of focusing on one aspect of the environment while disregarding others. "Pay attention!" is a phrase repeated by so many teachers all over world, and the explanation for this is simple [7]. Attention is the first step in the learning process. Students cannot learn or understand or even remember if they do not listen properly, so, they will fail in the learning process. For almost everyone it is easy to pay attention to subjects or things that are interesting or stimulating to them.

So the target of this paper is to propose the architecture of an ambient intelligent system aimed at capturing and measuring the level of the students' attentiveness in real scenarios and dynamically provide recommendations to the teacher. The students can learn in a relatively anonymous environment without the embarrassment of failure and/or socio-cultural prejudice from personal contact. All the students get the same standardized set of learning materials from the learning activities using technologies in an online environment. This environment is collaborative. The use of groups and teams working together in collaborative learning contexts allow interactions which enforce employability skills. Teaching other subjects they have learned help them to reinforce the learning process. The student usage of the learning materials can be monitored and early possible dropout can be discovered and counteractive support can be offered. The online environments can be used to give instant feedback from online self-assessment

or formative assessment particularly through multiple choice question formats. In addition, the students can be offered automatically marked self-assessment exercises to ascertain skill/knowledge levels and learning needs before engaging with course content. It can also assist students with certain disabilities.

2 Proposed Design

When students are affected by positive or negative states, they produce different kinds of thinking and this might hold important implications on the educational and training perspective. When that occurs it would be important to be able to notify and advise the teacher, so that she/he can be able to dynamically modify the teaching style according to students' feedback signals which include cognitive, emotional and motivational aspects.

While the student conscientiously interacts with the system and takes his/her decisions and actions, a parallel and transparent process takes place in which the TEL system uses the information. The architecture of the TEL system presented in Fig. 1 depicts the process through which the system operates.

The devices in which students work have software that generates raw data. These devices store the raw data locally until it is synchronized with the web server in the cloud, which occurs at regular intervals (normally 5 min).

In the storage layer the raw data received from the data generating devices is stored in a data store engine. The analytic layer provides powerful tools for performing analytics and analyses in real-time. The system calculates, at regular intervals, an estimation of the general level of performance and attention of each student.

The classification layer is where the indicators are interpreted. For example, interpreting data from the attentiveness indicators and building the meta-data that will support decision-making. When the system has a sufficiently large dataset that allows making classifications with precision, it will classify the inputs received into different attention levels in real-time. This layer has access to the current and historical state of the group from a global perspective, but it can also refer to each student individually creating each student learning profile.

With the results concerning the level of attention and the alternative strategies applied during a long period of time by the teacher, it is possible to obtain a profile of the learning style. This meta-data will allow the system to create the indicators used to perform decision-making.

The classification is also a very important aspect to have control of. This possibility allows analyses within longer time frames. This information will be used by another sub-module, the affective adaptive agent, to provide relevant information to the platform and to the mentioned personalization module.

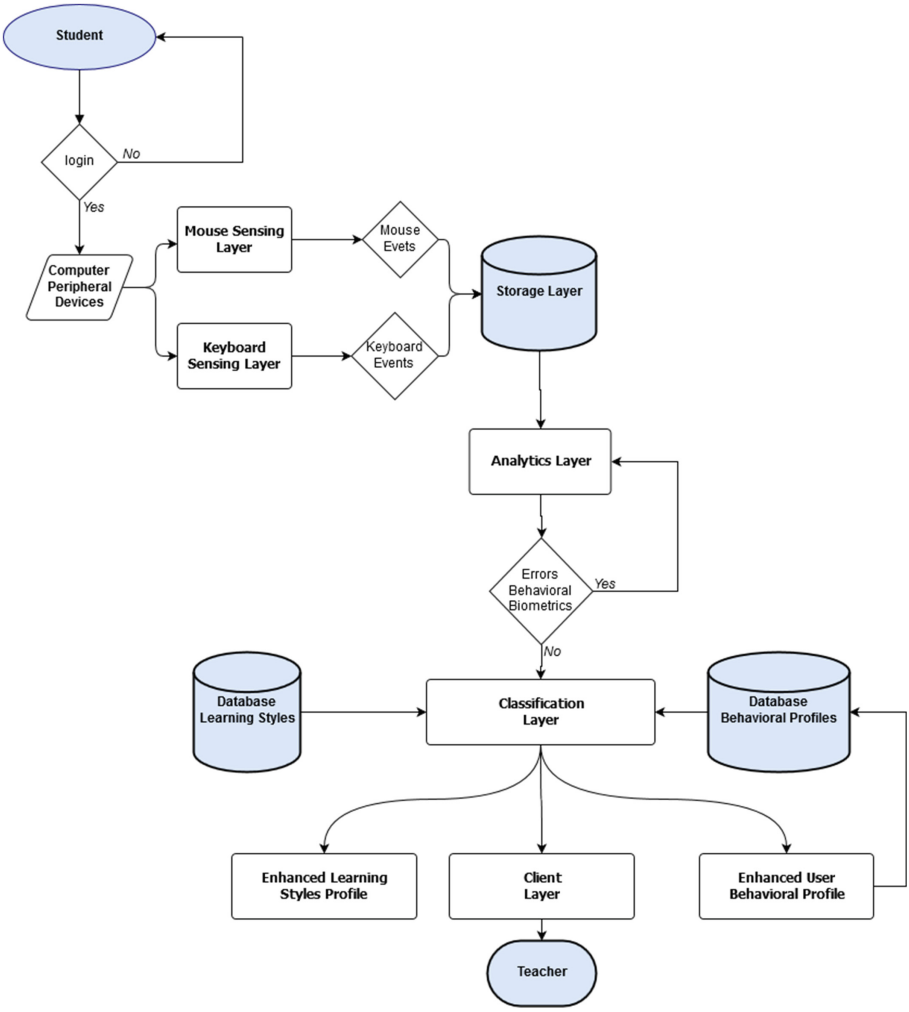


Fig. 1. Diagram of the system.

Finally, the Client layer is developed as a web app with an intuitive and visual representations (diagrams and graphical) of the attentiveness state of the group and of each student, abstracting from the complexity of the data level where they are positioned. At this point, the system can start to be used by the people involved, especially the teacher who can better adapt and personalize her/his teaching strategies. With a focus on individual and group performance and using real time analytics, the visual representation tools are intuitive, suggesting and facilitating the decision-making and human resources management. The actual students' attention information is displayed in the visualization layer, and can be used to personalize instruction according to the specific student, enabling the teacher to act differently with different students, and also to act differently with the same student, according to his/her past and present level of attention.

2.1 Methodology

In this work we compare two classes from the electronics vocational courses while performing an activity based on serious games at the High School of Caldas das Taipas, Guimarães, Portugal. We want to determine if classes from different level have a significant effect on mouse and keyboard dynamics, and attention level. With these results the TEL system can support future decision-making.

For this purpose, two groups of 10 (12I, all male students) and 7 (11J, all male students) electronic vocational students were selected to participate, whose average age is seventeen (12I) and sixteen (11J) years old. In the same week, they have a serious game lesson, where they have access to an individual computer and nineteen minutes to complete a task. For the 12I class, the lesson started at 8:30 and finished at 10:00 a.m, and for the 11J class, the lesson started at 11:30 a.m. and finished at 01:00 p.m. Students received, at the beginning of the lesson, a document with the goals of the task.

The first stage of the proposed system is the collection of the relevant data, which was designed and carried out using a logger application developed in previous work [8]. The data collected by the logger application characterizing the students' interaction patterns is aggregated in a server to which the logger application connects after the student logs in. The privacy of the students is ensured, since the necessary data that is collected to perform the login and the registration are an ID that does not personally identify the student, password, and the gender.

This application runs in the background, which makes the data acquisition process, a completely transparent one from the point of view of the student. The system was developed to acquire data from normal working routines, compiling information from the students' activities through the mouse and the keyboard, which act as sensors. The Mouse and Keyboard Sensing layers are responsible for capturing information describing the behavioral patterns of the students, and receiving data from the events generated by their mouse and keyboard [8]. These data are further processed, stored and then used to calculate the values of the behavioral biometrics.

2.2 Data Analysis

In this sub-section we will show the existence of different behaviors in the two different classes. Data was analyzed in two different ways. First, a general analysis was carried out in which statics methods are used to obtain preliminary conclusions. Second, an individual analysis was made in order to compare the different moments.

In a preliminary analysis of data, it was concluded that there are indeed different interaction patterns depending on the group analyzed. To conclude this, we looked at the distributions of the data collected and analyzed the statistical significance of their differences. To this end, we used the Kruskal-Wallis test. Table 1 details the mean value of each feature in each class (evaluation and normal). It also details the p-value of the Kruskal-Wallis test.

Table 1. Results of the Kruskall-Wallis test and mean values for each class and each feature.

Feature	Significance	12I	11J
kdt	0.17397	158.60	348.28
tbk	0.00013	867.95	483.58
mv	5.63×10^{-10}	0,49	0,59
ma	6.40×10^{-8}	0,53	0,61
cd	6.10×10^{-21}	93,95	509,40
tbc	3.39×10^{-10}	3981,61	2352,06
dbc	0.569996	213,39	206,13
ddc	4.93×10^{-21}	81,93	272,36
edbc	0.004164	395,33	306,82
aedbc	5.53×10^{-11}	2,84	1,89
ssdbc	0.0037082	-0,46	-10,03
asdbc	1.41×10^{-6}	6230,54	4557,29
tdc	0.632319	57,73	49,08
dplbc	2.41×10^{-6}	$1{,}543 \times 10^{7}$	$2{,}132 \times 10^{7}$
adplbc	2.09×10^{-13}	83112,50	170520,00

When data from the two classes is compared, the first conclusion is that the differences observed are statistically significant in nearly all features. Moreover, mean values of the features are consistently lower in the 12I class. In most of the features, this indicates an increased performance (e.g. a smaller average distance between clicks means that the student moved the mouse in a more efficient manner). In the case of mouse velocity, for example, a smaller velocity could point out a slower, and thus less efficient, movement.

In past work we concluded that a slower mouse velocity is indeed necessary for the student to achieve increased accuracy in mouse movement: moving the mouse too fast would make precise movements more difficult to carry out. A similar trend happens with mouse acceleration and click duration. The remaining features consistently show increased performance in the 12I class. Also for the keyboard features, the data are consistent with the features extracted from the mouse.

We analyzed all the applications used by all students and labelled each one of them as belonging to the task or not. We then quantified the amount of time that each student spent interacting with applications related to the task versus other applications.

For that, it is necessary to compute the level of attention of the user as detailed in Algorithm 1.

```
Data:
t - A list of triplets of the type (AppName, Timestamp,
Duration)
st - The starting time of the task
inter - The interval to update attention
set - the set of regular expressions
Class1=12I
Class2=11J
Task1={"*codecombat*", "*questionario*"}
Task2={"*code*", "*questionario*"}
Result: attention - A list of triplets of the type
(timestamp, attention%, others%)
attention←[];
work1←0;
others1←0;
work2←0;
others2←0;
time←st;
for (i←0;i < Size(t); i++)
  if (isWork(t_{i;1}, set)) then
    work1←work1 + t_{i;3};
  else
    others1←others1 + t_{i;3};
  end
if (t_{i;2} > time + inter) then
    AppendTo(attention, t_{i;2}, work1*100/(work1 + others1),
others *100/(work1 + others1));
Work1←0;
Others1←0;
time←t_{i;2};
end
end
for (i←0;i < Size(t); i++)
  if (isWork(t_{i;1}, set)) then
    work2←work2 + t_{i;3};
  else
    others2←others2 + t_{i;3};
  end
  if (t_{i;2} > time + inter) then
    AppendTo(attention, t_{i;2}, work2*100/(work2 + others2),
others2 *100/(work2 + others2));
    Work2←0;
Others2←0;
time←t_{i;2};
  end
end
```

To do this we measure the amount of time, in each interval, that the user spent interacting with work-related applications. The algorithm thus needs knowledge about the domain in order to classify each application as belonging or not to the set of work-related applications. This knowledge is provided by the teacher and is encoded in the form of regular expressions. The teacher uses a graphical interface to set up rules such as "started with 'codecombat'" or "contains 'code'", which are then translated to regular expressions that are used by the algorithm to determine which applications are work-related and which are not.

Whenever an application that does not match any of the known rules for the specific domain is found, the application name is saved so that the teacher can later decide if a new rule should or should not be created for it. By default, applications that are not considered work-related are marked as "others" and count negatively towards the quantification of attention.

3 Preliminary Results

During the lessons with the two classes, the monitoring system was used to assess the interaction of the students with the computer and to quantify their level of attentiveness as well. To quantify attentiveness the following methodology was followed. Asides from capturing the interaction of the students with the computer, the monitoring system also registers the applications with which students are interacting.

Attention is calculated at regular intervals, as configured by the teacher (e.g. five minutes). The teacher may also want to assess, in real-time or a posteriori, the evolution of attention of the whole class. Figure 2 presents the evolution of attention for each class in a lesson. If necessary, the teacher may also click on a student to analyze the temporal evolution of the attention for that specific student during the class.

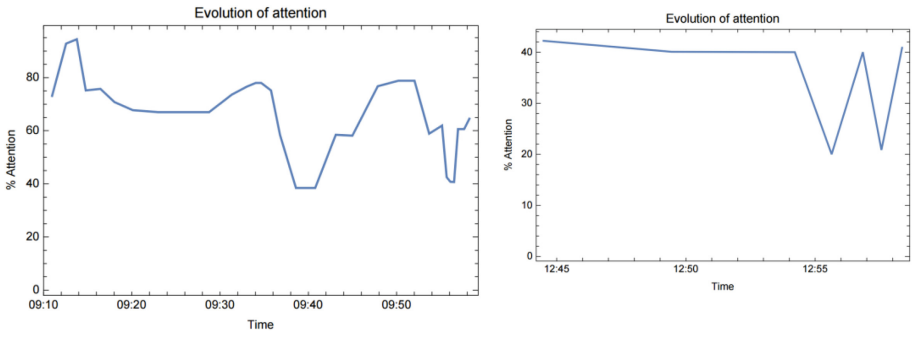


Fig. 2. Temporal evolution of attention in the two class (a) 12I (b) 11J.

The correlation between the level of attentiveness and the score of the students in the task in each class was also analyzed. In the two classes, there are positive correlations: 0.875 in 12I and 0.90 in 11J. We believe that these values are so high for two reasons: (1) these classes are mostly composed of two very different types of students (bad and

excellent students) and (2) the task was of high difficulty. Finally, there were many differences in the scores. In future work we will study this relationship in more detail, namely in different classes and with tasks of different levels of difficulty. Table 2 presents the percentage of work and the results obtained by the students in the two classes.

Table 2. Percentage of work and the scores obtained by the students: left - 12I; right - 11J

Student	% Attention	Results	Student	% Attention	Results
T2210001	78,2417	14	**T7110061**	0,0648998	6
T2210002	62,6977	12	**T7110062**	48,093	6
T2210003	52,1404	12	**T7110064**	95,9514	14
T2210004	82,0908	15	**T7110066**	84,2233	13
T2210006	49,5761	8	**T7110068**	4,21712	8
T2210007	44,2481	8	**T7110072**	100	14
T2210008	86,6918	13	**T7110073**	6,19405	6
T2210009	98,7981	16			
T2210010	94,3469	17			
T2210012	55,3309	13			

4 Limitations and Future Work

The work developed so far resulted in a very useful system for teachers to monitor, in real-time, the level of attention of their students. However, limitations were also identified. In fact, if a student opens an application that is not work-related and does not interact anymore with the computer until the end of the task a level of attention of 0% is calculated. Although the student was not actually carrying out the task, she/he was also probably not even standing in front of the computer. Lack of attention and absence must needs be distinguished. Similarly, if the student opens a work-related application and does not interact with the computer after that, the attention is quantified as 100% when she/he was most likely not even at the computer. These cases must, evidently, be pointed out. In order to correctly validate the efficacy of the system it is necessary to observe the amount of time that each student devoted to work and to other activities, but also the amount of time actually spent interacting with the computer.

To address this limitation, in future work we will implement a tighter integration between the previous and the new work. Specifically, in previous work we implemented the monitoring of the interaction of the users with the keyboard and the mouse. It is thus possible to know all the actions that each user performed both with the mouse or the keyboard, and at what time. We will thus generate a new feature that will quantify the level of activity of each user throughout time. This new feature will allow a more contextualized analysis of attention, improving the performance of its classification and quantification.

In this moment, the architecture of the proposed TEL system can only analyze the student's level of attention from the percentage of time spent interacting with work-related applications. A larger amount of data is necessary in order to analyze the profile of students and predict how each one will react in different types of lessons.

Regarding learning styles, the system only analyses the student's actions by the percentage of work-related tasks and the interaction with the mouse and the keyboard. It is possible to apply a preliminary questionnaire in order to detect the learning style of each student. When the system has enough data for each student, it will be possible to advise the teacher with the aim to improve the attention level. It will also be possible to analyze the students' profile, taking into account their individual characteristics, and to propose new strategies and actions. Given that the teacher is informed about the behavior of each student and each one's learning style, she/he will be able to maximize students' attention and, consequently, the performance of the teaching-learning process.

With the proposed system it is possible to detect potentially negative factors dynamically and non-intrusively, making it possible to foresee negative situations, and to take actions to mitigate them. This may, in turn, minimize issues such as stress and anxiety, which can negatively influence the students' results and are closely related to the occurrence of conflicts.

The next step in this work is to integrate smartphones and tablets, taking advantage of their new features such as several incorporated sensors and high resolution cameras. This may allow a wider characterization of the student, making it possible to enhance the learning experience, through better recommendation and personalization.

Acknowledgment. This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013.

References

1. Carroll, J.M.: Human-computer interaction: psychology as a science of design. Ann. Rev. Psychol. **48**, 61–83 (1997). doi:10.1146/annurev.psych.48.1.61
2. Clark, R.E.: The transfer of technology between nations. Interaction **1**(3), 185–203 (1983)
3. Milrad, M., Wong, L.H., Sharples, M., Hwang, G.-J., Looi, C.–.K., Ogata, H.: Seamless learning: an international perspective on next generation technology enhanced learning. In: Berge, Z.L., Muilenburg, L.Y. (eds.) Handbook of Mobile Learning, pp. 95–108. Routledge, New York (2013)
4. Pimenta, A., Gonçalves, S., Carneiro, D., Riverola, F., Novais, P.: Mental Workload Management as a Tool in e-Learning Scenarios (2014)
5. Hwang, K., Yang, C.: Automated inattention and fatigue detection system in distance education for elementary school students. J. Educ. Technol. **12**, 22–35 (2009)
6. Mancas, M.: Computational Attention – Toward Attentive Computers. Press Universitaire de Louvain, Louvain-la-Neuve (2007)
7. James, W.: Principles of Psychology, vol. 1, pp. 255–260. Dover Publications Inc., New York (1890)
8. Durães, D., Carneiro, D., Bajo, J., Novais, P.: Using computer peripheral devices to measure attentiveness. In: Advances in Intelligent System and Computing, pp. 147–155 (2016)

Towards a Medical Intensive Care Unit Decision Support System Based on Intuitionistic Fuzzy Logic

Hanen Jemal[1(✉)], Zied Kechaou[1,2], and Mounir Ben Ayed[1,3]

[1] National School of Engineers (ENIS), REGIM: REsearch Groups on Intelligent Machines, BP 1173, 3038 Sfax, Tunisia
{hanen.jemal, zied.kechaou, mounir.benayed}@ieee.org
[2] Computer Sciences Department, Higher School of Commerce (HSC), University of Sfax, Sfax, Tunisia
[3] Computer Sciences and Communication Department, Faculty of Sciences of Sfax (FSS), University of Sfax, Sfax, Tunisia

Abstract. Intensive Care Unit (ICU) medical processes can be so complex and unpredictable that physicians sometimes must make decisions based on perception. Both decision support system and Intuitionistic Fuzzy Logic (IFL) techniques can assist doctors to handle this complexity in a safe, harmless and efficient manner. To this end, we propose a prototype called Medical Intuitionistic Fuzzy Expert Decision Support System (MIFEDSS) based on IFL and the Modified Early Warning Score (MEWS) standard score. Moreover, the experimental results have been shown the efficiency of the proposed system.

Keywords: Intensive care unit · Decision support system · Accuracy · Modified early warning score · Medical intuitionistic fuzzy expert decision support system

1 Introduction

In real world tasks, we recurrently deal with vague information. Accessible information is occasionally vague, inexact or inadequate. Intuitionistic Fuzzy Sets (IFS) have been found to be practical to deal with ambiguity.

Intuitionistic Fuzzy Logic (IFL) affords a means for encapsulating the medical ambiguity in an appropriate algorithm for computer implementation. Thus, Intuitionistic fuzzy logic has a particular advantage especially suited for use in supporting Intensive Care Unit (ICU) medical decision making. This will support the executive of complex treatments in the ICU, decrease variability among physicians and aid us in attaining clinical endpoints. For that, we assume that IFL presents a very practical tool in various applied areas of medical science and healthcare technology. Intending to model the ICU process we propose in this study, the modeling and the implementation of Medical Intuitionistic Fuzzy Expert Decision Support System (MIFEDSS) this prototype is tested at Polyclinic ESSALEMA, Sfax, TUNISIA.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_59

This paper is organized as follows: At first we started with an introduction, subsequently Sect. 2 presents existing studies of fuzzy logic (FL) and IFL in healthcare. Section 3 presents theoretical background about FL and IFL. Then, we divulge the MIFEDSS proposed methods. Section 5, we presents the implementation and results details section and then we wind up by a conclusion with further suggestions for additional advancement.

2 Related Works

In the literature there are various studies used Fuzzy Logic (FL) for healthcare care were grouped into six medical fields: (i) Neuro domain [1], (ii) Blood glucose [2–5], (iii) Neck cancer [6], (iv) Breast cancer [7, 8], (v) Brain tumor [9] and (vi) Emergency decision system making [10].

Ever since Atanassov instigate the concept of IFS [11], this method has been utilized in several medical care case studies research. These studies concern the use and the comparison of different IFL measures of similarity in medical case study such as: the distance between interval-valued IFS, Mold Cosine similarity measure and the max-min-max composition rule. In fact, in [12] researchers use the max-min-max composition rule to identify the illness. In the same way [13], authors present an approach for medical healthcare diagnosis based on the smallest distance between symptoms. In [13] works, illness is introduced through different syndromes without aggregation of symptoms, to solve this problem, other studies [14–16] define a new interval-valued IFS measure based on the aggregation. In 2013 [17], authors survey other distance measures (Hamming distance, Euclidean distance, Normalized Hamming distance and Normalized Euclidean distance) as tools in pattern detection for IFL. In 2015, another IFL measure has been proposed, based on Mold Cosine similarity measure projection formula and present four kinds of pattern for Atanassov's intuitionistic fuzzy Values [18]. Table 1 summarize the existing works of IFL in medical domain and present the positive features of the MIFEDSS prototype.

It is inferred from above literature on IFL (Table 1), that the majority of these studies exhibits the IFL measures technique with a hypothetic medical case study. In the medical practice, at date only one the study of [19] introduces an intelligent method in bacteria classification issue based on IFL. Also we can assume that FL and IFL are successfully applied in several medical fields for detecting, supervising and monitoring of diverse diseases.

In this study we define an hybrid approach which combines Multi Agent System (MAS) and IFL decision making system [20, 21] tested in a real environment (Polyclinic ESSALEMA –Sfax-TUNISIA). Our system offers a model that combines the benefits of both MAS and IFL. The architecture is composed of a set of autonomous agents adapted to the interaction: the expert agents with IFL based software learning in order to assist agent doctor. In this paper, we present only the modelization and the evaluation of the intuitionistic fuzzy logic technique in medical decision (expert agents learning processes in the MAS organization).

Table 1. The contributions of existing works and the position of MIFEDSS system.

Authors	Contribution level
Jeong et al. [12]	A new approach for medical diagnosis using the distance between interval-valued intuitionistic fuzzy sets
Eulalia et al. [13]	A new approach for medical diagnosis by employing intuitionistic fuzzy Sets solution is obtained by looking for the smallest distance between symptoms
Pathinathan et al. [14]	A new technique to diagnose the symptom of the disease using Interval Valued Intuitionistic fuzzy set (IVIFS) with logical operators
Mohammed [15]	A new approach for medical diagnosis by interval valued intuitionistic fuzzy sets
Chetia et al. [16]	Present approach for medical diagnosis using interval-valued fuzzy soft sets and exhibit the technique with a hypothetical case study
Hoda et al. [17]	Present four new distance measures as tools in pattern recognition for intuitionistic fuzzy sets (IFSs)
Boquan et al. [18]	Propose a Mold Cosine similarity measure projection formula to medical pattern recognition and intuitionistic fuzzy decision making
Vahid [19]	An intelligent approach towards the bacteria classification problem by using five similarity measures of fuzzy sets (FSs) and intuitionistic fuzzy sets (IFSs) to examine their capabilities in encountering uncertainty in the medical pattern recognition
MIFEDSS contribution	A new Medical Intuitionistic Fuzzy Expert Decision Support System based Modified Early Warning Score (MEWS) in Intesive Care Unit (Plolyclinic ESSALEMA –Sfax)

3 Preliminaries

The beginnings of Fuzzy Logic (FL) go back to 1965 when Zadeh define FL "*As the complexity of a system increases, our ability to make precise and significant statements about its behavior diminishes until a threshold is reached beyond which precision and significance become mutually exclusive characteristics*" [22].

Let X is asset, with a generic element of X represented by x that is X = {x}. FS is defined in [23] as:

$$A = \{x, \mu A(x) | x \in X\} \tag{1}$$

Where $\mu A(x)$: X → (0, 1) is the membership function of the F Set $\mu A(x) \in (0, 1)$ is the degree of membership of the element x to the set A.

The intuitionistic fuzzy sets (IFSs) were introduced by Atanassov [24] as a generalization of fuzzy sets of Zadeh [22] "which look more accurately to uncertainty quantification and provides the opportunity to precisely model the problem based on the existing knowledge and observations", where besides the degree of membership $\mu A(x) \in [0,1]$ of each one element $x \in X$ to a set A, the degree of non-membership $\gamma A(x) \in [0,1]$ was also measured.

Let X is a non-empty fixed set. An intuitionistic fuzzy set (IFS) A is an object of the form:

$$A = \{x, \mu A(x), \gamma A(x) | x \in X\} \tag{2}$$

Where μ and γ are degrees of membership and non-membership of each $x \in X$, respectively, and $0 \leq \mu(x) + \gamma(x) \leq 1$ for each $x \in X$. A class of all the IFS's in X is denoted as IFS(X).

In addition to membership and non-membership functions, a function of hesitancy or uncertainty of x to A denoted by $\pi A(x)$ must be taken into consideration. $\pi A(x)$ is computed as:

$$\pi A(x) = 1 - \mu A(x) - \gamma A(x) \text{ with } 0 \leq \pi A(x) \leq 1 \tag{3}$$

The illustration of these degrees has been demonstrated in (Fig. 1):

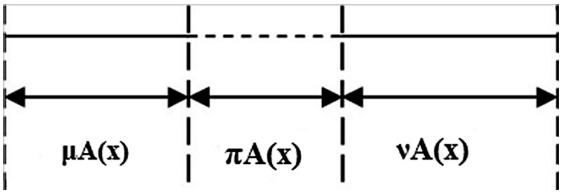


Fig. 1. IFS degrees of membership, non-membership and hesitation part

4 MIFEDSS Modelization

We define MIFEDSS as an intuitionistic fuzzy relation R from the physiological parameters P to the set of diagnoses D which present degree of relationship and degree of non relationship between physiological parameters and diagnosis. The methodology of IMFS contains three steps:

1. Define the medical physiological parameters.
2. Formulation of medical Intuitionistic Fuzzy Set knowledge.
3. Define the diagnosis (degree of risk).

In this study, the IFL are founded on the Modified Early Warning Score (MEWS), defined as a clear guide for caregivers in the emergency unit to find the level of sickness of a patient [25]. The Modified Early Warning Score (MEWS) is a physiological score for estimation and is based on five physiological parameters (Table 2).

The MEWS is unable to model and present the hesitancy of the complex situation due to insufficient facts, information lost, and ambiguity. To solve this problem, we suggest a novel extension of the MEWS model which is based on the IFL theory. A human expert knowledge, from Polyclinic ESSALEMA, was used for the determination of the input fields: systolic blood pressure (SBP), heart rate (HR), blood glucose (BG), patient temperature (TEMP), oxygen saturation (O2S) and Kalmy (KAM).

Table 2. MEWS standard [25]

MEWS	+3	+2	+1	0	+1	+2	+3
Systolic blood pressure	<70	70–80	81–100	101–199		≥ 200	
Heart rate		<40	41–50	51–100	101–110	111–130	>130
Respiratory rate		<9		9-14	15-20	21–29	≥ 30
Temperature		<35	35.1–36	36.1–38	38.1–38.5	>38.5	
AVPU/GCS score	<9	9-13	14	A/15	V/Confused	P	U

The score is calculated by measuring the five parameters as above and adding together the assigned score for each physiological value. AVPU = Alert, Verbal, Pain, Unresponsive! GCS = Glasgow Coma Scale

We add the Kalmy input to the MEWS shown in Table 2, in order to calculate the complete score Kalmy input. We dispense score for each parameter from 0 to 3 presented by different color ("Table 3").

Table 3. Intuitionistic MEWS system

Modified Early Warning Score MEWS							
Risk Level of risk	Low 3	Low 2	Low 1	Normal 0	High 1	High 2	High 3
sbp	SBP<75	70<SBP<85	80<SBP<100	95<SBP<199		SBP>185	
HR		HR<40	45<HR<60	53<HR<100	95<HR<110	105<HR<130	HR>125
O2S	O2S<85	83<O2S<90	87<O2S<92	O2S>92			
TEMP		T<36.5		36<T<38.5		T>38	
BG	BS<66	63<BS<72	-	70<BS<110	-	106<BS<150	BS>140
Kalmy		KLM<3.7		3.6<KLM<5		KLM>4.9	

The typical process (Fig. 2) of intuitionistic fuzzy logic is clarified in six steps:

1. Identify the linguistic variables: such as Heart rate, Temperature… provided by the medical caregivers of ICU.
2. Construct the membership functions of each linguistic variable such as trapezoidal, triangular…
3. Construct the rule base throw the aid of specialist medical expert knowledge.
4. convert the input to fuzzy values by the membership functions
5. The expert agent evaluate the results of each rule: calculate degree of membership, degree of non membership and hesitation part.
6. Convert the output variable to non fuzzy values: the expert agent transmits the degree of risk (normal, large and high) to the doctor agent to provide the suitable treatments to the patient.

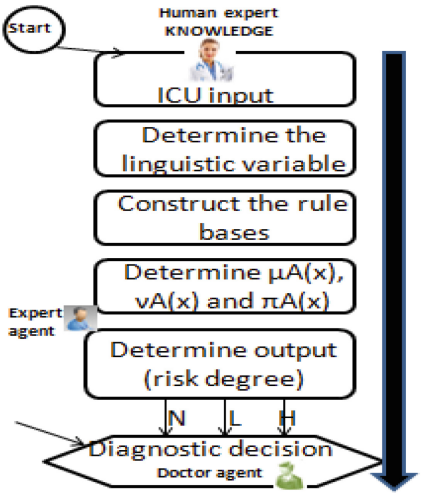


Fig. 2. MIFEDSS process

5 Implementation and Results

In this study, all the IFL parameter were implemented using Matlab 2013b and JFUZZY(a Java-based version of FuzzyLogic) [26].

Every input consists of two or three triangular or trapezoidal membership functions: Temperature (TEMP), O2S, Blood Glucose (BG), Kalmy (KAM): Systolic Blood Pressure (SBP) and Heart Rate (HR). In addition, the rules database that composes the linguistic parameter was formed from the membership functions, which set up a connection between the variables and membership functions. A total quantity of constructed fuzzy rules is 5400 rules calculated using (Eq. 4) that categorizes each input according to the explanation consulted by the specialist.

This number of rules is presented in the Eq. (4).

$$N = p_1 \times p_2 \times \ldots\ldots\ldots \times p_n \tag{4}$$

N: present the full number of possible rules and Pn present the number of linguistic parameter for the input fuzzy sets N.

Figure 3 illustrates membership plot for Blood pressure (SBP) in both Matlab modeling and in our Mobile cloud computing applications [27–29] consisting of five membership function values (Low 3, Low2, Low1, Normal, and High) (Table 4) five membership functions are trapezoidal.

The output parameter "degree of risk" is expressed by three fuzzy linguistic values (Normal, Large and High) as shown in "Fig. 4". It consists of three triangular membership function varies from 0 to 15 (Table 5).

The validation of the IFL is based on the comparison between the risk calculated by MIFEDSS, and the risk presented by the MEWS score. The experimentation presented

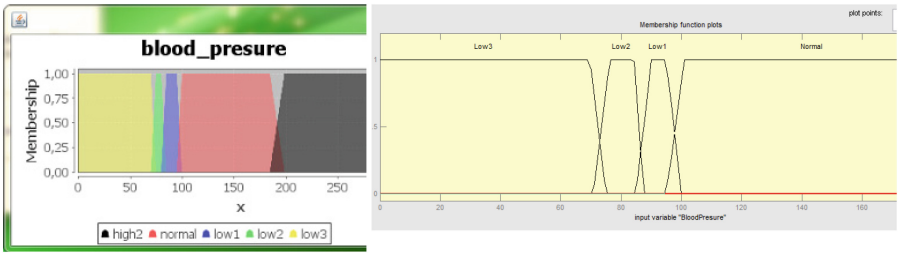


Fig. 3. The SBP membership functions

Table 4. The collections of SBP parameter

Input parameter	Collection	Fuzzy sets
Blood pressure BP	<75	Low 3
	70–85	Low 2
	80–100	Low 1
	95–199	Normal
	>185	High

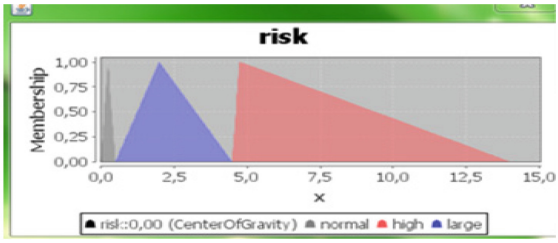


Fig. 4. The triangular membership functions of the output parameter

here (Fig. 5) include 16 patients, are from POLYCLINIC ESSALEMMA-Sfax and 4 anticipated patients (Normal patient).

Table 5. The output "Risk" collections

Output parameter	Arrange	linguistic variables
Risk	$0 < R < 0.5$	Normal
	$0.5 < R < 4.5$	Large
	$4.5 < R < 14$	High

Derived from the experimental results, IFL MEWS score has shown remarkable improvements in decision-making abilities over typical MEWS score. The MIFEDSS prototype presents a commanding capability to cope with the uncertainty issue in the

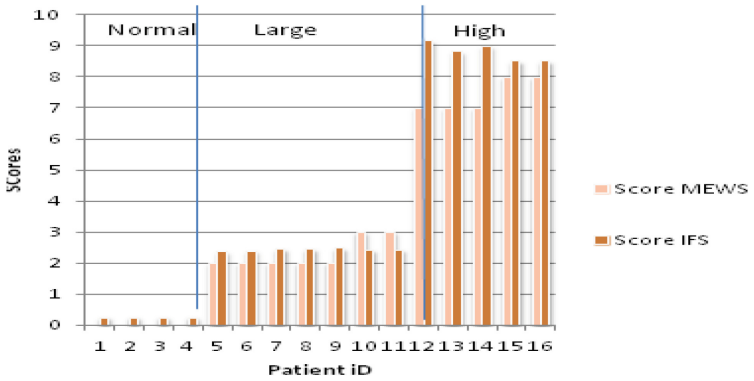


Fig. 5. Column chart for comparison between MEWS score and the MIFEDSS

ICU detection of patient's health status. The MIFEDSS prototype yields remarkable improvements in decision-making abilities over typical MEWS score (100% accuracy) and the results obtained match with the expert's opinion. To date, the prototype results are very encouraging and doctors interest in the latent medical helpfulness and utility of the system. Nerveless in this study we only evaluate the system for 16 patients may bee we can led to substantial improvement of the prototype if the number of patient increased, by adding other parameters in MEWS score like age and annury level.

6 Conclusion

Employing the use of decision support system and artificial intelligence using IFL techniques in medical applications could reduce the time, caregiver's expertise, cost and medical error. In the arena of healthcare diagnosis it acts like a potent tool to help physician to examine medical data.

In this paper, we present the MEWS based IFL, and through the experimentation of the MIFEDSS prototype, we came across different pretty points for IFL (Performance, Scalability and flexibility). In future work, we will reflect to apply the IFL to further similar medical decision system. Also, we plan to make a hybrid evaluation (subjective and objective features) in order to evaluate the tolerability of the applications.

References

1. Mahfouf, M., Abbod, M., Linkens, F.: A survey of fuzzy logic monitoring and control utilization in medicine. J. Artif. Intell. Med. **21**, 27–42 (2001)
2. Semotok, C., Andrysek, J., Basir, O., Otto, E.: An intelligent diabetes software prototype: predicting blood glucose levels and recommending regimen changes. Diab. Technol. Ther. **2** (4), 569–576 (2000)

3. Shereck, D., Jabur, F.: Implementation of a fuzzy logic based expert system to control insulin-pump doses. Mcgill University, ECE Department, Computer Architecture Laboratory (2005)
4. Yue, G., Yi, G.: Application study in decision support with fuzzy cognitive map. Int. J. Comput., 324–331(2007)
5. Bartolomeo, C.: Off-line control of the postprandial glycaemia in type one diabetes patients by a fuzzy logic decision support. Expert Syst. With Appl. **39**(12), 10693–10699 (2012)
6. Adeli, A., Nashat, M.: A fuzzy expert system for heart disease diagnosis. In: International Multi Conference of Engineers and Computer Scientists, vol. 1 (2010)
7. Kar, S., Majumder, D.: An investigative study on early diagnosis of breast cancer using a new approach of mathematical shape theory and neuro-fuzzy classification system. Int. J. Fuzzy Syst. **18**(3), 1–18 (2015)
8. Qu, Y., Shang, C., Shen, Q., Parthaláin, N.M., Wu, W.: Kernel-based fuzzy-rough nearestneighbour classification for mammographic risk analysis. Int. J. Fuzzy Syst. **17**(3), 471–483 (2015)
9. Jayachandran, A., Sundararaj, G.K.: Abnormality segmentation and classification of multi-class brain tumor in MR images using fuzzy logic-based hybrid kernel SVM. Int. J. Fuzzy Syst. **17**(3), 434–443 (2015)
10. Bingzhen, S., Weimin, M., Chen, X.: Fuzzy rough set on probabilistic approximation space over two universes and its application to emergency decision-making. Expert Syst. **32**(4), 507–521 (2015)
11. Atanassov, K.: Intuitionistic fuzzy sets, fuzzy sets and systems. Fuzzy Sets Syst. **20**, 87–96 (1986)
12. Jeong, Y., Kyung, S., Sun, Y., Chong, D.: An application of interval valued intuitionistic fuzzy sets for medical diagnosis of headache. Int. J. Innovative Comput. Inf. Control ICIC **7** (5(B)), 2755–2762 (2011)
13. Eulalia, S., Janusz, K.: Intuitionistic fuzzy sets in some medical applications. In: Fifth International Conference on IFSs. NIFS 7, pp. 58–64 (2001)
14. Pathinathan, T., Jon, A., Ilavarasi, P.: An application of interval valued intuitionistic fuzzy sets in medical diagnosis using logical operators. Int. J. Comput. Algorithm **3**, 495–498 (2014)
15. Mohammed, M.: Medical diagnosis via interval valued intuitionistic fuzzy sets. Ann. Fuzzy Math. Inf. (2012). ISSN 2093–9310
16. Chetia, B., Das, P.K.: An application of interval-valued fuzzy soft sets in medical diagnosis. Int. J. Contemp. Math. Sci. **5**(38), 1887–1894 (2010)
17. Hoda, D., Mohammadreza, A.: A novel application of intuitionistic fuzzy sets theory in medical science: Bacillus colonies recognition. Artif. Intell. Res. **2**(2), 1–17 (2013)
18. Boquan, L., Zhang, H., Li, Y.: The molds of intuitionistic fuzzy value and their applications. Int. J. Fuzzy Syst. **18**(2), 1–15 (2015)
19. Vahid, K.: Intuitionistic fuzzy set vs. fuzzy set application in medical pattern recognition. Artif. Intell. Med. **47**, 43–52 (2009)
20. Jemal, H., Kechaou, Z., Ayed, M.B., Alimi, A.M.: A multi agent system for hospital organization. Int. J. Mach. Learn. Comput. **5**(1), 51–56 (2015)
21. Jemal, H., Kechaou, Z., Ayed, M.B.: Swarm intelligence and multi agent system in healthcare. In: 6th International Conference of Soft Computing and Pattern Recognition, pp. 423–427. IEEE (2014)
22. Zadeh, L.: Fuzzy sets. Inf. Control. Inf. **8**(3), 338–353 (1965)

23. Hamid M., Dan, I., Jérôme, B., Jean, L., Lamine, B., Mohamed B., Bernadette D.: A fuzzy logic approach for remote healthcare monitoring by learning and recognizing human activities of daily living. In: Fuzzy Logic – Emerging Technologies and Applications, pp. 19–40 (2012)
24. Atanassov, K.: Intuitionistic Fuzzy Sets: Theory and Applications. Physica-Verlag, Heidelberg (1999)
25. Shaddel, F., Khosla, V., Banerjee, S.: Effects of introducing MEWS on nursing staff in mental health in patient settings. Prog. Neurol. Psychiatry **18**(2), 24–27 (2014)
26. JFuzzyLogic Plugins. http://jfuzzylogic.sourceforge.net/html/manual.html
27. Jemal, H., Kechaou, Z., Ayed, M.B.: An enhanced healthcare system in mobile cloud computing environment. Vietnam J. Comput. Sci. **3**(4), 267–277 (2016)
28. Jemal, H., Kechaou, Z., Ayed, M.B., Alimi, A.M.: Cloud computing and mobile devices based system for healthcare application. In: IEEE International Symposium on Technology and Society (2015). ISBN: 978-1-4799-8283-7
29. Jemal, H., Kechaou, Z., Ayed, M.B., Alimi, A.M.: Mobile cloud computing in healthcare system. In: Núñez, M., Nguyen, N., Camacho, D., Trawiński, B. (eds.) Computational Collective Intelligence. LNCS, vol. 9330, pp. 408–417. Springer, Heidelberg (2015)

M2Onto: An Approach and a Tool to Learn OWL Ontology from MongoDB Database

Hanen Abbes(✉) and Faiez Gargouri

MIRACL Laboratory, Higher Institute of Computer Science and Multimedia, Sfax University, Sfax, Tunisia
abbes.hanen@gmail.com, faiez.gargouri@isimsf.rnu.tn

Abstract. Ontologies provide shared and reusable pieces of knowledge about a specific domain. Building an ontology by hand is a very hard and prone to errors task. Ontology learning from existing resources provides a good solution to this issue. Databases are widely used to store data. They were often considered as the most reliable sources for knowledge extraction. NOSQL databases are more and more used to store data. MongoDB database is emerging as the fastest growing NOSQL database in the world. It belongs to the document oriented databases variant. This paper proposes an approach to learn OWL ontology from data in MongoDB database and describes a tool implementing transformation rules.

Keywords: NOSQL database · MongoDB · Ontology learning · OWL · Transformation rules

1 Introduction

The semantic Web remains one of the most appealing technologies in the world. Nevertheless, there is still a hard bottleneck in the generation of ontologies as the fundamental cornerstone of the semantic web. Ontologies provide shared and reusable pieces of knowledge about a specific domain [1]. They may be defined as a formalized vocabulary of concepts and their relationships or as an explicit assumption of a subject domain, that represent an agreed "universe of discourse" to link information structures [2].

By defining shared and common domain theories, ontologies help both people and machines to communicate concisely, supporting the exchange of not only syntax, but also semantics. Hence, the cheap and fast construction of domain specific ontologies is crucial for the success and the proliferation of the semantic Web [3].

More and more researchers are attracted by ontology research fields and many ontology editors have been developed to assist domain experts in developing and managing ontologies. Since manual creation of ontologies is a laborious task, there has been a great inclination towards semi-automatic as well as automatic generation of ontologies over the past decade. Ontology learning [4] emerged then as a field of ontology engineering to propose a solution to this issue. It remains an active research area with considerable potential to facilitate the work of ontology engineers [5]. The main goal of ontology learning methods is to derive automatically an ontology from

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_60

existing data [6]. To provide an extended automated support to facilitate the production of high quality ontologies from databases, adequate ontology learning methods should be elaborated [7].

Databases are often considered as the most reliable sources for knowledge extraction. NOSQL databases[1] are emerging nowadays as a new kind of databases developed in response to the requirements of modern applications.

NOSQL databases permit to store large volumes of structured, semi-structured, and unstructured data. Moreover, they provide high speed access to semi-structured and unstructured data and are very flexible. There are four types of NOSQL databases namely, key/value stores, column family stores, document oriented databases and graph databases [7]. We choose to deal with document oriented databases. They correspond to an extension of the well-known key-value concept where the value consists of a structured document. A document contains hierarchically organized data similar to XML and JSON. This allows to represent one-to-one as well as one-to-many relationships in a single document. Therefore a complex document can be retrieved or stored without using joins. Since document oriented databases are aware of stored data, they enable to define document field indexes as well as to propose advanced query features. The most popular document oriented databases are MongoDB[2] and CouchDB[3].

Due to the abundance of NOSQL databases in the storage of data nowadays, recent attempts have been oriented towards automatic extraction of ontologies from these databases. The results obtained by these methods need yet to be improved.

To the best of our knowledge, there is almost no research works that deal with ontology learning from NOSQL databases in the real sense of the word except some research works dealing with NOSQL databases integration. We are particularly interested to MongoDB as a document oriented NOSQL database.

This paper proposes an automatic ontology learning approach, which acquires OWL[4] ontology based on data in MongoDB database. The resulting ontology is to be used as a conceptual view over the data stored in the MongoDB database.

This paper is outlined as follows. Section 2 describes briefly the related work in terms of ontology based NOSQL databases integration. Our approach to learn an ontology from MongoDB database is detailed in Sect. 3. It first gives an overview of the MongoDB database followed by transformation rules and implementation details. Finally Sect. 4 draws conclusions and suggests further research.

2 Related Work

In [8], authors propose a data integration framework where the target schema is represented as an OWL ontology and the sources correspond to NOSQL databases namely MongoDB and Cassandra. The first step of this work consists of generating a local schema for each integrated source using an inductive approach. This approach uses non

[1] http://nosql-database.org/.
[2] http://www.mongodb.org/.
[3] http://couchdb.apache.org/.
[4] http://www.w3.org/TR/owl-features/.

standard description logic reasoning services like Most Specific Concept (MSC) [9] and Least Concept Subsumer (LCS) [10] in order to generate a concept for a group of similar individuals and to define hierarchies for these concepts. The second contribution enables the specification of a global ontology based on the local ontologies generated for each data source. This global ontology results from the correspondences discovered between concept definitions present in each local ontology. For the source ontology generation, authors use containers, i.e. collections in MongoDB and column families in Cassandra to deduce a schema. Authors consider that each container defines a DL concept and that each key label corresponds to a DL property that can be either a dataTypeProperty or an objectProperty and whose domain is the DL concept corresponding to its container. These concepts are hierarchically organized by means of Formal Concept Analysis methods [11]. Finally an incremental schema generation approach is implemented. At the end of this step, a schema for each NOSQL data source is created. To discover alignments between ontologies and to build the global ontology, the first step is to enrich the two ontologies to be aligned using the IDDL reasoner [12], the second step detects the simple correspondences using three classical alignment processes and the last step detects the complex correspondences.

Authors in [13] develop an ontology based semantic integration system for a column-oriented NoSQL data store (HBase). The proposed system is based on four main steps namely schema generation, schema to ontology conversion, ontology alignment-merging, and semantic querying. In the first step, schema has to be generated for the data stores considering the fact that millions of individuals exist in the store and there is no restriction on the number of attributes for each individual. Two methods for schema generation have been proposed in this work: an on-line Schema generation which generates and maintains schema for a data store that is being populated, and an offline schema generation which generates and maintains schema for a data store that is already populated. The proposed system creates a lookup table in HBase for each table whose schema has to be extracted. For this a map-reduce task is launched on the source table. Then, schema extraction is done by finding the fittest individual by means of a genetic algorithm. The schema obtained gives an idea of the column families and columns presenting the greatest common subsumer of the data store which is an individual that can represent every other individual in a given set. This information is converted to suitable OWL primitives and mapped into OWL ontology. The system extracts column details from the persisted schema of the table and applies conversion rules to convert them to ontology entities. In the third step, the local ontologies are aligned to find semantic similarities. COMA++ 3.0 CE tool[5] was used to find lexical similarities between ontology entities. After generating alignments, a global ontology is formed from local ontologies by adding the new relationships found between entities of local ontologies. Finally, The global ontology serves as T-Box for OWL reasoners that query the SPARQL endpoint. The RDF triples obtained as result of SPARQL query are used for inference and more matching triples are generated as output of query. A custom made SPARQL endpoint with reasoning for HBase is the main contribution in this module.

[5] http://dbs.uni-leipzig.de/en/research/projects/schema_and_ontology_matching/coma_3_0/coma_3_0_community_edition.

3 How to Learn OWL Ontology from MongoDB?

3.1 Overview of MongoDB

MongoDB is the fastest growing new database in the world. It provides a rich document oriented structure with dynamic queries and allows to compartmentalize data into collections to divide data logically. Thus, the speed of queries can increase dramatically by querying on a subset of the data instead of all it. Collections are analogous to tables in a relational database. Each collection contains documents that can be nested in complex hierarchies but still easy to query and index. A document is seen as a set of fields, each one being a key-value pair. A key is a string and the associated value may be a basic type, a document, or an array of values. MongoDB is schema-less. The documents stored in the database can have varying sets of fields, with different types for each field. MongoDB offers a tree-like structure in documents by storing references to "parent" nodes in children nodes to model large hierarchical data relationships. MongoDB does not support joins. However, in some cases it is better to store related information in separate documents, typically in different collections or databases. This is insured by means of DBRef references. DBRefs are references from one document to another using the value of the first document's_id field, collection name, and, optionally, its database name. By including these names, DBRefs allow documents located in multiple collections to be more easily linked with documents from a single collection. MongoDB can manage data of any structure without expensive data warehouse loads, no matter how often it changes. So, we can cheap new functionalities without redesigning the database.

3.2 Transformation Rules

Our approach to learn an ontology from MongoDB database is based on transformation rules to map MongoDB constructs into OWL ontology [14]. Our choice is oriented to OWL as the ontology representation language since it is the standard recommended by the World Wide Web Consortium (W3C)[6] to represent ontologies. OWL facilitates greater machine interoperability of web content than that supported by other ontology representation languages such as XML and RDF by providing additional vocabulary along with formal semantics. We are interested particularly to OWL-DL since it supports the maximum expressiveness while retaining computational completeness and decidability. Our approach is based on five main steps.

The first step is the creation of the ontology skeleton. It consists of defining ontology classes and detecting subsumption relationships between them. To do this, every collection in the database must be transformed into an ontology class. Then, A subsumption relationship is extracted from the field "parent" in every document. It must be transformed into a "subClassOf" relationship in the ontology.

[6] https://www.w3.org/.

The second step is to learn concepts properties (dataTypeProperties and objectProperties). Each basic field in a document is transformed into a dataTypeProperty in the class corresponding to the collection containing this document into the ontology (Fig. 1).

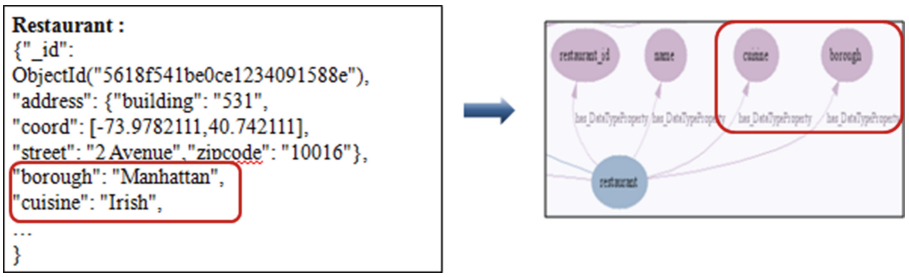


Fig. 1. Learning dataTypeProperties

ObjectProperties are extracted either from embedded documents or from references with DBRef (Fig. 2). An embedded document gives rise to an objectProperty whose "domain" is the class corresponding to the embedding document and the range is the class corresponding to the embedded document. The name of the objectProperty relationship is the concatenation of the word 'has' with the name of the document B (Fig. 2a). A DBRef reference is transformed into objectProperty for which the domain is the class corresponding to the referencing document and the range is the referenced class (corresponding to the referenced document). The name of the objectProperty relationship is the concatenation of the word 'has' with the name of the referenced document (Fig. 2b).

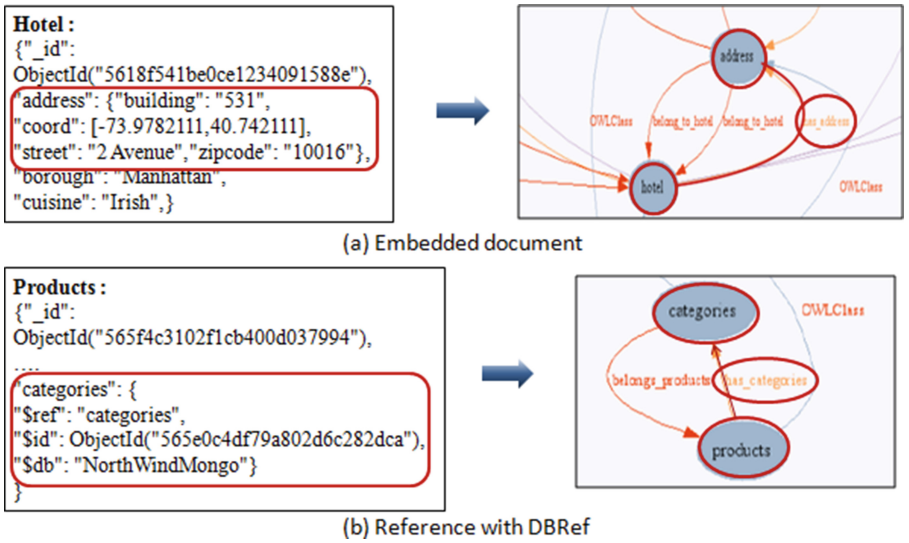


Fig. 2. Learning objectProperties

Individuals are identified in the third step. The values of fields in the database documents are transformed to individuals into the ontology.

In the fourth step, we deduce class axioms (equivalence and disjoining), property axioms (inverseOf) and constraints (cardinality constraints, value constraints). To deal with class axioms, we consider that two classes C1 and C2 are equivalent if both class extensions contain exactly the same set of individuals, else C1 is disjoint with C2. As for the property axiom inverseOf, for each objectProperty corresponds an inverse objectProperty with exchanged domain and range. To learn cardinality constraints, a maxcardinality constraint is calculated from the size of DBList fields. Mincardinality = 0 if the document's fields are empty, else mincardinality = 1. The owl:cardinality constraint is defined if mincardinality = maxcardinality. The value constraint stipulates that all individuals related to an objectProperty must belong to its range.

Finally, we enrich the ontology with classes definition operators (union, intersection, complement).

Classes definition operators (union, intersection, complement) are borrowed from Description Logic. They can be viewed as representing the AND, OR and NOT operators on classes. The intersection class constructor describes a class whose set of individuals contains precisely those ones that are simultaneously individuals of all class descriptions in the list. It is assimilated to logical conjunction. The union class constructor describes an anonymous class whose individuals are those ones that occur in at least one of the class extensions of the class descriptions in the list. It is assimilated to logical disjunction. The complement class constructor describes a class whose individuals are exactly those ones that do not belong to the class extension of the class description that is the object of the statement. It is assimilated to logical negation.

3.3 Implementation

We implemented a tool (M2Onto: MongoDB to ontology) to perform these tasks. This work is directed by means of the JAVA programming language. To access MongoDB database we used the mongo-java-driver-2.13.0 API and to perform the ontology creation we used the OWL-API version 4.0.1. Our tool allows to select an existing database. By clicking on the "Create Ontology" button (Fig. 3), the M2Onto tool executes transformation rules and generates the corresponding OWL ontology. Our tool provides an OWL file as output as well as a graphical representation of the resulting ontology. The following figure shows an OWL file as output to the ontology learning process.

By clicking the "Show Graph" button, a graphical representation of the resulting ontology appears. The graphical representation is available thanks to the GraphViz API. It is an API that allows to generate graphs from a text file describing nodes and arrows of the graph. GraphViz places the nodes in a way that it minimizes arrows intersection. It allows to visualize concepts and relationships of the ontology. We used GraphViz version 2.24.

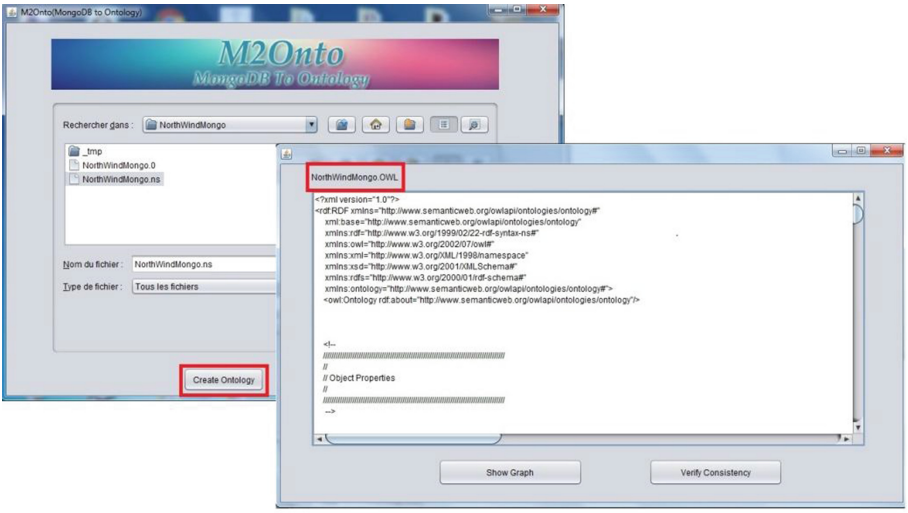


Fig. 3. Visualization of an OWL file with the M2Onto tool

In order to test the validity of our proposal, it was compulsory to use a sample MongoDB database. Unfortunately, we didn't find a sample database that covers all MongoDB constructs. We finally resorted to the "NorthWind" database which is a sample database provided with the Microsoft Access DataBase Management System. It describes all commercial transactions of the "Northwind Traders" company which is specialized in the import-export business of aliments. The relational modeling of the "NorthWind" database is provided in Fig. 4. The database describes all sale transactions that take place between the company and its customers and all purchase transactions between the company and its suppliers. The "Northwind" database is available in the JSON format for MongoDB with the name "NorthWindMongo". It may be downloaded from https://github.com/raynaldmo/northwind-mongodb. It contains 13 collections namely: "categories", "companies", "customers", "employees", "northwind", "order_details", "orders", "products", "regions", "shippers", "suppliers" and "territories".

The "NorthwindMongo" database is a simple data dump from the "Northwind" database to a data set in the JSON format. It consists on a "flat" transformation that ignores both notions of embedded documents and references with DBRef. It simply transforms relational tables to collections and ignores joins and integrity constraints which have to be transformed into references and embedded documents. Moreover, this data dump fails to implement the "parent/child" relationship between collections. Thus, it is not adequate to test the totality of transformation rules that we proposed. To this end, we modified an excerpt of the database that carries on the collections highlighted in Fig. 4. Moreover, to represent the «parent/child» relationship, we created a collection "companies" and added the field "parent" in the collections "customers" and "shippers". As an example, a document containing the field "parent" in the collection "customers" becomes as follows: {"_id": ObjectId("566602bd442e0ce009000029"), "parent":

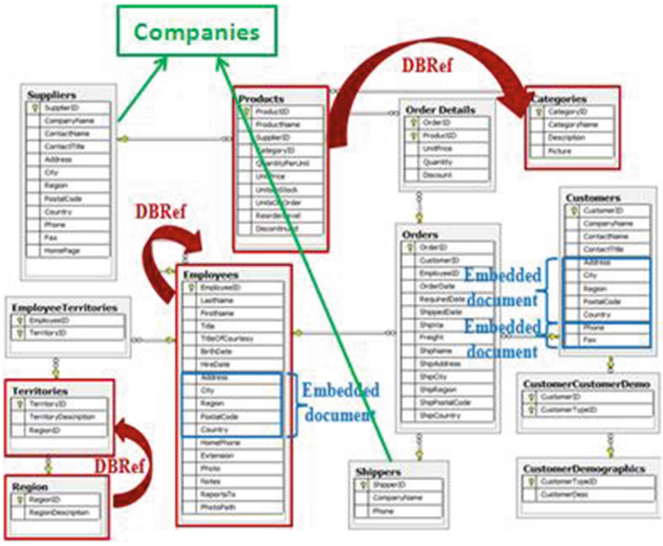


Fig. 4. Summary of the changes performed against the "NorthWindMongo" database schematized over the relational model of the "NorthWind" database

"companies"}. We also eliminated null valued fields in order to preserve the shemalessness aspect of the MongoDB database.

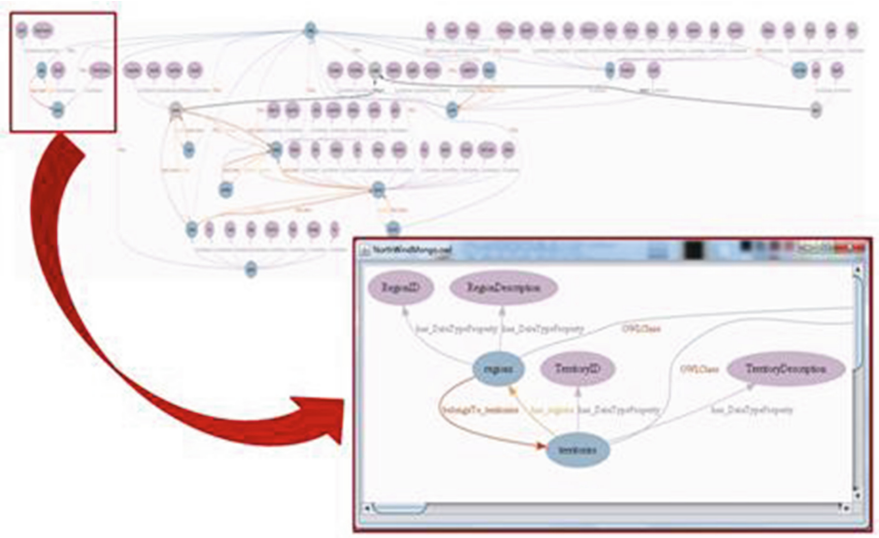


Fig. 5. The "NorthWindMongo" ontology graph visualized by the M2Onto tool

The execution of the M2Onto tool over the "NorthWindMongo" database gives a consistent Owl ontology as output. Figure 3 provides the resulting OWL file and Fig. 5. shows the graphical representation of the "NorthWindMongo" ontology.

Classes are represented by blue nodes. Sub-classes are represented by grey nodes. DataTypeProperties are represented by purple nodes. ObjectProperties are represented by red edges and their inverses by orange edges.

The evaluation of the resulting ontology was performed by checking the consistency of the generated ontology. To do this, we integrated the Pellet[7] reasoner to the M2Onto tool.

4 Conclusions and Future Work

Ontologies play an important role for knowledge intensive applications as a set of formally defined terms to allow concise communication. Research on ontology is becoming increasingly widespread in the computer science community. Ontology learning is a good solution to minimize mass of handwork met by the ontology engineer. This paper dealt with ontology learning from MongoDB database to address the semantic aspect of the stored data. We discussed state of art regarding NOSQL databases integration. An ontology learning approach is proposed to automatically construct OWL ontology based on data stored in MongoDB database. The related learning rules are discussed. It can be seen that the approach is practical and helpful to automate the task of ontology building from MongoDB database. A tool implementing the proposal is also described.

As future work, we envisage to integrate other ontology learning techniques into our approach to obtain better learning result and to propose an approach to update the resulting ontology taking into account the updates of the MongoDB database.

References

1. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquis. **5**(2), 199–220 (1993)
2. George, D: Understanding structural and semantic heterogeneity in the context of database shema integration. In: Proceedings of 6th Conference in the Department of Computing, (J. Dept. Comput. UCLan, (4), pp. 29–44 (2005). ISSN 1476-9069
3. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intell. Syst. **16**(2), 72–79 (2001)
4. Maedche, A., Staab, S.: Ontology learning. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 173–190. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24750-0_9
5. Cimiano, P., Mädche, A., Staab, S., Völker, J.: Ontology learning. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 245–267. Springer, Heidelberg (2009). doi: 10.1007/978-3-540-92673-3_11

[7] https://www.w3.org/2001/sw/wiki/Pellet.

6. Bontcheva, K., Sabou, M.: Learning ontologies from software artifacts: exploring and combining multiple sources. In: Proceedings of 2nd International Workshop on Semantic Web Enabled Software Engineering (SWESE) (2006)
7. Nayak, A., Poriya, A., Poriya, A.: Type of NOSQL databases and its comparison with relational databases. Int. J. Appl. Inf. Syst. (IJAIS) **5**(4), 16–19 (2013). Foundation of Computer Science (FCS), New York, USA, ISSN 2249-0868
8. Curé, O., Lamolle, M., Le Duc, C.: Ontology based data integration over document and column family oriented NOSQL. The Computing Research Repository (Corr) (2013)
9. Baader, F., Calvanese, D., McGuiness, D.L., Nardi, D., Patel-Schneider, P.: The Description Logic Handbook: Theory, Implementation, Applications. Cambridge University Press, Cambridge (2003)
10. Baader, F., Sertkaya, B., Turhan, A.Y.: Computing the least common subsumer w.r.t. a background terminology. J. Appl. Log. **5**, 400–412 (2004). Springer
11. Obitko, M., Snasel, V., Smid, J.: Ontology design with formal concept analysis. In: CLA, vol. 110 (2004)
12. Zimmermann, A., Le Duc, C.: Reasoning with a network of aligned ontologies. In: Proceedings of 2nd International Conference on Web Reasoning and Rule Systems (ICWRRS), pp. 43–57 (2008)
13. Kiran, V.K., Vijayakumar, R.: Ontology based data integration of NoSQL datastores. In: 9th International Conference on Industrial and Information Systems (ICIIS) (2014)
14. Abbes, H., Boukettaya, S., Gargouri, F.: Learning ontology from big data through MongoDB database. In: 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA) (2015)

Comparison of Hard and Probabilistic Evidence in Bayesian Model

Rebai Rim(✉), Maalej Mohamed Amin, Mahfoudhi Adel, and Abid Mohamed

ENIS, Embedded Computer System Lab, University of Sfax, Sfax, Tunisia
rebai_rima@yahoo.fr, m.maalej@gmail.com,
adel.mahfoudhi@fss.rnu.tn, mohamed.abid@enis.rnu.tn

Abstract. Bayesian networks are powerful tools for probabilistic reasoning with uncertain evidences. Evidence originates from information based on the variables of observation. In this paper, we focus on two types of evidences: hard evidence and probabilistic evidence. We were interested in updating an evidence represented by a Bayesian model. This paper presents the application of probabilistic evidence in an adaptive user interface. Then, we compare the Bayesian model using probabilistic evidence with the Bayesian model using hard evidence.

Keywords: Bayesian networks · Probabilistic evidence · Adaptive user interface

1 Introduction

Bayesian networks are probabilistic graphical models specified by two parts representing the state of knowledge about the system [6,10,11,13].

Bayesian inference is used to calculate posterior probabilities, given a set of variables of observations. The inference, in Bayesian networks, is used to update the following variables of model given new information about other variables. Thus, the value of a variable, at a given moment is called evidence. An evidence usually means that this variable is instantiated. Evidences are the starting point of inference methods in a Bayesian network, also called findings. In the literature, we find the term "deterministic or likelihood evidences" and "uncertain or probabilistic evidences [14]".

In this paper, we clarify the different concepts of probabilistic evidence, also called soft evidence, on a variable in a Bayesian network and we present the application of this probabilistic evidence in our Bayesian user modeling. Using Bayesian approach, we interested in a comparison between the adaptive user interface using soft evidence and the adaptive user interface using hard evidence.

The paper is organized as follows: Sect. 2 is devoted to present the basics of Bayesian network and the concepts and definitions related to hard and probabilistic evidence. In Sect. 3, we present an example of application of soft evidence using the associated propagation algorithm and we compare between the Bayesian model using soft evidence and the Bayesian model using hard evidence. Section 4 presents the conclusion.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_61

2 Definition of Types of Evidence: Hard and Soft Evidence

A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph. The Bayesian model is a graph composed of nodes, arcs, and probability distributions [3,14,22].

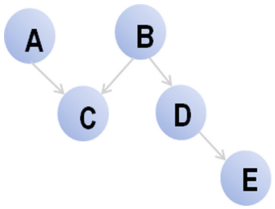


Fig. 1. Example of Bayesian networks structure.

An example of graph is shown in Fig. 1. An arc come from node A and enter node C. Thus, the node A is a parent of the node C. Each node, that has a parent, is presented with a conditional probability distribution $P(C|parents(C))$.

A Bayesian Network is a directed acyclic graph and n random variables $(X_1, X_2, .., X_n)$ such that it is a bijection between the set of random variables and the set of parents:

$$P(X_1, X_2, .., X_n) = \prod_{i=1}^{n} P(X_i|pa(X_i)) \tag{1}$$

where $pa(X_i)$ is the set of parents of X_i.

Historically, the first notion of uncertain evidence in Bayesian networks is due to Pearl [17]. It is specified using a likelihood ratio and propagated using the method of virtual evidence. It is referred to by the term "likelihood finding", "likelihood evidence" or "virtual evidence".

In [4], Chan and Darwish compare in detail the method of Pearl and the rule of Jeffrey [9], which allows to review a joint probability distribution.

More recently, Valtorta introduces the concept of "soft evidence" [22]. We also call "probabilistic evidence" [1]. This type of evidence is propagated by methods which aim to revise the probability distribution of the Bayesian network from probabilistic constraints on a set of variables [2,18,19,22].

Bayesian Networks can be used to manage uncertainty in user modeling [15]. Bayesian user modeling is an effective approach in diagnosing a user's objectives [3,5,7,8,12]. As we will see, soft evidence or "fixed probabilistic evidence [1]" significantly extends the power of Bayesian networks, and makes them much more widely applicable [1,22]. The soft evidence is considered as a direct generalization

of evidence, where the soft evidence is interpreted as information provided from outside the framework of the process modeled by the Bayesian network.

Inference in the Bayesian networks entails the updating of the probability distribution over all findings on variables given the available evidence. Evidence is presented by a collection of findings on variables. A finding can be hard or soft.

A hard finding is a value that a variable X is in. The hard finding is specified by the instantiation of variable X to the value x_i. This hard finding is characterized by $P(X = x_i|f) = 1$ where f is a collection of findings [1]. The soft finding is represented by a distribution R(X) [1,22].

The term soft evidence was proposed by Valtorta et al. [22] and it is presented as a probability distribution R(X). This soft evidence can be seen in some works [1,16,18,19,22]. For example, we cannot observe the precise state of variable but we can know its probability distribution.

3 Comparison Between the Application of Soft and Hard Evidence in Bayesian User Modeling

3.1 The Structure of Bayesian Model

We extracted information from our web interface to built the data set. This web interface allows finding thesis topics, conferences, publishing courses and exchanging documents between teachers and students (Fig. 2).

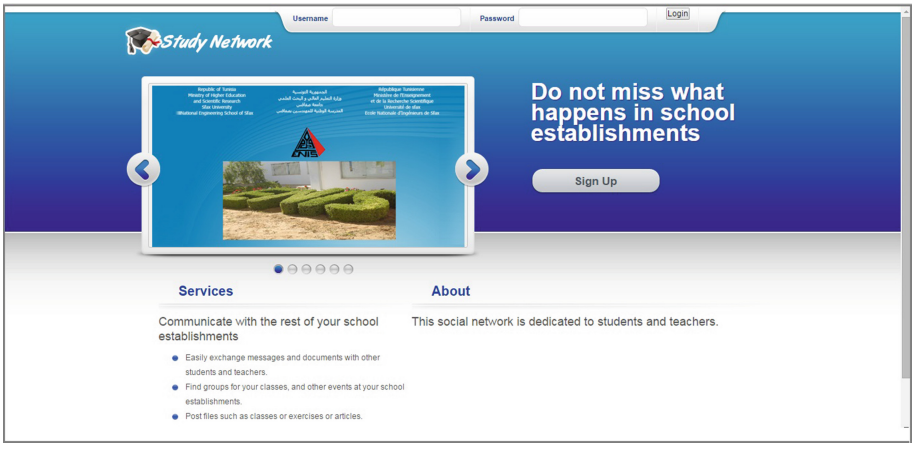


Fig. 2. The web interface.

We defined the structure of user model with simulated annealing algorithm [20]. Given the small number of users in the data, we used the method of cross-validation. Figure 3 shows the structure of the Bayesian model.

In our Bayesian model, the variables age, profession, gender and focus could be represented as evidence to predict the user's preferences.

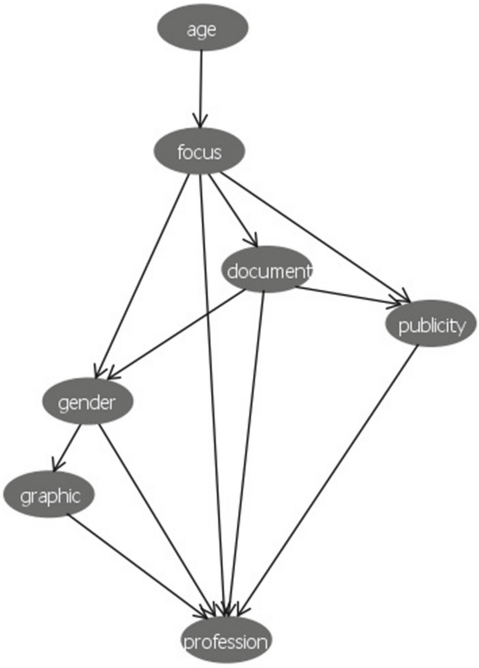


Fig. 3. Structure of Bayesian user model.

3.2 Soft Evidence in Our Bayesian Model

The solution proposed in [21] to propagate evidence on the continuous variables in a Bayesian network is to choose a discretization with a few intervals, and to spread soft evidence on those statements to represent some evidence on continuous variables.

Thus, some evidence on the continuous variables will be replaced by probabilistic evidence on the same variables after discretization. Probabilistic evidence are obtained by fuzzy discretization.

A fuzzy partition on a universe Ω is a set of fuzzy states $f_1, ..., f_p$ such that:

$$\forall x \in \Omega, \sum_{i=1}^{p} X_{f_i}(x) = 1 \tag{2}$$

where X_{f_i} is the function membership of f_i, that is:

$$X_{f_i} : \Omega \mapsto [0, 1] \tag{3}$$

The set of intervals of discretization of the continuous variable is replaced by a set of fuzzy states $f_1, ..., f_p$ as defined above. The initial observation on the continuous variable X is transformed into a distribution probability:

$$R(X_d) = (X_{f_1}(x), ..., X_{f_p}) \tag{4}$$

In our Bayesian model, the variable A represents the age of a person. Its a continuous variable with a value in $D_A = \{junior, adult, oldPerson\}$.

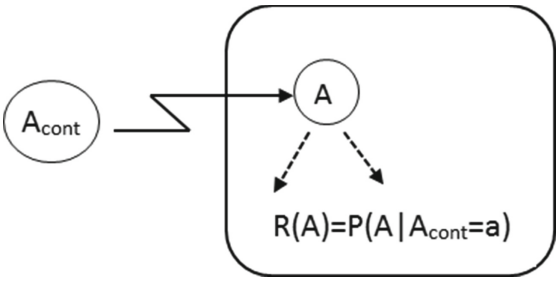


Fig. 4. Soft finding on the variable A.

The soft finding on variable A is represented by a probability distribution $R(A)$ coming from a fuzzy discretization of this variable (Fig. 4).

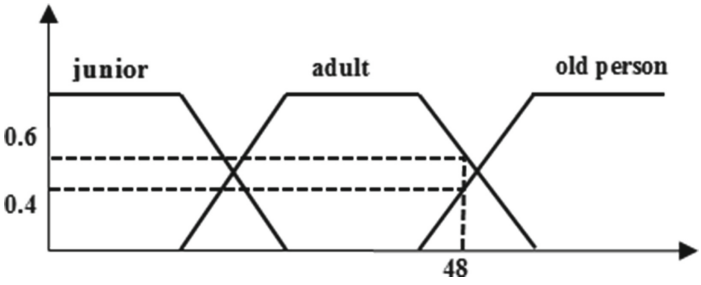


Fig. 5. The membership function of the age A.

For example, the information (the user is 48 years old) is represented by a soft finding on age, by specifying the knowledge of the fuzzy function (Fig. 5). Thus, we obtained: $R(A) = (0, 0.6, 0.4)$.

The variable P represents the profession of a person. The value of P is in $D_P = \{student, teacher\}$. The soft finding P on variable P is represented by a probability distribution R(P).

3.3 Results

We compare the results of classification of the Bayesian network BN1 and BN2 in only propagating hard evidence (in Bayesian network BN1) or soft evidence on the variables Age and Profession and hard evidence on the variables focus and gender (in the Bayesian network BN2).

Table 1 shows the two Bayesian networks used in the tests with the types of evidence used on these variables: hard evidence or soft evidence (fixed probabilistic evidence).

Table 1. The two Bayesian networks used in the tests (number of intervals of discretization and type of evidences).

	Discretization of the variable age	Type of evidence for age and profession variables
BN1	3 intervals	Hard evidence
BN2	3 intervals	Probabilistic evidence

The objective of this comparison is to study the interest of propagating a soft evidence.

After the study of 126 users, the results, presented in Fig. 6, show that 96.43% of the cases were correctly classified. The results provided by applying soft evidence are comparable to the requirements specified by users. The results give an accuracy of 93.61% of the classification of user's preferences where we used only hard evidence in Bayesian model.

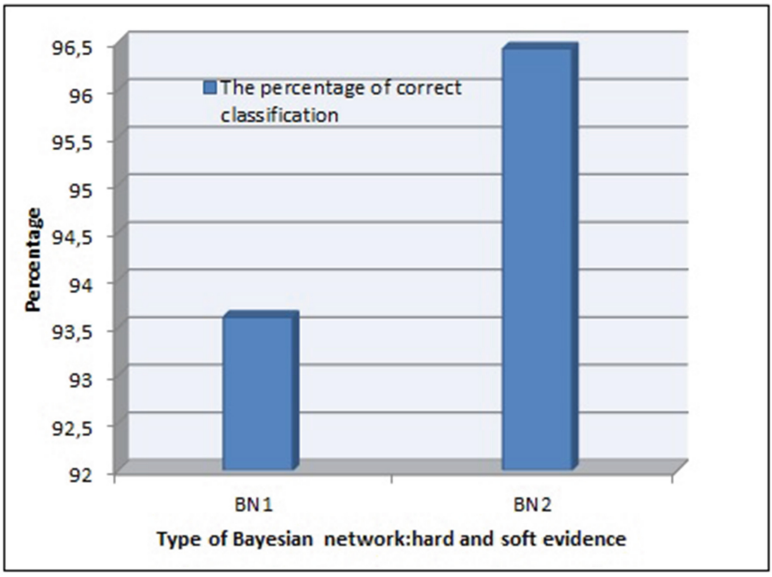


Fig. 6. User classification.

3.4 Propagating Soft Evidence to Predict User's Objectives

Many algorithms with soft evidence have been proposed for Bayesian network reasoning. Most of them can be the combination of the Jeffrey's rule, the Pearl's virtual evidence method and the iterative proportional fitting procedure algorithm (IPFP algorithm) [1].

In our bayesian model, we implemented the BN-IPFP1 algorithm in user interface. This algorithm can implemented with any inference algorithm. The inference algorithm, starting from observed evidence, calculates the probability related to all non observed nodes [1]. In our approach, we reconstruct this probability using the algorithm of Junction Tree.

4 Conclusion

In this paper, we have proposed a using of soft evidence in Bayesian user modeling. Soft evidence extends the power of Bayesian networks. Thus, Bayesian user model is an effective approach in diagnosing a user's preferences. We show the interest of propagating a soft evidence by comparing the two bayesian networks with or without probabilistic evidence.

Later, we can improve the Bayesian user modeling with better training. The user's activities was used to represent the Bayesian user modeling.

References

1. Ben Mrad, A., Delcroix, V., Piechowiak, S., Leicester, P., Abid, M.: An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence. Appl. Intell. 1–23 (2015)
2. Bloemeke, M.: Agent encapsulated Bayesian networks. Ph.D. thesis, Department of Computer Science, University of South Carolina (1998)
3. Brusilovsky, P., Millán, E.: User models for adaptive hypermedia and adaptive educational systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web. LNCS, vol. 4321, pp. 3–53. Springer, Heidelberg (2007). doi:10.1007/978-3-540-72079-9_1
4. Chan, H., Darwiche, A.: On the revision of probabilistic beliefs using uncertain evidence. Artif. Intell. **163**(1), 67–90 (2005)
5. Conati, C., Gertner, A., Vanlehn, K.: Using Bayesian networks to manage uncertainty in student modeling. J. User Model. User-Adap. Inter. **12**(4), 371–417 (2002)
6. Darwiche, A.: Modeling and Reasoning with Bayesian Networks. Cambridge University Press, Cambridge (2009)
7. Fischer, G.: User modeling in human-computer interaction. J. User Model. User-Adap. Inter. **11**, 65–86 (2001)
8. In-Jee, S., Sung-Bae, C.: Bayesian and behavior networks for context-adaptive user interface in a ubiquitous home environment. Int. J. Expert Syst. Appl. **40**, 1827–1838 (2013)
9. Jeffrey, R.C.: The Logic of Decision, 2nd edn., p. 246. University of Chicago Press, Chicago (1990)

10. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge (2009)
11. Korb, K., Nicholson, A.: Bayesian Artificial Intelligence. Chapman and Hall/CRC Computer Science and Data Analysis, Boca Raton (2010)
12. Korpipaa, P., Koskinen, M., Peltola, J., Makela, S.M., Seppanen, T.: Bayesian approach to sensor-based context awareness. Pers. Ubiquit. Comput. **7**(2), 113124 (2003)
13. Koski, T., Noble, J.: Bayesian Networks: An Introduction. Wiley Series in Probability and Statistics. Wiley, Hoboken (2009)
14. Naim, P., Wuillemin, P.H., Leray, P., Pourret, O., Becker, A.: Rseaux Baysiens (2007)
15. Nguyen, L., Do, P.: Combination of Bayesian network and overlay model in user modeling. In: Allen, G., Nabrzyski, J., Seidel, E., Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2009. LNCS, vol. 5545, pp. 5–14. Springer, Heidelberg (2009). doi:10.1007/978-3-642-01973-9_2
16. Pan, R., Peng, Y., Ding, Z.: Belief update in Bayesian networks using uncertain evidence. In: ICTAI, pp. 441–444 (2006)
17. Pearl, J.: Probabilistic Reasoning in Intelligent Systems - Networks of Plausible Inference. Morgan Kaufman, San Mateo (1988)
18. Peng, Y., Zhang, S., Pan, R.: Bayesian network reasoning with uncertain evidences. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **18**(5), 539–564 (2010)
19. Peng, Y., Ding, Z., Zhang, S., Pan, R.: Bayesian network revision with probabilistic constraints. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **20**(3), 317–337 (2012)
20. Rim, R., Mohamed Amin, M., Adel, M.: Bayesian networks for user modeling: predicting the user's preferences. In: International Conference on Hybrid Intelligent Systems, HIS 2013, pp. 144–148 (2013)
21. Di Tomaso, E., Baldwin, J.F.: An approach to hybrid probabilistic models. Int. J. Approx. Reason. **47**(2), 202–218 (2008)
22. Valtorta, M., Kim, Y.G., Vomlel, J.: Soft evidential update for probabilistic multiagent systems. Int. J. Approx. Reason. **29**(1), 71–106 (2002)

Towards Activity Theory - Preliminary Report: Ambient Intelligence Applied to Smart Education

Frederica Gonçalves[1(✉)], Eduardo Fermé[2], Ana Lúcia Martins[3], and João C. Ferreira[4]

[1] Madeira Interactive Technologies Institute/University of Madeira, Funchal, Portugal
frederica.goncalves@m-iti.org
[2] University of Madeira/NOVA LINCS, Funchal, Portugal
ferme@uma.pt
[3] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisbon, Portugal
almartins@iscte.pt
[4] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisbon, Portugal
joao.carlos.ferreira@iscte.pt

Abstract. The limitations and work processes for how people work and interact are suffering changes due to the very fast development of technologies. The use of computers as daily instruments has a reflective effect on people's activity and attitudes towards work. It is a challenge to take advantage of emergent technologies and also to overcome these technologies such as pervasive and smart places. We have changed the way we work shifting often our physical boundaries and how we operate. In this position paper we describe a simple activity analysis of a teacher's class from a high school, based on activity theory, observations and semi-structured interviews. Specifically, we outline our perspective of ambient intelligence in smart education.

Keywords: Activity theory · Ambient intelligence · ICT · Smart education

1 Introduction

Over the last decades the proliferation of human civilization has significantly shifted from desktop computers to new devices such tablet computers, large displays, smart mobiles among other ways which are used to exacerbate work in all its forms. The increasing adoption of computerized tools caused people to express themselves more easily and more often in a diverse ways. Today, people are surrounded by intelligent and intuitive interfaces embedded in all kinds of objects. It is a true challenge to design applications that support users of technology in complex and emergent organizational and work contexts [1]. Information and communication technologies (ICT) and all artefacts would fade into the background while people are immersed in a digital

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_62

environment [2]. In this position paper we outline our perspective as future environments of ambient intelligence in smart education using one of our cases - teachers, which is part of a long term study, of a daily work routine and the use of ICT, based on activity theory to enhance the emerging opportunity to use pervasive technology among other technologies in the teaching-learning process.

2 Background

Activity theory holds that the human mind is the product of our interaction with people and artefacts in the context of everyday activity [3]. This conceptual framework was originally developed by Leontev's [4] and Vygotsky [5] and the foundational concept is "activity", which is understood as purposeful, transformative and developing interaction between actors ("subjects" and the world ("Objects") [6]. A new version (Fig. 1) of activity theory was proposed by Engeström [7] and it is today a global multidisciplinary research approach [7]. The theory provides a number of constructs by focusing on activities as the unit of analysis, activities as goal-directed of "focused" interaction of a subject with an object through the use of tools [8]. The "representation" of these tools are forms of mental processes manifested in constructs whether physical or psychological. This theory identifies both internalization and externalization of cognitive processes involved in use of tools, as well the transformation or development that results from the interaction [8].

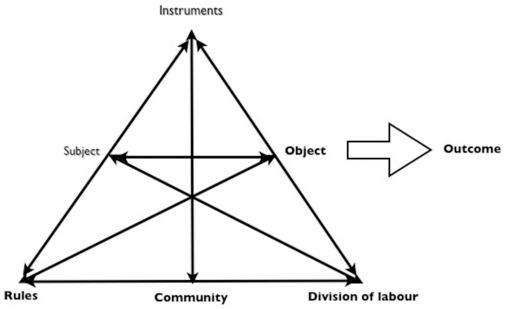


Fig. 1. Model example of human activity by Engeström [7].

As stated by Bazerman [9] as useful and ubiquitous as computers have become in human activity and as innovative as people have been in inventing new uses for computers, it is not easy to think in principled ways about how new computer systems can be made more useful for human proposes. Bardram et al. [1] states that human activity is executed in a hierarchy of three levels: activity-action-operation. They report that an activity is always directed towards a motive, and can be divided into a series of actions, each of which are directed towards more specific goals. These goals are conscious; humans are aware of the goals they want to achieve. Actions, in turn, are decomposed into lower-level units of execution called operations, which are executed

in accordance with, and adjusted to, the specific conditions of the context, in which the operation (and hence the action) takes place. Finally, operations are often automated (internalized) and people are typically not aware of how operations are performed – they become routine [1]. Bannon [10] criticizes the most contextual aspects in traditional HCI models like the human-information processor [11] and the use of the framework of activity theory for new methods, can overcome this limitation [12, 13].

Ambient Intelligence (AmI) is a vision of the future information society stemming from the convergence of ubiquitous computing, ubiquitous communication and intelligent user-friendly interfaces [14]. As stated by Aarts et al. [15] ambient intelligence environments will be characterized by their ubiquity, transparency and intelligence. In classrooms will not only students learn about technologies such as AmI, but ambient intelligence can also help improve the learning experience for these students [2]. The rapid adoption of technologies improve the interactions between students and teachers. Chen et al. [16] have developed a system for test taking using handheld devices. Others such as Abut et al. [17] present a new classroom environment to conduct digital signal processing and communication systems courses. The smart classroom of Shi [18] and the intelligent classroom [19] employs devices such as video and microphones that recognize a set of gestures, motions and speech that can be used to retrieve information or focus attention on appropriate displays and material. Others studies [20] targeted childhood education using a smart table for teachers to observe learning progress of children in the class. Our work revolves around the use of pervasive technology among other ubiquitous solutions in smart education to generate new environments in schools and formulate new ways in the teaching-learning process, as well enhancing exploration and alternatives.

3 Study

The aim of the study, was to examine how teachers go through the process of ideation with ICT and their interaction with their lectures in the teaching-learning process. To contribute towards understanding the "big picture" we applied activity theory as a lens to analyze the information which is being produced. The study applies traditional qualitative sociological methods of study, including participant observation and interviews from 28 teachers.

3.1 Analysis

We interviewed 28 randomly-chosen teachers from different grades and disciplines from a high school[1]. There were 16 females and 12 males aged between 29 and 54 years of age (M = 40.0; SD = 6.38). 14.3% of the teachers that were interviewed had more than twenty years of experience teaching classes, 60.7% had between eleven years and twenty years of experience and 21.4% had between six years to ten years of

[1] Escola da APEL: www.escola-apel.com.

experience. Only 3.6% of teachers had at least three years of experience of giving classes. All teachers use computers, laptops, tablets or smart phones, pen drives or external drives, Internet, personal storage solutions such as could solutions, google docs, google drive, and different software like video editing programs, PowerPoint, learning management system and collaborative work environment (Moodle), among other software in their daily routine at work.

All teachers that were interviewed answered that ICT has changed completely for the better the way they work, expressing reasons such as *"systematization, control, immediate updating of records and observations of students. The fault reviews and all other potential situations inherent in the teaching-learning process, that made us "force" an immediate update of the student."*[T1]; *"It was very rewarding. Allowing us to make the classes more dynamic and interactive. With the technologies I ended up using less photocopies."*[T5]; *"Replacing paper, and on the other hand allowing all information to be stored and if necessary accessed very quickly."*[T6]; *"Technology has greatly facilitated my work, from the calculation of averages of student evaluations to the preparation of more active and dynamic classes, without forgetting the production of new materials in Word and PowerPoint formats in preparation classes. It is also worth highlighting the ease we have in designing audiovisual resources for the classroom, as well as drafting summaries quicker, marking faults and releasing grades through the Platform Place. I refer also to the Moodle platform which is also an advantage, especially in the educational training, I have attended, the level of availability of materials and interaction in the forums."*[T7]*; "In time savings. Collecting data and sharing knowledge."*[T9]; "*Altered since it can be stored in a computer, avoiding the agglomeration of paper. It also becomes useful for designing information in classes for the flat pattern, as we just need to fill in information, there is no need to create root unfolds. In addition, the reduced use of paper, contributes to less pollution."*[T10]; "*Greatly improved. It is more practical, fast. At any time we can project an image regarding the subject!"*[T12].

To present an at-a-glance overview of the answers about how ICT changed their working methods, we have created a word cloud using IBM'S Word Cloud Generator [21] (Fig. 2).



Fig. 2. Word could for the answers of each teacher interviewed.

From semi-structured interviews, we could ascertain that 92.9% of teachers considered that ICT makes work faster, beneficial, efficient and changed the way they worked before. Only 7.1% of teachers considered that ICT did not have an impact on the way they work because it is still a luxury in some education systems, e.g., *"unfortunately not much has changed in the schools that are not adequately equipped to be able to do work that integrates the technologies in an effective way. It is still a luxury."*[T20].

They were all in agreement that it would be a step backwards in the teaching-learning process if they no longer had access to technologies or smart solutions in their work, and they comment, e.g., *"I cannot imagine my classes without technological resources, I no longer know how to work without them (…), for example, how would I talk about a place in historical context without going to the Internet (Google Maps) and show my students, where it is located?"*[T24]; *"I think it would have fewer resources that are important for the proper functioning of the teaching-learning process. The e-docs are essential as well as platforms, since the books are increasingly less used, losing importance to documents produced and stored virtually."*[T1]; *"My work would certainly be much more time consuming and classes would lose a greater ability to motivate students and to manage the time available."*[T7]; *"Difficult, after having computerized processes from a simple summary of class attendance control to the planning of lessons and their respective ratings."*[T15]; *"That scenario is unimaginable, it would be like going back in time, when the school was only centered on the figure of the teacher as transmitting element of knowledge, and the passivity of the student as receiver using only a didactic tool, the book(…)"*[T24]; *"Since the information (technological resources) greatly multiplies the information transmission capacity and knowledge (faster, updated and close to reality), I think that without access, the completion of the work would take more than twice as long. No doubt the banks would change our relationship with the world."*[T26].

From our qualitative data that we obtained in the semi-structured interviews, we can see that interactive technologies are entering all aspects of this teacher's routines and which allows them to have access to all the information they need directly in their classroom. They all agreed that there is need in future education, to exist more interactive and technological resources that promote curiosity, creativity and the development of increased cognitive skills in students as well the teachers.

We randomly took one teacher as the start of the experiment. *She starts her class, by welcoming the students. Later, she starts the computer in the classroom and the software that contains all class data (this software is able to manage courses, classes, students and online summaries) to make the class attendance call. If any student is missing, the teacher marks the absence. After this, the class begins, the teacher projects the slides prepared for this purpose. First she has to turn on the projector with the command and asked for one student to assist her with the slide projection screen on the wall. The slides were in her pen drive. After a while, one student knocks on the door and enter the class, sitting in his place. The teacher writes a quick note in her paper.*

Within some 10 min a string of five rather distinctive actions were observed: (1) initiation of the computer with user name and password and the class data software (2) make the class attendance call and marks the absences (3) turn on the projector and slide the projection screen (4) choose the PowerPoint slides (5) initiate the class.

However, even a slight temporal and spatial extension of this observations reveals that the teacher is not only working on the computer when she initiate the class. She is simultaneously opening the book and reading her notes on paper.

In this observation we have the classical set-up of human-machine interaction: a human operator working on a machine and since the teacher was simultaneously opening her book and reading her notes on paper while she was working with the computer, we have also interaction between human and multiple mediating artefacts in corresponding representational modalities.

To support the reader in understanding the analysis, we illustrate briefly the general aspects of each element in the activity system under study according to Engeström's model [22].

Action 1**:** *Initiate the computer with user name and password and the class data software*. Object: the primary object of this system in action one is the class data software. Subject: the teacher. Instruments: computer system. Outcome: as related to this object the outcome is the image of the task as also the image of the specific data from class data software.

Action 2: *Make the class attendance call and marks the absences.* Object: Class data software and students. Subject: the teacher. Instruments: computer system, questions. Outcome: preliminary verification of students in class and marking absences.

Action 3: *Turn on the projector and slide the projection screen.* Object: projector and projection screen. Subject: teacher. Community: teacher and students. Division of labor: student assists with the teacher. Instruments: computer system, projector and projector command, projection screen. Outcome: projection on the wall.

Action 4: *Choose the PowerPoint slides.* Object: PowerPoint slides. Subject: teacher. Instruments: computer system, projector, projection screen, pen drive. Outcome: image of the slides projections on the wall.

Action 5: *Initiate the class.* Object: reusable knowledge for teaching process. Subject: teacher. Instruments: teacher knowledge, questions, computer, projector. Community: teacher and students. Division of labor: teacher and students interactions. Rules: pedagogy of teaching-learning process. Outcome: lectures knowledge.

Based on the model of human activity Fig. 1 [7] the teacher's first action and the second action are represented with the importance of the computer. It is not the book or her notes she is focused on, it is the class data software on the screen. The book or notes will be her focus of her attention if she need to support their claims or monitoring her class after initiate it. In the third action, the object of attention shifts from the data class software and presences of students on the computer to the projector and the projection screen. The outcome is the projection on the wall. A mediating factor in this action is the division of labor between the teacher and the student. In the fourth action the focus shifts to computer and the pen drive. The teacher has to choose which PowerPoints she wanted to use in her pen drive to project on the projection screen. Finally, in the fifth action the teacher began her class with the assistance of the computer, the book and her notes.

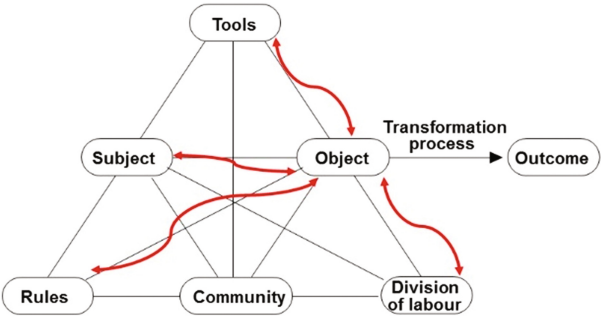


Fig. 3. Conflicts in the activity system - teacher class.

In these observation a large number of repeated disturbances (Fig. 3) were encountered such as student entered late in the class, which resulted the teacher lost class time in the presentation of informative content and the late student did not learn the previous information. As stated by Csikszentmihalyi [23], concentration is an important ingredient to engage participants in an activity with high involvement, enjoyment and intrinsic motivation. Taking into account the occurrences, also the concentration of the students that were already in the class was also disturbed. During the class, both students and teacher were in interaction as the exposure of content and doubts that were emerging, leading to questions and subsequent answers. One tension in rules on pedagogy teaching-learning process, and it was raised, was when two different students provide different exposures of the content, leading to several arguments for the same purpose. In this specific case, the teacher capture the ideas and exposes them more concisely and foment ideas for discussion and increased creativity. Students were later asked to refine their ideas in solving the questions that were raised and from that point forward the teaching-learning process became more human-computer interaction.

4 Discussion and Conclusions

We must to look at how people work in the teaching-learning process to design new environments with pervasive and ubiquitous computing in classrooms. As Gross et al. [24] states, making computationally embedded things demands cross-disciplinary creativity. Ambient Intelligence is designed for real-world, physical environments, effective use of sensors are vital [2]. Taking into account the cycle of expanding learning by Engeström [22], in the case studied, the use of different technologies to minimize the interruptions such as to open a door during the teaching-learning process are a way to make surroundings flexible and adaptive. Based on students' mobile devices it is possible to find student presence in a classroom and student path in a school campus. Example of this is our work [25], applied to check the numbers of persons in a public transportation. This mobility patterns are collected passively without user intervention based on the probe request of mobile devices. These probes

request are sent periodically broadcasting packets, which contain the unique MAC address of the client and (sometimes) the name of a network to which it had previously connected. This can be used to determine the owner of mobile device in a range of an Access Point (AP). We propose a passive Wi-Fi tracking system based on AP in each class zone and a central tracking cloud server. Alternatives to this approach are Bluetooth based on mobile device signal received from beacon, working in a low energy mode (BLE), in this case on beacon per classroom. Basically this BLE have the advantage of no installation need, working in a simple configuration, but Wi-Fi have the advantage of providing more user interaction. This approach Wi-Fi we can track student movements at school campus, identify times spent at lunch, library in classes. With this the process of teaching-learning process it was more involved in the activity of the community and its consequent practice with interrelations that could enable them to learn from each other. Since the classroom had desks with computers incorporated, the teachers could improve their classes using cloud computing solutions to delivering more economical, securable and reliable education services [26]. Another perspective is the use of crowdsourcing resources to share information and to facilitate learning among colleges students with rewards mechanics associated to promote users participation, with create success in social network programs. Example of this is our work applied to the exchange of student among universities [27]. The use of heterogeneous devices such as smart watches, mobile phones, augmented reality among other technologies, could be an emerging engine to improve the teaching-learning process. There is an interesting question that we can raise for future work, facing the high speed evolution in the advances of technologies, that is: how can we developed solutions that can be part of future environments in smart education that are not so expensive as to be considerate a luxury? As stated by Kinshunk et al. [28], advances in technologies can bring enormous potential for improving education through more efficient, different and innovative methods.

As stated previous, one possible scenario of usage could be described as, e.g.: the teacher starts her class. Based on mobile device with sensor information (accelerometer) it is possible to check student class attention with mobile device key pressing detection. Also as stated before Wi-Fi probe request or Bluetooth can be used to identify student presence in the classroom. If they are late, when they sit, students will be able to consult what has been covered in the matter, through a system (e.g. Google Classroom, Microsoft Classroom or other system available) on devices such as tablet or surfaces, in order to follow what was proposed for the day. During the class, to inspire students about the subject being addressed and to facilitate the collaborative learning, the teacher suggests to students the use of their available system, that is connected to a crowdsourcing engine. By doing that she shows different perspectives of that specific subject using images, videos, or even if with an immersive moment with augmented reality. As part of the progression of teaching-learning process students are encouraged to contribute using the app for that proposed which is connected to the crowdsourcing engine with any information they considered relevant to the subject they were learning. They can do this using their mobile phone, smart watch's or any other device.

From a simple activity as with the presented case, a significant problem to design systems that have focus on supporting human activity and involve a multimodal interactive incorporated systems is the *(a)* to understand the physical boundaries of

human works; *(b)* the interactions, relations and context of people's activity; *(c)* when and how to apply pervasive, smart technologies to enhance their work activity.

As future work, we will explore how we can improve future environments in schools using different interactive or multisensory systems to enhance the creativity and productivity levels of students during the teaching-learning process and trying to keep students in a creative experience peak for a longer period of time.

This position paper was written with the intention to spark an initial discussion around future school environments and solutions for ambient intelligence, with the simple focus of analyzing the current challenges in the pedagogy of ideation in teaching-learning process. One solution could be integrating in classrooms more ICT such as interactive surfaces, cloud computing solutions, ubiquitous solutions in such a way that they could improve the quality of instruction for a smarter education. Therefore we propose a pervasive and hopefully more inspiring workplace for all.

Acknowledgements. We would like to thank the teachers who participated in this research. The institutional support from Escola da APEL was generously provided by the director Pe. Fernando Gonçalves, by the pedagogical director Gonçalo Faria and by the administrative secretary Rubina Mendonça. We would also like to thank the previous pedagogical director Marco Gomes. Thanks to ARDITI - Regional Agency for the Development of Research Technology and Innovation through the support provided by the M14-20 Project - 09-5369-FSE-000001- PhD Scholarship. E.F. was supported by FCT MCTES and NOVA LINCS UID/CEC/04516/2013.

References

1. Bardram, J.E., Doryab, A.: Activity analysis: applying activity theory to analyze complex work in hospitals. In: CSCW, Hangzhou, China (2011)
2. Cook, D., Augusto, J., Jakkula, V.: Ambient intelligence: technologies, applications, and opportunities. Pervasive Mob. Comput. **5**, 277–298 (2009). Elsevier
3. Kaptelinin, V., Nardi, B.A.: Acting with Technology: Activity Theory and Interaction Design. The MIT Press, Cambridge (2009)
4. Leontev, A.: Activity, Consciousness, and Personality. Prentice-Hall, Upper Saddle River (1978)
5. Vygotsky, L.: Mind in Society: The Development of Higher Psychological Processes. Harvard Universiy Press, Massachusetts (1978)
6. Kaptelinin, V.: The Encyclopedia of Human Computer Interaction, 2nd ed., Interaction Design Foundation (2015)
7. Engeström, Y.: Activity theory as a framework for analyzing and redesigning work. Ergonomics **43**, 960–974 (2000)
8. Appiah, E., Cronjé, J.: Exploring information and communication theory in graphic design education with activity theory. Int. J. Comput. Appl. **84**, 15–22 (2013)
9. Bazerman, C.: Cyborgs in action. In: Context and Consciousness: Activity Theory and Human-Computer Interaction. Regents of the University of California on behalf of the Laboratory of Comparative Human Cognition, pp. 73–75 (1998)
10. Bannon, L., Bodker, S.: Beyond the interface: encountering artifacts in use. In: Designing Interaction: Psychology at the Human-Computer Interface. Cambridge University Press, pp. 227–253 (1991)

11. Card, S., Moran, T., Newell, A.: The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, Mahwah (1983)
12. Kaptelinn, V., Nardi, B.: Acting with Technology - Activity Theory and Interaction Design. MIT Press, Cambridge (2006)
13. Kuutti, K.: Acitivity theory as a potential framework for human-computer interaction research. In: Context and Consciousness, pp. 17–44 (1996)
14. Friedewald, M., Costa, O., Punie, Y., Alahuhta, P., Hienonen, S.: Perpectives of ambient intelligence in the home environment. Telematics Inform. **22**, 221–238 (2005). Elsevier
15. Aarts, E., Marzano, S.: The New Everyday: Views on Ambient Intelligence. 010 Publishers, Rotterdam (2003)
16. Chen, F., Myers, B., Yaron, D.: Using handheld devices for tests in classes. Technical report CMU-HCII-00-101, CMU-CS-00-152 and Human Computer Interaction Institute (2000)
17. Abut, H., Öztürk, Y.: Interactive Classroom for DSP/Communications Courses. In: Proceedings of ICASSP, pp. 15–18 (1997)
18. Shi, Y., Xie, W., Xu, G., Shi, R., iChen, E., Mao, Y., Liu, F.: The smart classroom: merging technologies for seamless tele-education. In: IEEE Pervasive Computing, pp. 47–55 (2003)
19. Franklin, D.: Cooperating with people: the intelligent classroom. In: Proceedings of the National Conference on Artificial Intelligence, pp. 555–560 (1998)
20. Steurer, P., Srivastava, M.: System design of smart table. In: Proceedings of the IEEE International Conference on Pervasive Computing Communications (2003)
21. IBM: Word Cloud Generator (2016). http://www.ibm.com/us-en/
22. Engeström, Y.: Learning by expanding: an activity theoretical approach to development research. Orienta-Konsultit, Helsinki (1987)
23. Csikszentmihalyi, M.: Flow: The Psychologic of Optimal Experience. Harper Perennial, New York (1991)
24. Gross, M., Yi-Luen Do, E.: Educating the new makers: cross-disciplinary creativity. Leonardo – J. Int. Soc. Arts Sci. Technol. **42**, 210–215 (2009)
25. Baeta, N., Fernandes, A., Ferreira, J.: Tracking users mobility at public transportation. In: Bajo, J., Escalona, M.J., Giroux, S., Hoffa-Dąbrowska, P., Julián, V., Novais, P., Sánchez-Pi, N., Unland, R., Azambuja-Silveira, R. (eds.) PAAMS 2016. CCIS, vol. 616, pp. 224–235. Springer, Heidelberg (2016). doi:10.1007/978-3-319-39387-2_19
26. Jeong, J., Kim, M., Yoo, K.: A content oriented smart education system based on cloud computing. Int. J. Multimed. Ubiquit. Eng. **8**, 313–328 (2013)
27. Ferreira, J.C., Porfírio, F.: Towards an academic social network for Bologna process. In: Proceedings of the International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE 09), December 2009
28. Kinshunk, S., Sutinen, E., Goh, T.: Mobile technologies in support of distance learning. Asian J. Distance Educ. **1**, 60–68 (2003)

Emerging Opportunities for Ambient Intelligence in Creativity Support Tools

Frederica Gonçalves[1(✉)], Eduardo Fermé[2], and João C. Ferreira[3]

[1] Madeira Interactive Technologies Institute/University of Madeira, Funchal, Portugal
frederica.goncalves@m-iti.org
[2] University of Madeira/NOVA LINCS, Funchal, Portugal
ferme@uma.pt
[3] ISCTE - Instituto Universitário de Lisboa, Lisbon, Portugal
joao.carlos.ferreira@iscte.pt

Abstract. The fundamental challenge in developing and evaluating creativity support tools is that we are not able to detect when a person is being creative. In this position paper we described our perspective of ambient intelligence in creativity support tools specially in the use of creative writing environments. Starting with the activity theory, we describe a simple analysis of writing sessions involving 100 students from a higher school, and recorded using a keystroke logging program called Inputlog and the program iTALC to support it. Specifically, we outline the writing activity goals and the interaction design goals.

Keywords: Future environments · Ambient intelligence · Creativity support tools

1 Introduction

Interactive technologies and new environments are used by a crowd of users with diverse characteristics, needs and requirements [1]. Humans are able to distinguish different scenes, objects or places with all their senses and also have the ability to recall those to spark creativity, e.g. while writing a text. Today, people can take small yet powerful computers anywhere at any time. The rapid adaptation of smart and ubiquitous or pervasive computing enable people to have access to information much faster that previous generations.

With the new trends in natural interaction between computer and human there is a challenge to empower users to harness and embrace their creativity through the use of creativity support tools [2]. Robison [3] defines creativity as the process of having original ideas that have value. Creativity it is also a complex and multifaceted phenomenon [4]. As stated by Shneiderman [5] creativity includes discovery or invention of a significant idea, pattern, method, or device that gains recognition from accepted leaders in a field. Creative workers use artificial stimuli as inspirational guides, e.g., listening to music while writing or looking at images while drawing. Indeed, well-designed creativity support tools can help users in generating multiple levels of creativity for example

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_63

during the process of writing, particularly those tools that can also generate different stimuli. Ambient intelligent environments can be characterized by their ubiquity, transparency and intelligence [6], and also eventually intelligent agents or algorithms creativity support tools, would help decisions that automatically serve a person or notify to make a decision or to carry out an action [7].

In this position paper we outline our perspective of ambient intelligence in creativity support tools, especially in the user interfaces to support creative writing. Based on activity theory we enhance the emerging opportunity to use ubiquitous or pervasive technology in a creative process. We present a study using different modalities - such as sound and smell - on a creative writing activity of 100 students in a high school[1]. In a between-subjects design we tried to discover patterns in the influence of smell and sound on the participants writing activity while using a word processor in different environments. These modalities were chosen because they do not distract the users' attention from the main creative task.

2 Background

Ambient Intelligence (AmI) is an emerging discipline that brings intelligence to our everyday environments and makes those environments sensitive to us [8]. It is a multidisciplinary paradigm that draws a new kind of relationship between humans, their environment and the technology [9]. Weiser [10] states that the computer of the 21th century should be invisible to its users, also it would be embedded in the environment. Such invisibility is related to the capacity of the technology to help users to reach their goals in a less obtrusive way. As stated by Hopper [11] applications can be made more responsive and useful by observing and reacting to the physical world and it is particularly attractive in a world of mobile devices and ubiquitous computers. The development of technologies including facial expression [12], emotion [13], speech [14], or gesture recognition, motion tracking [15], facilitate normal interactions with intelligent environments. Essentially, AmI systems should know when it is convenient to interrupt a user when to make a suggestion but also when it is more convenient to refrain from making a suggestion [8]. Our work revolves around creativity support tools and using different modalities, such as sound and smell in a creative writing process, to improve and formulate new guidelines for design for this type of user interfaces, as well as enhancing exploration of future environments.

Interaction with technology gives fresh possibilities to use it in a creative way while also leading to the evolution and sometimes to a transformation of that specific technology. Emerging computer-based tools can develop better and more creative solutions to the problems that we face in our day to day [16]. The success during software development, despite of being a conceptually complex, knowledge intensive and cognitive activity, depends on the creativity of software engineers [17]. There's an effort for developing creativity support tools, which enable us to explore, discover, imagine, innovate, compose and collaborate [18]. And also it is a challenge to construct

[1] Escola da APEL www.escola-apel.com.

information technologies that support creativity and the goal of developing new creativity supported tools can be obtained by building upon an adequate understanding of the creative process [5]. A creativity support tools is any tool that can be used by people in the open-ended creation of new artifacts [19]. Creativity and motivation enhancement can easily be aligned with the design of high-quality human-computer interaction and also creativity might be viewed as any process which results in a novel and useful product [20]. Wang et al. [21] and Andolina et al. [22], targeted stimuli such as the visual stimuli, images and text, to support creativity, increasing both originality and diversity of ideas during brainstorming. Other researches [23] studied UI Zen-based themes, composed of sound and images to foster inspiration, focus and immersion on creative writing tasks.

The model of activity theory (Fig. 1) was proposed by Engeström [24] that provides a number of constructs by focusing on activities as the unit of analysis, activities as goal-directed of intensive interaction of a subject with an object through the use of tools [25]. The illustration of these tools are forms of mental processes manifested in constructs whether physical or psychological. As stated by Appiah et al. [25] this theory identifies both internalization and externalization of cognitive processes involved in use of tools, as well the transformation or development that results from the interaction.

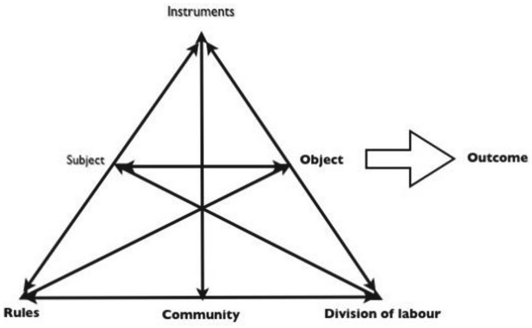


Fig. 1. Model example of human activity by Engeström (Engeström [24]).

Human activity is executed in a hierarchy of three levels: activity-action-operation [26]. An activity is always directed towards a motive, and can be divided into a series of actions, each of which are directed towards more specific goals [26]. The major key terms in activity theory include internalization, mediation, subject, object, tool, process (transformation), rules community, division of labor and outcomes [25]. As stated by Kuuti [27] activity theory provides a philosophical and cross disciplinary framework for studying different forms of human practices as development processes.

3 Study

The aim of the study, was to examine how students go through a creative process with computers in a special environment - using modalities such as sound and smell, and their interaction in a creative task. This study applies a between-subject design experience,

and we focused on the creative task activity of 100 students. We used a creativity activity such a writing task, since they were students and during their school curriculum they have to write compositions in various disciplines as further evaluation. We recorded the writing sessions using a keystroke logging program called Inputlog [28] and the program iTALC [29] to support the observations. Also to contribute towards understanding we applied activity theory [24] and semi-structured interviews.

3.1 Participants, Conditions and Task

The experiment was conducted in a classroom in the secondary high school, during two weeks, and we recruited a group of students. Our experimental design was based on a between-subjects design, in which seven groups of students were randomly assembled according to different conditions such as gender, age, computing experience, and Internet experience, to ensure that the groups were similar as possible.

This experience involved 100 students from the institution's population, aged between 15 and 19 years old (M = 15.9; SD = 0.94). There were 45 females and 55 males. All subjects were naïve to the experimental conditions. We lead one session per group of students, making up a total of seven sessions for the experience. The classroom was equipped with desks with integrated computers (Fig. 1).

The creative activity in this experiment included writing in four creative writing conditions. Microsoft Word was the word processor in this experiment. We chose this creative writing user interface because the participants were more familiar with it, has a large formatting options, the ability to change font-weight and many other things that allows users an essay and near-instant input.

Condition 1 (N = 14). Our baseline (*session 1*) without any cues.

Condition 2 (N = 22). We used two types of olfactory cues: an alerting smell (*session 2*) was achieved by the actual fragrance of hot coffee that was spread around the room in small cups, and relaxing smell (*session 3*) – for which we used a laurel fragrance.

Condition 3 (N = 31). In this condition we used also two auditory cues: an alerting sound (*session 4*) - users might feel that they would be sitting in a cafe with the constant bustle of movement of people, machines and dishwashers, and a relaxing sound (*session 5*), achieved using a natural soundtrack featuring water, birds and foliage.

Condition 4 (N = 33). We used the combination of sound and smell in each condition – alert and relaxed. One group of students could have the smell of the real coffee and the sound of a café (*session 6*), and the other group of students could had the relaxing fragrance and sound of nature (*session 7*).

Task. We presented a map to students and they had to choose two points. Students had to write a short story using the starting point as a map. Through their own imagination and creativity, they wrote a story of the path from point X to Y. They initiated their writing from their own ideas. The time limit was 15 min to complete the writing task.

3.2 Results

To study the activity system according to Engeström model [24], we randomly took one writing session as starting point. *Students initiate their writing as if they had their ideas already predefined. After a while, they were looking at the map while writing or staring vaguely at the horizon, pausing as if to think and then they started writing*. The action in this activity was a writing task using as instruments the software (Microsoft Word) in the computer for that purpose. One rule existed, students could not use the internet while writing. With a disturbance such as the absence of ideas and resulted in a pause while writing, made the object (in this case the student) activate a contradiction in the initially exposed rule. The outcome in this case, is the final draft of a short story.



Fig. 2. One group of students performing the study in the moment of the experience.

It was also monitored this activity through the iTALC program, and one example of disturbance in this activity, was the use of image search when students stopped writing (Fig. 2). When we interviewed them, enquiring why they did this search, students

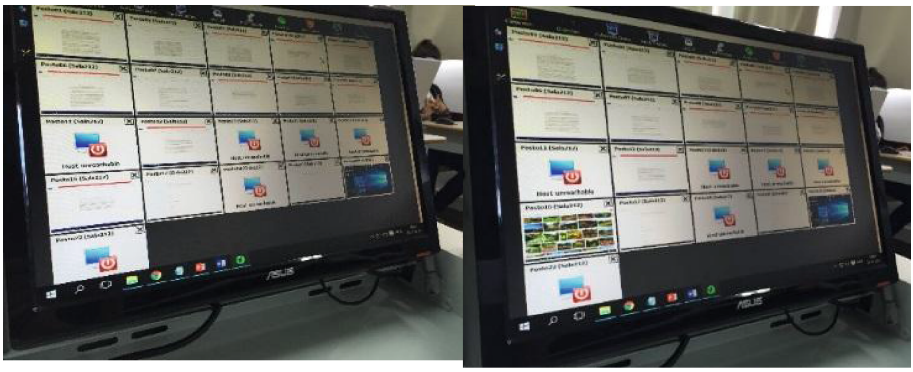


Fig. 3. Fragment of monitoring during the writing activity: example of the use of image search during the writing task.

assumed they were out of ideas to write and wanted to find inspiration, or others said that they wanted to enrich their text with a more elaborate definitions (Fig. 3).

Using the keystroke logging program called Inputlog, we had an overview of information such as the average time in each session on the writing activity using the keyboard (Table 1), the average time of pauses in each session while writing and the average of switches between mouse and keyboard in each session (Table 2).

Table 1. Average time in each session on the writing activity using the keyboard

	Condition 1	Condition 2		Condition 3		Condition 4	
	Session 1	*Session 2*	*Session 3*	*Session 4*	*Session 5*	*Session 6*	*Session 7*
Mean (seconds)	00:12:03 (72,180)	00:12:02 (72,120)	**00:10:53** (63,180)	00:11:44 (68,640)	00:12:28 (73,680)	00:11:27 (67,620)	00:12:19 (73,140)
Std. deviation	0,00233	0,00298	0,00305	0,00260	0,00235	0,00266	0,00230

Table 2. Average time pauses in each session while writing and average switches between mouse and keyboard.

	Condition 1	Condition 2		Condition 3		Condition 4	
	Session 1	*Session 2*	*Session 3*	*Session 4*	*Session 5*	*Session 6*	*Session 7*
Mean (pauses)	00:06:16 (36,960)	00:04:33 (25,980)	00:04:48 (26,880)	00:04:19 (25,140)	00:04:10 (24,600)	00:04:45 (26,700)	00:03:55 (21,300)
Std. deviation (pauses)	0,00287	0,00100	0,00083	0,00112	0,00107	0,00104	0,00125
Mean (switches)	33,714	10,333	23,500	26,118	18,429	28,476	19,154

In Table 2 results shows that the average time on the writing activity using the keyboard in session 3 (Condition 2) had lower values than the other sessions, but results were not of statistical significance.

Regarding the average switches between mouse and keyboard, results in Table 3 shows that the participants in Session 1 (baseline) had changed more often than the other participants in the other Conditions. From the results, we also noted that participants in Session 2 had changed fewer times between the keyboard and the mouse, as well participants in Session 5 and Session 7. Considering the average time pauses while writing,

Table 3. Statistical analysis for stories written in each condition.

	Condition 1	Condition 2		Condition 3		Condition 4	
	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6	Session 7
Mean	230,10	412,60	210,10	258,40	273,70	279,20	277,10
Std. error	39,383	30,768	24,897	32,245	41,226	30,234	17,669
Median	217,50	429,50	192,00	232,50	294,50	275,00	264,00
Std. deviation	124,541	97,298	78,732	101,969	130,369	95,609	55,876
Minimum	87	211	124	136	54	117	201
Maximum	493	533	364	399	463	433	365

participants in Session 1 stopped more often than the others in each session, but results were not of statistical significance.

Regarding the stories written and the data dispersion, results (Table 3) shows that the participants in the Condition 2 (Session 2 and 3) wrote an average more words, which contrasts with our baseline - Condition 1 (Session 1), and with Condition 3 (Sessions 4 and 5) and Condition 4 (Session 6 and 7).

Using Friedman's ANOVA approach to testing differences between each condition results were statically significant $Fr(6) = 15.77$, $p < .05$. Therefore the non-parametric Wilcoxon test was applied to follow up this finding. For students in Condition 2 the results were significantly higher than students in Condition 1 (our baseline), ($T = 1$, $z = -2.59$, $p < .05$, $r = -.43$). When comparing Condition 1 with Condition 3 and Condition 4, results were not statistically significant. In general, we observed that the participants were concentrated on the writing moment. Triangulating the results with semi-structured interviews it was clear that all participants felt somewhat creative during the experiment. In the semi-structured interview, 50% of students consider that they did feel more creative in Condition 1, 54.8% in the Condition 3 and 63,6% in the Condition 4. 72.7% of the students in the Condition 2 considered themselves more creative in this condition. One of the factors that are believed to play an important role in human creativity is our sensory richness. To improve and foster creative tasks, our study suggests that the combination of different modalities such as sound and smell, should increase research on the development of creativity support tools and environments.

4 Discussion

Cook et al. [8] suggest that ambient intelligence is maturing and the resulting technologies promise to revolutionize daily human life by making people's surroundings flexible and adaptive. In the field of ambient intelligence applied to education there are few studies referring to smart classrooms that use an interactive whiteboard that stores content in a database [30] or that employs different devices that recognize interactions such as gestures, motions and speech to capture information or focus attention on appropriate displays and material [31] but more research is needed in this field and creativity support tools especially with the rise of new technologies.

In this position paper we investigate how students go through a creative process with computers in a special environment (using olfactory and auditory cues), and their interaction in a creative task such as creative writing. Today technology is shifting the way people express themselves. The pervasiveness of new technologies in our day to day lives, especially digital interfaces, has increased our access to and interaction with the written word, resulting in both more reading and writing. Whether you're script-writing a novel or creating a new story at school, sometimes writing might not be as easy as it seems. Chang et al. [32] refer that a good storyteller usually needs good inspiration during the story construction process. Other researchers [33] have studied the role of storytelling technologies to encourage collaboration and to reflect design suggestions made by children themselves. To improve and foster creative tasks, our study suggests that interactions with multisensory systems can have an important role in a creative process of users.

5 Conclusions

In this position paper we presented a study that addresses the use of different modalities such as sound and smell on the creative writing process of users to enhance their creative activity in four different conditions. This position paper was written with the intention to spark an initial discussion around ambient intelligence and creativity support tools with the simple focus of how to design new tools that involved multisensory interactions, and when to apply the multisensory stimulus in the user's creative process. New creativity support tools could provide ubiquitous solutions whereby new technologies can enhance the user's creative process.

Acknowledgements. We would like to thank the students who participated in this research, as well as the teachers and the institutional support from Escola da APEL. Thanks to ARDITI - Regional Agency for the Development of Research Technology and Innovation through the support provided by the M14-20 Project - 09-5369-FSE-000001- PhD Scholarship. E.F. was supported by FCT MCTES and NOVA LINCS UID/CEC/04516/2013.

References

1. Stephanidis, C.: Design for all. In: The Encyclopedia of Human Computer Interaction, 2nd edn., pp. 2453–2550. Interaction Design Foundation (2015)
2. Carrol, E.A.L.C.: Triangulating the personal creative experience: self-report, external judgments and physiology. In: Proceedings the Graphics Interface 2012 (GI 2012), pp. 53–60. Canadian Information Processing Society, Canada (2012)
3. Robinson, K.: Out of our Minds, Learning to be Creative. Capstone Publishing Ltd., Mankato (2011)
4. Ward, T.: Creative cognition as a window on creativity. Elsevier Sci. Direct Methods **42**(2007), 28–37 (2006)
5. Shneiderman, B.: Accelerating discovery and innovation. Commun. ACM **50**, 12 (2007)
6. Aarts, E., Marzano, S.: The New Everyday: Views on Ambient Intelligence. 010 Publishers, Rotterdam (2003)
7. Friedewald, M., Costa, O., Punie, Y., Alahuhta, P., Hienonen, S.: Perspectives of ambient intelligence in the home environment. Telematics Inform. **22**, 221–238 (2005). Elsevier
8. Cook, D., Augusto, J., Jakkula, V.: Ambient intelligence: technologies, applications, and opportunities. Pervasive Mob. Comput. **5**, 277–298 (2009). Elsevier
9. Dibitonto, M.: New Challenges in HCI: Ambient Intelligence for Human Performance Improvement (2011)
10. Weiser, M.: The computer for the twenty-century, Sci. Am. 94–104 (1991)
11. Hopper, A.: Satient Computing. University of Cambridge, Cambridge (1999)
12. Pantic, M.: Face for ambient interface. In: Cai, Y., Abascal, J. (eds.) Ambient Intelligence in Everyday Life. LNCS (LNAI), vol. 3864, pp. 32–66. Springer, Heidelberg (2006). doi: 10.1007/11825890_2
13. Nakatsu, R.:Integration of multimedia and art for new human-interaction communications. In: Proceedings of the Pacific Rim International Conference of Artificial Intelligence (2002)

14. Xu, R., Mei, G., Ren, Z., Kwan, C., Aube, J., Rochet, C., Stanford, V.: Speaker identification and speech recognition using phased arrays. In: Cai, Y., Abascal, J. (eds.) Ambient Intelligence in Everyday Life. LNCS (LNAI), vol. 3864, pp. 227–238. Springer, Heidelberg (2006). doi: 10.1007/11825890_11
15. Ge, S., Yang, Y., Lee, T.: Hand gesture recognition and tracking based on distributed locally linear embedding. Image Vis. Comput. **26**, 1607–1620 (2008)
16. Elam, J.J., Mead, M.: Can software influence creativity? Inf. Syst. Res. (1990)
17. Hegde R., Walia, G.: How to enhance the creativity of software. In: Proceedings of the 26th IEEE International Conference on Software Engineering and Knowledge Engineering, Vancouver, Canada (2014)
18. Shneiderman, B.: Leonardo's Laptop: Human Needs and the New Computing Technologies. MIT Press, Cambridge (2002)
19. Cherry, E., Latulipe, C.: Quantifying the creative support of digital tools through the creativity support index. ACM Trans. Comput.-Hum. Interact. **21**(4) (2014)
20. Selker, T.: Fostering motivation and creativity for computer users. Int. J. Hum.-Comput. Stud. **63**(4–5), 410–421 (2005)
21. Wang, H., Fussell, S., Cosley, D.: From diversity to creativity: stimulating group brainstorming with cultural differences and conversationally-retrieved pictures. In: Proceedings of CSCW. ACM (2011)
22. Andolina, S., Klouche, K., Cabral, D., Ruotsalo, T., Jacucci, G.: InspirationWall: supporting idea generation through automatic information exploration. In: Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition (C&C 2015), New York, USA (2015)
23. Gonçalves, F., Campos, P., Hanna, J., Ashby, S.: You're the voice: evaluating user interfaces for encouraging underserved youths to express themselves through creative writing. In: C&C 2015 Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition, New York, USA (2015)
24. Engeström, Y.: Activity theory as a framework for analyzing and redesigning work. Ergonomics **43**, 960–974 (2000)
25. Appiah, E., Cronjé, J.: Exploring information and communication theory in graphic design education with activity theory. Int. J. Comput. Appl. **84**, 15–22 2013
26. Bardram, J.E., Doryab, A.: Activity analysis: applying activity theory to analyze complex work in hospitals. In: CSCW, Hangzhou, China (2011)
27. Kuuti, K.: Activity theory as a potential framework for human-computer interaction research. In: Context and Consciousness, pp. 17–44 (1996)
28. InputLog: InputLog. http://www.inputlog.net/
29. iTALC: iTALC - Intelligent Teaching and Learning with Computers. http://italc.sourceforge.net/
30. Shi, Y., Xie, W., Xu, G., Shi, R., Chen, E., Mao, Y., Liu, F.: The smart classroom: merging technologies for seamless tele-education. IEEE Pervasive Comput. **2**, 47–55 (2003)
31. Franklin, D.: Cooperating with people: the intelligent classroom. In: Proceedings of the National Conference on Artificial Intelligence, pp. 555–560 (1998)
32. Chang, Y., Chen, S., Li, T.: A computer system aiming to stimulate creativity for narrative. In: Proceedings in Creativity and Cognition (2011)
33. Benford, S., et al.: Designing storytelling technologies to encourage collaboration between young children. In: Proceedings of CHI2000 (2000)

Time Series Data Mining for Energy Prices Forecasting: An Application to Real Data

Eliana Costa e Silva[1(✉)], Ana Borges[1], M. Filomena Teodoro[2,3], Marina A.P. Andrade[4], and Ricardo Covas[5,6]

[1] Centro Algoritmi, University of Minho, 4800-058 Guimarães, Portugal
{eos,aib}@estg.ipp.pt
[2] CINAV, Escola Naval - Marinha, 2810-001 Almada, Portugal
[3] CEMAT - CEMAT, Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Universidade de Lisboa, 1048-001 Lisboa, Portugal
[4] ISCTE-IUL/UNIDE, 1649-026 Lisboa, Portugal
[5] CMA - Centro de Matemática e Aplicações, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
[6] EDP - Energias de Portugal, 1249-300 Lisboa, Portugal

Abstract. Recently, at the 119th European Study Group with Industry, the Energy Solutions Operator EDP proposed a challenge concerning electricity prices simulation, not only for risk measures purposes but also for scenario analysis in terms of pricing and strategy. The main purpose was short-term Electricity Price Forecasting (EPF). This analysis is contextualized in the study of time series behavior, in particular multivariate time series, which is considered one of the current challenges in data mining. In this work a short-term EPF analysis making use of vector autoregressive models (VAR) with exogenous variables is proposed. The results show that the multivariate approach using VAR, with the season of the year and the type of day as exogenous variables, yield a model that explains the intra-day and intra-hour dynamics of the hourly prices.

Keywords: Data mining · Multivariate time series · Vector autoregressive models · Electricity prices forecasting

1 Introduction

Due to the unique behavior of time series data, the existing research is still considered inadequate and one of the 10 challenges in data mining [12]. In particular, the discovery of interesting patterns, also called motif discovery, is non-trivial and has become one of the most important tasks in data mining. In fact, motif discovery can be applied to many domains [4].

Under the 109th European Study Group with Industry (ESGI)[1], the company EDP - Energias de Portugal submitted the mathematical challenge of simulating

[1] http://www.estgf.ipp.pt/esgi/.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_64

electricity prices not only for risk measures purposes but also for scenario analysis in terms of pricing and strategy. EDP Group is an Energy Solutions Operator which operates in the business areas of generation, supply and distribution of electricity, and supply and distribution of gas. EDP with nearly 14 000 MW[2] of installed capacity in the Iberian Electricity Market – MIBEL[3], is the only company in the Iberian Peninsula with generation, distribution and supply (both electricity and gas) activities in Portugal and Spain.

A large number of methods have been proposed to deal with Electricity Price Forecasting (EPF). In [11] Weron presents, in an exhaustive review article, the complexity of the available methods, revealing their strengths and weaknesses. Weron reduces the existing methods into five major categories: (i) multi-agent models, (ii) fundamental models, (iii) reduced-form models, (iv) statistical models and (v) computational intelligence models. The frontier between these categories is not always clear. For instance, there are several statistical methods that are classified as computational intelligent models. In this paper we contribute with a statistical approach to data mining procedure in the context of EPF.

Most of the EPF approaches consist in methods that forecast the current electricity price by using a mathematical combination of the previous prices and/or previous or current values of exogenous factors, such as consumption and production figures, or weather variables (see [11] for further detail).

Statistical EPF models are mainly inspired from economics literature (for example, game theory models and time-series econometric models). Murthy et al., in [7], present an extremely relevant summary of selected finance and econometrics inspired literature on spot electricity price forecasting (see Table 3 in [7]).

Conejo et al. [2] make use of time series analysis, along with neural networks and wavelets, to predict the 24 market-clearing prices of a day-ahead electric energy market. They clarify that time series techniques, which are a relevant part of the statistical models approach to this king of problems, revealed themselves through many realistic studies as the most efficacious tools for day-ahead market clearing price forecasting.

Synthesizing, in EPF, and confining this study mainly to short term price forecasting, the following models are widely used: the univariate AutoRegressive model (AR); the AutoRegressive Moving Average model (ARMA) – a standard time series model that takes into account the random nature and time correlations of the phenomenon under study; the AutoRegressive Integrated Moving Average (ARIMA) – an extension of ARMA that enables a transformation of the series to the stationary form; or even the seasonal ARIMA model (SARIMA) that captures a possible existence of seasonality. The forecasting of ARMA-type models can be conducted via the Durbin-Levinson algorithm or the innovations algorithm, or by using the Kalman filter for models specified in state space form. As electricity prices can be influenced by the present and past values of various exogenous factors, such as generation capacity, load profiles and ambient

[2] 2012 update and excluding wind power.
[3] http://www.mibel.com/.

weather conditions [11], extensions of the previous mentioned time series models with exogenous or input variables such as ARX, ARMAX, ARIMAX and SARIMAX are used (see e.g. [10] for an excellent manual in Time Series Analysis).

Reported works have already been developed on electricity price forecasting, applying statistical techniques concerning the Portuguese and Spanish electricity market - the market operated by the group EDP. For example, [3] provides a method to predict next-day electricity prices of mainland Spain market based on the ARIMA methodology. Also, [6] apply forecasting factor models to the market framework in Spain and Portugal. More recently, [8] proposes an enhanced hybrid approach composed of an innovative combination of wavelet transform, differential evolutionary particle swarm optimization, and an adaptive neuro-fuzzy inference system to forecast short-term electricity market prices signals in different electricity markets (and also wind power) in Portugal.

In an energy market, the production units are dispatched by considering the bid offers that each agent made for energy. There is not a "consumption" concept "as is". Instead, there are "ask offers" for energy that are made and that are fulfilled only if the price is met. A day-ahead market consists in a system where agents submit their bids and offers for the delivery of electricity during each hour of the next day before a certain market closing time [11].

EPF literature has mainly concerned on models that use information at daily level, however the present analysis is interested in forecasting intra-day prices using hourly data (disaggregated data), and therefore, it is necessary to consider models that explore the complex dependence structure of the multivariate price series. For that, a vector autoregressive structure (VAR) approach has been recently proposed [11].

This paper is organized as follows: Sect. 2 presents the problem statement and a summarized description of the database; Sect. 3 explains the methodology and exposes the main results; finally in Sect. 4 conclusions are drawn and suggestions for future work are pointed.

2 Problem Statement and Database

The daily market electricity prices can be given as a strip of prices (one for each hour of the day), all simultaneously observed once at a given time of each day:

$$Y_t = [y_{1t}, y_{2t}, \ldots, y_{nt}], \quad n = 1, \ldots, 24 \quad \text{(or 23 or 25)}, \quad t = 1, 2, \ldots$$

Therefore, the daily market prices can be interpreted as a multivariate time series. Figure 1 presents the hourly prices for the 31 days of January 2016. This figure suggests a common pattern among the different hours of the day.

The data that will be analyzed in this work consists of disaggregated data, i.e., hourly prices and average day price, from January 2008 to June 2016, in a total 3102×24 observations (see Fig. 2). Particularity, in the present study, data corresponding to the hourly prices from 01/01/2014 to 28/06/2016 is analyzed.

Figure 3 shows the MIBEL daily prices of the 24 h from 01/01/2014 to 28/06/2016, in a total of 21837 ($= 910 \times 24 - 3$) observations. Days 30/03/2014,

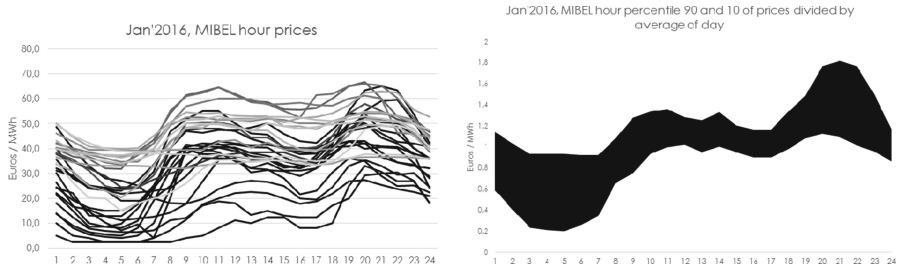


Fig. 1. On the *left*, MIBEL prices from January 2016. On the *right*, 10^{th} up to 90^{th} percentiles MIBEL hourly prices divided by price average of each day.

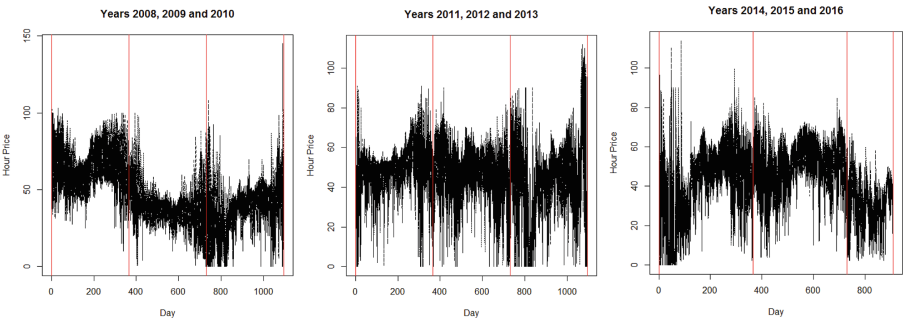


Fig. 2. Hourly prices for the period from January 2008 to June 2016 (data provided by EDP during the 119th ESGI).

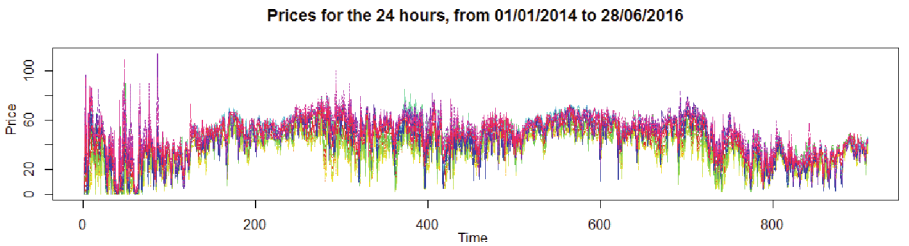


Fig. 3. MIBEL hourly prices for each of the 24 h from 01/01/2014 to 28/06/2016 (data provided by EDP during the 119th ESGI).

29/03/2015 and 27/03/2016 only presented 23 legal hours, therefore, these missing values were filled with the previous hour price value, with the assumption that the current data will be similar to the previous ones (see also [1]). Furthermore, there were several zero hourly prices in the first observations of this time series, more precisely on the first 68 days in 2014. For this reason, the observations for those first 68 days were removed, remaining a total of 20205 observations, that were considered in the present analysis.

3 Model Estimation and Forecast

In this Section a multivariate time series approach is applied to the data corresponding to the hourly prices from 10/03/2014 to 28/06/2016. The data analysis was performed using RStudio (version 0.99.9902) and R Statistical Software (version 3.3.0) [9].

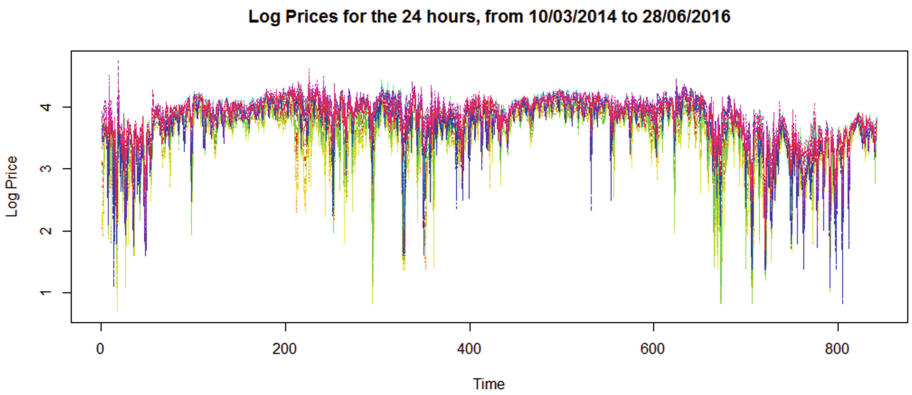


Fig. 4. MIBEL Log hourly prices for each of the 24 h from 10/03/2014 to 28/06/2016 (data provided by EDP during the 119th ESGI).

For stabilizing the variance the log transformation was applied (see Fig. 4). The existence of linear dynamic dependence in the data is supported by the multivariate LjungBox test (not presented here for simplicity). For the Augmented Dickey-Fuller Test one may assume that the time series are stationary (see Table 1).

Table 1. Results for the stationarity tests without the logarithm transformation (*on the left*) and after the transformation (*on the right*). Recall that for the Augmented Dickey-Fuller the alternative hypothesis is stationarity.

	ADF	p		ADF	p
1	-4.65	<0.01	13	-3.89	0.01
2	-5.00	<0.01	14	-3.97	0.01
3	-5.10	<0.01	15	-4.13	< 0.01
4	-4.94	<0.01	16	-4.37	< 0.01
5	-4.86	<0.01	17	-4.25	< 0.01
6	-4.91	<0.01	18	-3.92	0.01
7	-4.98	<0.01	19	-3.49	0.04
8	-4.70	<0.01	20	-3.31	0.07
9	-4.47	<0.01	21	-3.27	0.08
10	-4.21	<0.01	22	-3.44	0.05
11	-3.91	<0.01	23	-3.54	0.04
12	-3.84	0.02	24	-3.89	0.01

	ADF	p		ADF	p
1	-4.96	< 0.01	13	-4.13	< 0.01
2	-5.49	< 0.01	14	-4.21	< 0.01
3	-5.56	< 0.01	15	-4.39	< 0.01
4	-5.39	< 0.01	16	-4.68	< 0.01
5	-5.27	< 0.01	17	-4.53	< 0.01
6	-5.41	< 0.01	18	-4.19	< 0.01
7	-5.38	< 0.01	19	-3.74	0.02
8	-5.00	< 0.01	20	-3.50	0.04
9	-4.71	< 0.01	21	-3.40	0.05
10	-4.50	< 0.01	22	-3.57	0.04
11	-4.07	< 0.01	23	-3.64	0.03
12	-4.03	< 0.01	24	-4.02	< 0.01

In order to explore the intra-day correlation structure, the hourly spot prices, $\boldsymbol{Y}_t = (Y_{1,t}, \ldots, Y_{24,t})^\top$, can be represented as a Vector Autoregressive model of order p, VAR(p) [10]:

$$\boldsymbol{Y}_t = \phi_0 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{Y}_{t-i} + \boldsymbol{a}_t, \tag{1}$$

where ϕ_0 denotes the constant vector, $\boldsymbol{\Phi}_i$ are 24×24 matrices of autoregressive parameters and $\boldsymbol{a}_t$ the residuals. This approach allows to model the hourly spot prices jointly, capturing the correlation between the hourly price, as has been reported in previous studies (see e.g. [11]). The drawback is that it requires the estimation of $24 \times (24 \times p + 1)$.

To capture the seasonal pattern of the process mean, exogenous variables, $\boldsymbol{X}$, were added yielding a VARX model (see e.g. [10]). The general form of a VARX of orders p and s is:

$$\boldsymbol{Y}_t = \phi_0 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{Y}_{t-i} + \sum_{j=0}^{s} \boldsymbol{\beta}_j \boldsymbol{X}_{t-j} + \boldsymbol{a}_t. \tag{2}$$

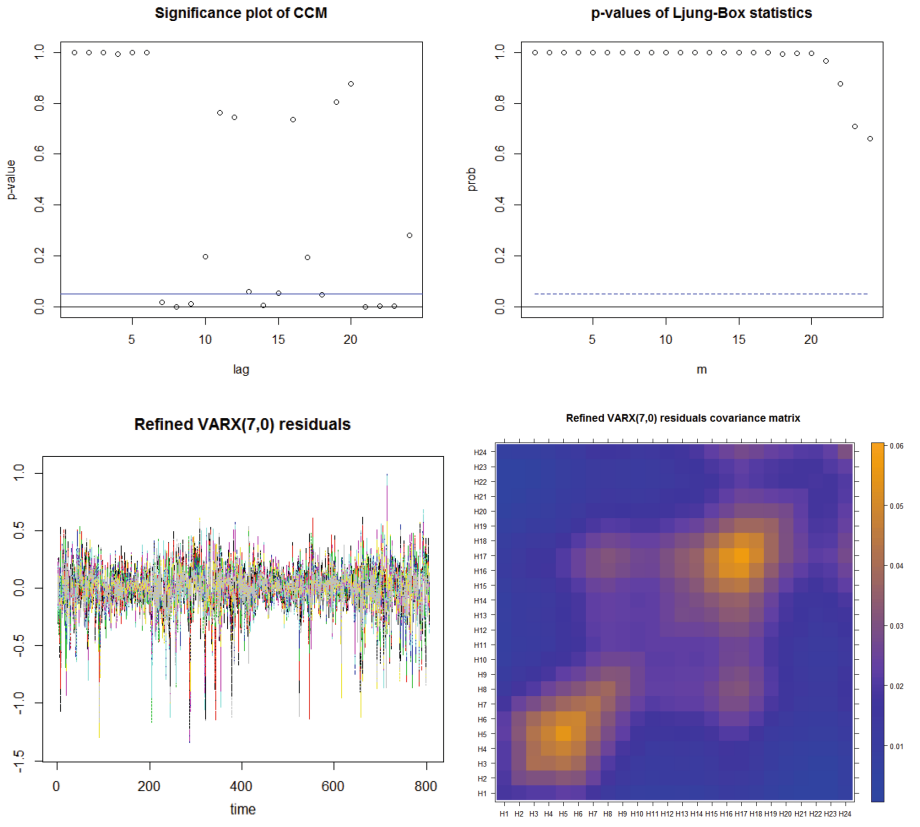


Fig. 5. VARX(7,0) residual analysis.

A similar but simpler model, with $p = 7$, was presented by [5], since $\boldsymbol{\Phi}_i$ were considered as diagonal matrices. In that work the authors use a dummy representing the day type (working day vs. weekend) and the number of daylight hours (in order to mimic the annual seasonality).

In the present analysis, two exogenous variables are considered: (i) day type (working day vs. weekend) and (ii) the annual seasons defined in terms of meteorological conditions[4]. A VARX$(7, 0)$ was estimated using hourly prices Y_{kt}, $k = 1, \ldots, 24$, from 10/03/2014 to 29/05/2016, in a total of 19461 $(= 811 \times 24 - 3)$ observations. Note that $p = 7$ is chosen in several of the EPF studies presented in literature [11]. The prices from 30/05 to 28/06/2016 were used for comparison between the real hourly prices and the forecast of the VARX$(7, 0)$ model. A total of 4104 $(= 24 + 24 \times 24 \times 7 + 24 \times 2)$ parameters were estimated using functions provided by the `MTS` package from R. The full model was reduced by removing simultaneously all estimates with t-ratio less than 0.5. Figure 5 suggests the adequacy of the model. The p-values of the Ljung-Box statistics are all >0.05, and therefore the null hypothesis of zero cross-correlations is not rejected.

The autoregressive coefficients of the VARX(7,0) explain the existence (or not) of dependence within the hourly prices. The element (k, j) with $k \neq j$ of the $\boldsymbol{\Phi}_i$ matrices shows the linear dependence of Y_{kt} on $Y_{j,t-i}$ in the presence of $Y_{k,t-1}$. Figure 6 shows that there are several non-diagonal elements of Φ_i that are nonzero, therefore there are several hourly prices that are dynamically correlated to prices of other hours in previous days (yellow colour).

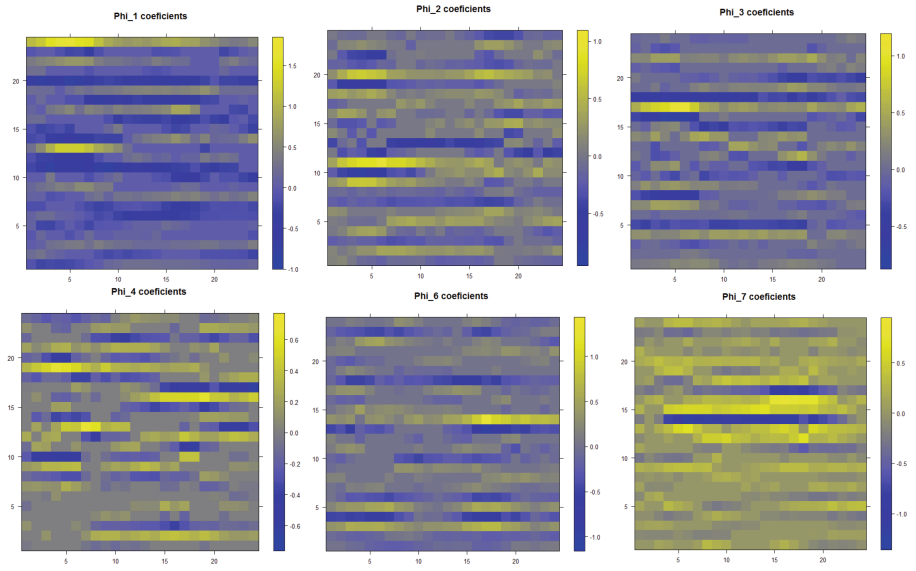


Fig. 6. Autoregressive coefficients of the VARX(7,0). For simplicity the coefficients of $\boldsymbol{\Phi}_5$ were omitted.

[4] See http://www.calendario-365.pt/epocas-estacoes-do-ano.html.

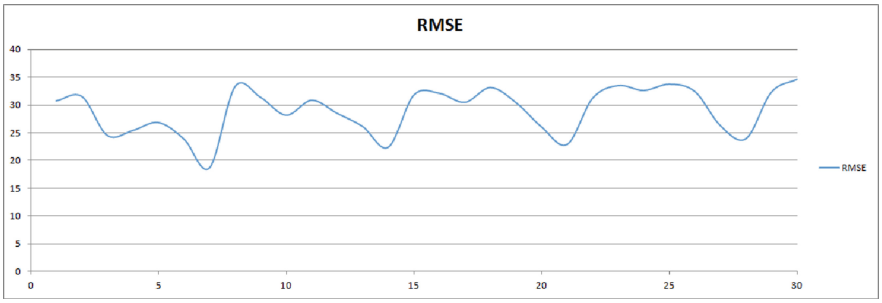


Fig. 7. RMSE for the forecasts from 30/05 to 28/06 using the reduced VARX(7,0).

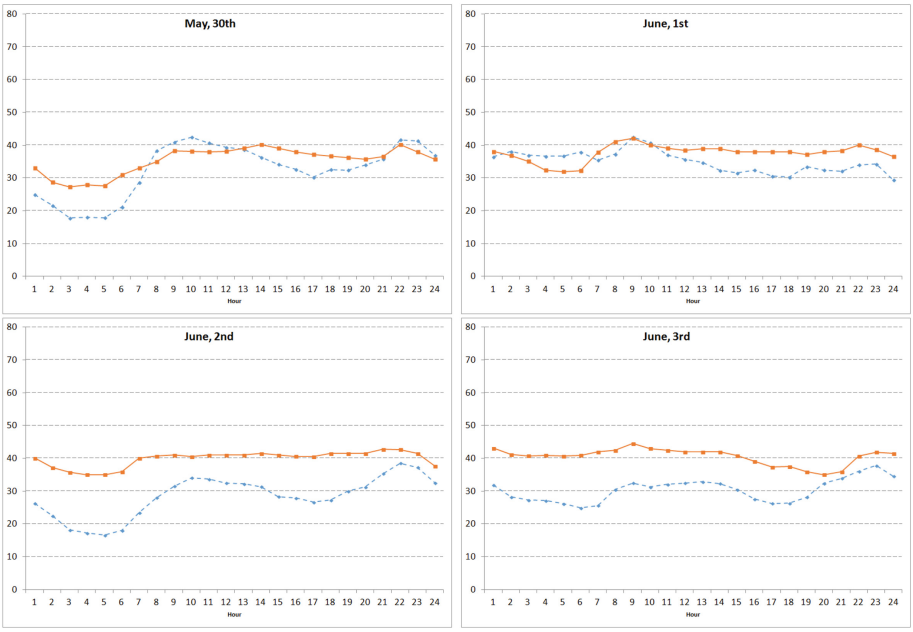


Fig. 8. Real (*orange full line*) and forecast (*blue dot line*) hourly prices using VARX(7,0).

For evaluating the point forecasts the root mean square errors (RMSE) was used. According to [11] this is perhaps the most popular measure for this purpose. Figure 7 presents the RMSE for the 30 days forecast. The forecasts were obtained from a VARX(7,0) estimated using log of the hourly prices. These forecasts were then transform back by applying exp. As an example, in Fig. 8 the real hourly prices and forecasts for the first four days are depicted.

4 Conclusions and Future Work

The challenge proposed by EDP consisted in simulating electricity prices, not only for risk measures purposes, but also for scenario analysis in terms of pricing and strategy. Data provided by EDP, concerning hourly electricity prices from 2014 to 2016, was analysed making use of the VAR methodology.

The multivariate VAR approach, considering as exogenous variables the meteorological seasons and the type of day, yield a model that explains the intra-day and intra-hour dynamics of the hourly prices.

Although the forecast does not exactly replicate the real price they are quite similar. For instance, the introduction into the model of exogenous variables that better explain the meteorological conditions could improve the model estimates. Moreover, further analysis of the orders of the VARX should be performed.

Nowadays data mining approaches to such type of data, consisting in multiple large time series, adopts mainly neural network approaches [4,11]. The analysis here presented aims to contribute with a statistical approach to data mining methodology, particularly in the context of EPF.

Note that, as previously stated, there are several other statistical approaches that can be applied, such as univariate time series analysis. However, several authors that presented models using different univariate approaches, failed in presenting, for example, a diagnostic analysis and selection of the order of the models. Also, to our knowledge, longitudinal modelling has not yet been addressed in EPF, and is an approach that we consider that deserves further attention, if we group identical hourly time series, and thus consider the groups of hourly time series independent.

In conclusions, EPF is a growing area that groups different approaches that can be applied. In fact, other approaches from multi-agent models, fundamental models, reduced-form models and computational intelligence models, such as the statistical approach here presented, should be openly consider for EPF.

References

1. Chujai, P., Kerdprasop, N., Kerdprasop, K.: Time series analysis of household electric consumption with ARIMA and ARMA models. In: Proceedings of the IMECS Conference, Hong Kong (2013)
2. Conejo, A.J., Contreras, J., Espínola, R., Plazas, M.A.: Forecasting electricity prices for a day-ahead pool-based electric energy market. Int. J. Forecast. **21**(1), 435–462 (2005)
3. Contreras, J., Espínola, R., Nogales, F.J., Conejo, A.J.: ARIMA models to predict next-day electricity prices. IEEE Trans. Power Syst. **18**(3), 1014–1020 (2003)
4. Tak-Chung, F.: A review on time series data mining. Eng. Appl. Artif. Intell. **24**(1), 164–181 (2011)
5. Maciejowska, K., Weron, R.: Forecasting of daily electricity prices with factor models: utilizing intra-day and inter-zone relationships. Comput. Stat. **30**(3), 805–819 (2015)

6. Muñoz, M.P., Corchero, C., Heredia, F.J.: Improving electricity market price forecasting with factor models for the optimal generation bid. Int. Stat. Rev. **81**(2), 289–306 (2013)
7. Murthy, G.G.P., Sedidi, V., Panda, A.K., Rath, B.N.: Forecasting electricity prices in deregulated wholesale spot electricity market - a review. Int. J. Energy Econ. Policy **4**(1), 32 (2014)
8. Osório, G.J., Gonçalves, J., Lujano-Rojas, J., Catalão, J.: Enhanced forecasting approach for electricity market prices and wind power data series in the short-term. Energies **9**(693), 1–19 (2016)
9. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2016)
10. Tsay, R.S.: Multivariate Time Series Analysis: With R and Financial Applications. Wiley, Hoboken (2014)
11. Weron, R.: Electricity price forecasting: a review of the state-of-the-art with a look into the future. Int. J. Forecast. **30**(4), 1030–1081 (2014)
12. Yang, Q., Xindong, W.: 10 challenging problems in data mining research. Int. J. Inf. Technol. Decis. Making **5**(4), 597–604 (2006)

Efficient Parameterization for Automatic Speaker Recognition Using Support Vector Machines

Rania Chakroun[3(✉)], Mondher Frikha[2], and Leila Beltaïfa Zouari[1]

[1] Advanced Technologies for Medicine and Signals (ATMS) Research Unit, National School of Engineering of Sousse, Sousse, Tunisia
[2] Advanced Technologies for Medicine and Signals (ATMS) Research Unit, National School of Electronics and Telecommunications of Sfax, Sfax, Tunisia
[3] Advanced Technologies for Medicine and Signals (ATMS) Research Unit, National School of Engineering of Sfax, Sfax, Tunisia
chakrounrania@yahoo.fr

Abstract. Recent advances in the field of speaker recognition have proved to highly outperform algorithms. However this performance degrades when limited data are presented. This paper presents examples on how SVM can improve speaker recognition. The main contribution in this approach is the use of new low-dimensional vectors when training data are limited. We show how different kernels function of Support Vector Machines (SVM) can be used to deal a new approach for speaker recognition system. We achieved remarkable results using TIMIT database.

Keywords: Support vector machines · Limited data · Speaker recognition · Speaker identification

1 Introduction

Speaker recognition is a biometric modality [1] that depends on speech information to determine a person identity [2]. It is widely employed in various kind of applications such as banking over a telephone network, voice mail, voice dialing, database access services, security control for confidential information, remote access to computers, and information and reservation services.

The field of speaker recognition can be divided into two main applications: speaker identification and speaker verification. The task of Speaker identification consists on determining the identity of an unknown speaker based on his/her speech utterances. However, the task of speaker verification concerns the use of the voice of an unknown speaker to verify the identity claimed by the speaker.

The process of speaker recognition operates on two different modes. In fact, recognition of speakers can be text-dependent or text-independent. For text-dependent applications, the system supposes that the presented speaker say exactly an utterance which is determined by the system. However, in independent mode of the text, the speaker is free to pronounce any sentence and there is no constraint on the speech content of the speaker.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_65

Over the past decade, a substantial research has been made on speaker recognition and various speaker recognition algorithms have been developed [3–7]. The well known and most successful techniques are those based on Gaussian Mixture Model (GMM)-based system [26]. However, the performances of these speaker recognition systems have usually been drastically degraded when limited data are presented, which explains the necessity of using new methods to adapt the GMM for this purposes or also the necessity of using new techniques to decrease the problem of short utterances. During recent years, the increasing adoption of SVM for speaker recognition task is well perceived [28, 29]. In this article, we introduce a new speaker recognition system based on Support Vector Machines (SVM) [13, 27] with different type of SVM kernels. The proposed method used a complexity enhancement technique based on new reduced feature vectors able to improve the identification performance of the system when limited data are available for training and resting.

The rest of this paper is organized as follows. Section 2 introduces Support Vector Machines and shows the utility of using SVM for speaker recognition in the literature. The proposed system is described in Sect. 3. Experimental results are demonstrated in Sect. 4 and conclusions are drawn in Sect. 5.

2 SVM for Speaker Recognition

Speaker recognition is the process of authenticating individuals with speaker specific information extracted from his speech signal.

The task of speaker recognition is accomplished through different phases. The fundamental processes are those concerning feature extraction, speaker modelling and decision defining.

In fact, the phase of feature extraction serves to capture from the speech signal the important information that represents each speaker. Typically, the most successful extracted speaker features are Mel-frequency cepstral coefficients (MFCCs) [3]. For speaker modelling, different algorithms were developed to describe the feature distributions of the speakers [3]. Recognition decisions are usually made after comparing the observing test utterance feature vectors with the models of the different speakers trained with the system.

For classification problems, we find that Gaussian Mixture Models (GMM) are widely used in the literature [3, 10]. More recent works assert the increasing adoption of SVM for speaker recognition [11, 27, 29]. This technique has demonstrated to be a novel effective method for speaker recognition [13]. In fact, owing to the kernel which represent the main design component in an SVM, this classifier is able to find an appropriate metric in the SVM feature space relevant to the classification problem [14]. SVM-based systems conduct to comparable or superior performances than those based on GMM with much less training data [15].

The proposed SVM based methods for speaker recognition have demonstrated encouraging performance improvements. However, they still require more improvements when available data are limited. That's why we present in this work a novel

parameterization for SVM-based speaker identification system to improve the performance of the system.

Although the existence of different kernel functions, the following functions are the most known for SVM:

- Linear: $K(x_i, x_j) = x_i^T x_j$
- Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0.$
- Radial Basis Function (RBF): $K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \gamma > 0.$

  Where γ, r and d are kernel parameters.

3 Methodology

In order to evaluate the performance of the proposed system, we start by adopting baseline SVM systems so that comparison can be clearly explicated. The classification is realized by Support Vector Machines (SVM) which proved their efficiencies with regard to the other systems of classification in our domain [3, 18]. We used two SVM baseline systems using two kernel functions in our experiments. The first baseline system uses the linear kernel and the second baseline system uses the radial basis function kernel.

Unlike other works such as Dehak [19] who used NIS-TSRE 2006 corpus [31] where the train and test utterances contain 2.5 min of speech on average, we used short utterances with a mean duration of only 3.28 s from TIMIT database. Besides, we used MFCC features which prove their efficiency in speaker recognition [3] instead of other popular features like Linear Frequency Cepstral Coefficients (LFCC) [3] which are widely criticized because of the non linear character of the speech [21].

The experiments performed with the baseline systems are evaluated using cepstral features extracted from the speech signal using a 25 ms Hamming window with an overlap of 10 ms. Figure 1 explains the process of extracting acoustic feature vectors

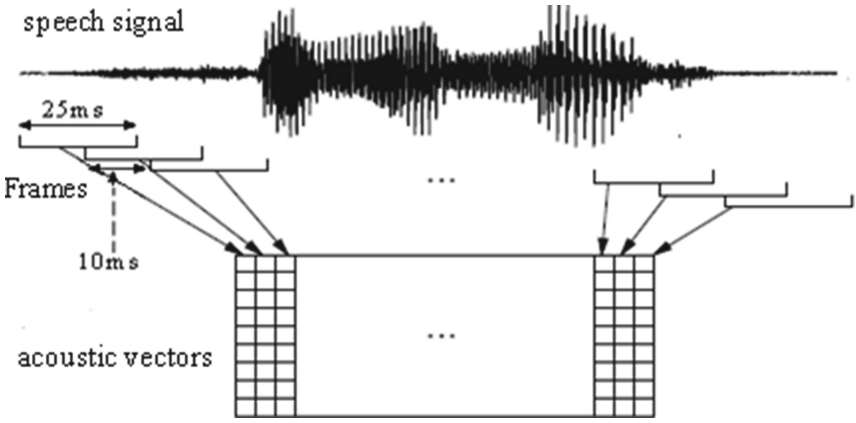


Fig. 1. The process of extracting acoustic feature vectors from the speech signal.

from the speech signal. 19 Mel Frequency Cepstral Coefficients (MFCC) [3] together with log energy are calculated every 10 ms. Delta and double delta coefficients were then calculated to obtain a 60-dimensional final vector. In fact, this feature vector is widely used for the state of the art applications [24, 25, 30].

Unlike our previous works in [32] which try to develop a new approach including additional information, which is the dialect of the speaker to improve the task of speaker identification, we perform in this paper a new speaker identification system based on new reduced feature vectors to enhance the performance of the system. In fact, for the proposed system, experiments are evaluated with different SVM kernel systems with low-dimensional feature vectors and limited training data. In fact, we use reduced MFCC features extracted from the speech signal. Indeed, we used only 12 MFCC together with log energy are calculated using a 25 ms Hamming window with an overlap of 10 ms. Delta and double delta coefficients were then calculated to obtain a 39-dimensional final vector.

In order to increase the effectiveness of the proposed system, we decide to include new cepstral featutes. In fact, in this work, we propose new speaker recognition systems using new features by introducing Cepstral Mean and Variance Normalization (CMVN). Indeed, for a given feature vector of MFCC coefficients:

$$X = \{x[1], x[2], \ldots, x[N]\} \tag{1}$$

The vector presenting the CMVN coefficients is calculated as follows:

$$\hat{x}[n] = \frac{x[n] - \bar{x}}{\sigma_x} \tag{2}$$

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x[n] \tag{3}$$

$$\sigma_x^2 = \frac{1}{N} \sum_{n=1}^{N} (x[n] - \bar{x})^2 \tag{4}$$

Where N represents the number of MFCC coefficients in a feature vector and n is the order of the MFCC coefficient in this vector.

Referring to the protocol suggested in [3], we use 64 speakers from TIMIT corpus for this work. The sentences were subdivided into two subsets: for the first set of experiments, we divide the utterance spoken by each speaker in our work into 8 utterances per speaker for training and 2 utterances for testing. In order to more closely exanimate the system performance with more reduced data, we use only three utterances for the train task and two sentences for the test task for the second set of experiments. We use correct identification rate as metrics for performance evaluation.

4 Experiments and Results

4.1 Test Database

The experiments are evaluated using the TIMIT Dataset. This corpus is comprised of recordings of 438 male and 192 female [16] to obtain a total of 630 speakers coming from eight different regions of the United States and speaking eight different dialects of American English. Each speaker pronounces ten different utterances where the mean duration of the utterance is about 3.28 s.

4.2 Experimental Results

We examine the performance of the speaker recognition systems described previously by carrying out experimental evaluations as follows. The two baseline systems for speaker recognition were implemented and evaluated with 60-dimensional MFCC feature vectors (original MFCC) using the two kind of kernel functions. The two proposed systems are afterward evaluated with reduced MFCC feature vectors using the two kernel functions. In order to more improve the system performance, additional two proposed systems are evaluated with reduced CMVN feature vectors with the two kernel functions. The best speaker identification rate (IR) found for each system with linear and RBF kernels are presented in Fig. 2.

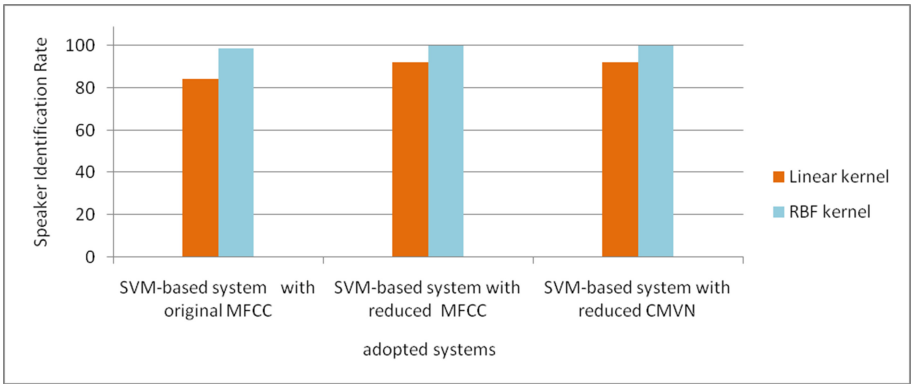


Fig. 2. Speaker identification rates with SVM-based systems with 8 utterances for training and 2 utterances for the test using RBF and Linear kernels

From the experimental results, we notice that the use of the SVM based system with RBF kernel achieves better results than the use of the linear kernel. If we compare between the results obtained with the system using original MFCC feature vectors with the result obtained with reduced MFCC feature vectors, we can remark that our system attain an identification rate of 100% which outperform the baseline system that achieves only 98,43% with the system using RBF kernel. The use of reduced CMVN feature vectors maintains the same performance of 100% of correct identification rate in this case.

For further comparison with more limited data, a second set of experiments was developed with only 3 utterances for training and 2 utterances for testing. The results are illustrated with Fig. 3.

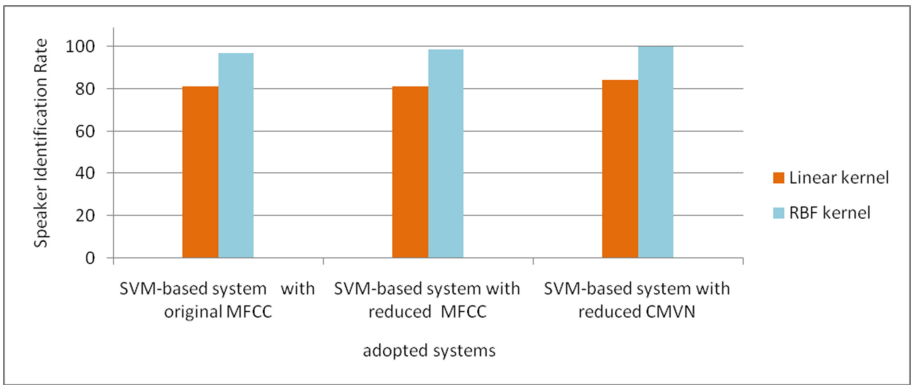


Fig. 3. Speaker identification rates with SVM-based systems with 3 utterances for training and 2 utterances for the test using RBF and Linear kernels.

The obtained results highlight the influence of the use of limited data in the training phase on the system performance. Compared to the results obtained previously it is clear to remark that the system performance degrades when less amount of speech data are used.

If we compare between the results obtained with the baseline system using original MFCC feature vectors with the results obtained with reduced MFCC feature vectors, we can remark that the proposed system achieved an identification rate of 98,43% of correct identification rate which outperform the baseline system that achieves only 96,88% with RBF kernel. The use of reduced CMVN feature vectors help to improve the system performance and achieves 100% of correct identification rate. This system can then compensate the reduction of the amount of speech data and remarkably increase the system performance.

5 Conclusions and Perspectives

In this paper, we present a new method for SVM based speaker recognition systems. This method has focused on the formulation of an algorithm looking for recognizing speakers with very reduced information. In fact we don't need to use important training dataset as in traditional algorithms. Besides, we don't require high-dimensional feature vectors. This is an interesting advantage especially when we need to reduce the learning and time complexity of our system. We have done experiments on two different SVM kernel functions. We obtained better performance in terms of correct identification rate with RBF Kernel than with the linear one and we demonstrated the effectiveness of the use of the new reduced CMVN features when using a limited amount of speech data. We plan to use this approach with other SVM kernels and with different features.

References

1. Jain, A., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Trans. Circuits Syst. Video Technol. **14**(1), 4–20 (2004)
2. Reynolds, D.: An overview of automatic speaker recognition technology. In: Proceedings of the IEEE International Conference on Acoustics Speech Signal Processing (ICASSP), vol. 4, pp. 4072–4075 (2002)
3. Togneri, R., Pullella, D.: An overview of speaker identification: accuracy and robustness issues. IEEE Circuits Syst. Mag. **11**(2), 23–61 (2011). ISSN: 1531-636X
4. Li, Q., Huang, Y.: An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. IEEE Trans. Audio Speech and Lang. Process. **19**(6), 1791–1801 (2011)
5. Wa Maina, C., Walsh, J.M.L.: Joint speech enhancement and speaker identification using approximate Bayesian inference. IEEE Trans. Audio Speech Lang. Process. **19**(6), 5491–5510 (2011)
6. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process. Lett. **13**(5), 308–311 (2006)
7. Xiang, B., Berger, T.: Efficient text-independent speaker verification with structural Gaussian mixture models and neural network. IEEE Trans. Speech Audio Process. **11**(5), 447–456 (2003)
8. Cortes, C., Vapnick, V.: Support vector networks. Mach. Learn. **20**, 1–25 (1995)
9. Kamppari, S.O., Hazen, T.J.: Word and phone level acoustic confidence scoring. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (2000)
10. Reynolds, D.A., Quatieri, T.F., Dunn, R.: Speaker verification using adapted Gaussian mixture models. Digital Signal Process. **10**(1–3), 19–41 (2000)
11. Keshet, J., Bengio, S.: Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods. Wiley, Hoboken (2009)
12. Louradour, J., Daoudi, K., Bach, F.: Feature space mahalanobis sequence kernels: application to SVM speaker verification. IEEE Trans. Audio Speech Lang. Process. **15**(8), 2465–2475 (2007)
13. Campbell, W.M.: Generalized linear discriminant sequence kernels for speaker recognition. In: Proceedings of the International Conference on Acoustics Speech and Signal Processing, pp. 161–164 (2002)
14. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support vector machine using GMM supervectors for speaker verification. IEEE Signal Process. Lett. **13**(5), 308–311 (2006)
15. Dehak, R., Dehak, N., Kenny, P., Dumouchel, P.: Kernel combination for SVM speaker verification. In: Proceedings of the Speaker and Language Recognition Workshop, Odyssey (2008)
16. Reynolds, D.A.: Automatic Speaker Recognition Using Gaussian Mixture Speaker Models. Linc. Lab. J. **8**(2), 173–192 (1995)
17. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: Hidden Markov model toolkit (HTK) version 3.4 user's guide (2002)
18. Jourani, R.: Reconnaissance automatique du locuteur par des GMM à grande marge, UT3 Paul Sabatier (2012)
19. Dehak, R., Dehak, N., Kenny, P., Dumouchel, P.: Linear and non linear kernel GMM supervector machines for speaker verification. In: Proceedings of the Interspeech, Antwerp, Belgium, pp. 302–305 (2007)

20. Reynolds, D.: Speaker identification and verification using Gaussian mixture speaker models. Speech Commun. **17**(1–2), 91–108 (1995)
21. Pitsikalis, V., Maragos, P.: Some advances on speech analysis using generalized dimensions. In: ISCA Tutorial and Research Workshop on Non-Linear Speech Processing (NOLISP) (2003)
22. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification. In: Technical report, Department of Computer Science and Information Engineering, University of National Taiwan, Taipei, pp. 1–12 (2010)
23. Keerthi, S.S., Lin, C.-J.: Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput. **15**(7), 1667–1689 (2003)
24. Dehak, N., Karam, Z., Reynolds, D., Dehak, R., Campbell, W., Glass, J.: A channel-blind system for speaker verification. In: Proceedings of the ICASSP, Prague, Czech Republic, pp. 4536–4539, May 2011
25. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011)
26. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digital Signal Process. **10**(1), 19–41 (2000)
27. Wang, J.C., Lian, L.X., Lin, Y.Y., Zhao, J.H.: VLSI design for SVM-based speaker verification system. IEEE Trans. Very Large Scale Integr. VLSI Syst. **23**(7), 1355–1359 (2015)
28. Mishra, P., Lotia, P.: Speaker recognition using dynamic time warping polynomial kernel SVM with confusion matrix. i-Manager's J. Comput. Sci. **3**(3), 23 (2015)
29. Wang, J.C., Wang, C.Y., Chin, Y.H., Liu, Y.T., Chen, E.T., Chang, P.C.: Spectral-temporal receptive fields and MFCC balanced feature extraction for robust speaker recognition. Multimedia Tools Appl., 1–14 (2016)
30. Mat, P., Cernock, J.H.: Analysis of DNN approaches to speaker identification. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 5100–5104. IEEE, March 2016
31. http://www.nist.gov/speech/tests/spk/2006/
32. Chakroun, R., Zouari, L.B., Frikha, M., Hamida, A.B.: A novel approach based on support vector machines for automatic speaker identification. In: 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), pp. 1–5 (2015)

Security Incident Response: Towards a Novel Decision-Making System

Samih Souissi[1(✉)], Ahmed Serhrouchni[1], Layth Sliman[2], and Benoit Charroux[2]

[1] Telecom ParisTech, Paris, France
{souissi, serhrouchni}@telecom-paristech.fr
[2] EFREI Engineering College, Villejuif, France
{sliman, charroux}@efrei.fr

Abstract. Cyber-attacks have become more complex and unpredictable. Due to their devastating impacts, choosing the appropriate response has become a priority for corporations. This paper introduces an incident response system based on a supervised machine learning model. It offers a framework to process alerts and enhance them to classify and defend against sophisticated attacks. Our method helps security analysts handle alerts and apply the most appropriate response mechanisms, thanks to a high level of abstraction of attack description and supervised learning model. The proposed system is flexible and takes into account several attack properties in order to simplify attack handling and aggregate defense mechanisms. The originality of our work is the ability of this system to provide a response to an attack the system face for the first time.

Keywords: Cyber-attacks · Attack classification · Incident response · Enhanced alerts · Alert processing · Machine learning · Decision making

1 Introduction

In the last few years, the cyber-attack landscape has tremendously evolved lately, as attacks have become more complex and sophisticated. Security deals with protecting company resources from increasing malicious behaviors. The World Economic Forum has recognized cyber-attacks as a global risk in its 2013 Global Risk report [1]. The impact of regulatory interventions, related to the negative public perception and the bad publicity, forces companies to take immediate steps in order to better understand the sensitive information they hold, how it is controlled, and apprehend solutions to detect and respond to the attacks. Most companies are concerned with the protection of confidential data against cyber-attacks, or even human error, and securing IT systems. Thus, security has become a major concern and has gained more interest for companies and corporations. Security aims at protecting firm resources from undesired access by users and applications. Improving information systems security relies on analyzing threats, risks and vulnerabilities, in order to set up suitable countermeasures. One of the new challenges is attack response.

With the growing complexity of attacks and the advent of new ones, many solutions, such as Intrusion Detection and Prevention Systems (IDS/IPS) or Web

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_66

Application Firewalls (WAF), have been used in order to counter these attacks. These solutions are hardware or computer programs designed to detect attacks at a network/information system level, providing information about threats to handle existing breaches. Hundreds of published works [2, 3] on increasing efficiency and reliability of intrusion detection have been published, though there are still areas to explore, especially in incident response, in order to know how to prioritize incidents and what response to apply. In this context, many constraints to security can be found within corporations, related to the heterogeneity of frameworks, platforms, or used software or tools. New architecture, mobility, Bring Your Own Device (BYOD), cloud computing, etc., can widen the perimeter and the type of security that should be set up. Security costs have consequently greatly increased during last few years, but this investment does not guarantee better protection. This situation has led to the emergence of solutions like SIEM (Security Information and Event Management) or SOC (Security Operations Center), in order to analyze different attack alerts, mitigate and then fix the damage caused by the attacks. 2016 SANS Incident Response Survey [4] indicates that the key impediment to effective incident response is staffing and skills shortage. Hence the corporations' vital need for more security experts.

In response to this shortage, creating a decision-making tool that would simplify alert handling and help security analysts is becoming a priority. In attack response systems, the choice of defense mechanisms, policies, decision parameters, as well as attack modeling, is a crucial issue. This paper focuses on a novel machine learning-based decision making system in a heterogeneous environment, facing different attacks and providing appropriate responses. We have already dealt with modeling and architectural issues in a previous work [5, 6]. The objective of this work is to bring a level of abstraction that makes the response of complex attacks more feasible, and the choice of defense mechanisms easier, even for attacks faced for the first time. To this end, we introduce a novel machine learning-based attack response system, while providing a simple approach to describe attacks. Here, we focus on the use of our system in a SOC context.

This paper is organized as follows: Sect. 2 details the related work concerning existing attack response solutions. In Sect. 3, we present our proposition, describing the model and the workflow, with the different functionalities. Section 4 deals with systems feasibility. Finally, Sect. 5 presents the conclusion and perspectives for future work.

2 Related Work

Over the last decade, many solutions and mechanisms have been proposed to detect and respond to computer and network attacks with passive or active response. Most of the deployed systems are IDS and IPS. Snort [7], one of the most widespread IDS, Suricata [8] or Bro [9] allow attack detection more on a packet level through predefined rules to generate events and report by various means. However, it may not be adapted to detect complex attack or to allow the defining of mitigation scenarios. In this case, response is basic and local.

For deeper applicative level detection, WAF are often used. ModSecurity [10] is a signature-based attack detection solution and has relatively good performances. It allows defining advanced responses like virtual patching, although, this system is strongly related to some types of web servers and its policy defining is very complex, needing a high expertise in HTTP protocol and regular expressions. Besides, the attack response is local and does not take into account what happens on other areas of the supervised network.

In [11], Wu et al. propose an attack classification for automatic response systems. Based on this three-dimensional response-oriented classification (Source: attack origin, Technique: method used by the attacker, Result: outcome of the attack), a correspondence matrix for every attack technique is defined, taking into account different sources and results as matrix parameters, so as to define automatic defense techniques. This approach is interesting as the classification helps to describe the attack and allows the aggregation of defense mechanisms. However, blended and complex attacks are difficult to classify and thus to counter. Target type and its criticality are not taken into account.

Simmons et al. propose an issue resolution system IRS [12] based on an ontology and a cyber-attack taxonomy called AVOIDIT [13] used to identify and characterize attacks and their defense mechanisms. Using attack components, a set of metrics are defined and used by an attack defense performance taxonomy. It helps classify and detect blended attacks. It helps make an intelligent decision when defending against attacks relying on a game decision system. However, the taxonomy lacks defense strategies.

In [14], Dasgupta and Gonzalez describe a decision support for IDS system that uses multi-level parameter monitoring. The system observes user, system and process information levels using them in a Genetic classifier-based IDS. It is an adaptive learning system that evolves ruleset, so as to cope with the environment. Rules are generated from a general knowledge base. Genetic algorithms are used following natural evolution metaphor. This system is interesting as it can perform real-time monitoring and analyzing, and provide appropriate response with a global overview of threat landscape. Nevertheless, modifying parameters in order to adapt to defined security policies is not an obvious task. Golling et al. [15] define a multi-layered detection system, which uses a manager that communicates with different types of IDS/IPS: flow-based, protocol-based, statistics-based and DPI-based ones. Each IDS is used based on the data stream to monitor. The manager has an important role within the system as it helps find indications, rate them, investigate them in more details, evaluate result and eventually react to malicious traffic. This system is interesting in detection because it is adapted to high speed networks as the most appropriate detection system is used. Thus, attack detection is faster. However, policy definition in such hierarchical system is not obvious to set up and response is still basic.

Besides, Security Information and Event Management systems, or SIEM, are increasingly used nowadays, due to simplified incident handling. These systems are able to collect logs/alerts from different probes, aggregate, prioritize and correlate them to decrease the number of alerts that the analyst has to deal with. OSSIM [16] and Prelude [17] are widely used open source SIEM. They are rule-based and take into account several security tools (IDS, scanners, probes, etc.). They are interesting as they

allow a global view of the threat landscape within the secured scope, and are easy to configure. Though, attack mitigation and remediation need human expertise. Without it, a response cannot easily be found and can last long. As corporations suffer from a shortage in security expertise, security analysts would benefit much from a decision-making tool that helps guide them to respond effectively to attacks.

Researchers have done promising work in the field of attack detection and automated intrusion response. Nevertheless, no model that covers attack description and response issue using machine learning is entirely practicable. As mentioned before, several challenges need to be faced to have a flexible, complete, expressive, and manageable decision making system capable of responding to attacks.

3 Contribution: Machine Learning Based Security System

In our context, we don't address detection issues related to false positives. We focus on response, and we consider each alert as a real attack that needs to be mitigated. Another hypothesis is that each received alert contains CVE [18] reference of the vulnerability exploited by the attack. This will help enhance and normalize the information contained in the alert.

The challenge is how to guarantee a good and fast attack response or even automatic response decision, while providing accurate attack description and integration simplicity and flexibility. This allows to interface with other security tools and to respond even to new attacks according to a user defined security policy. To overcome these issues, we present in this section our machine learning based security system. This solution should satisfy a set of criteria that will be listed first. Then, we describe our proposition that is mainly composed of two parts: preprocessing functions and supervised learning classifier model. We present the preprocessing part of our system, its different modules, and how it works. Then, we introduce the proposed machine learning approach in order to define appropriate defense mechanisms according to the result of preprocessing functions.

3.1 System Requirements

Figure 1 shows how the decision-making system is used within a SOC. It takes input alerts from different security devices. Input alerts can be already aggregated or correlated before their analysis by our system.

To tackle the different issues mentioned before, our solution should satisfy different criteria:

- **Attack completeness:** the system must be complete, able to provide response to all currently known attacks.
- **Stability:** new attacks must not challenge the decision-making system.
- **Flexibility and Reusability:** our system is independent of the runtime environment, topology and security devices that are used. It can be reused in different contexts, after a period of adaptation.

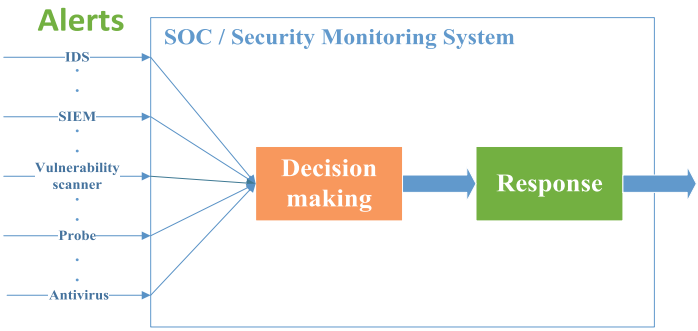


Fig. 1. Decision making system in context

- **Availability:** our system stays available for attack visualization and response purposes. Our system is proactive, as it helps the other areas of the network be aware of what is happening globally.
- **Extensibility:** the user can define his own module to upgrade the system services. He can also add or update different preprocessing functions, learning algorithms, attack scenarios and security policies.
- **Multi-devices:** our system is adapted to different devices. Specification of input from each device is needed. It can handle security tools from different constructors.

Taking into consideration these different requirements, we define our decision-making system.

3.2 Decision Making System

The attack response decision making system, shown in Fig. 2, is responsible for alert analysis, attack description and response. It is composed of preprocessing modules and a supervised learning classification model.

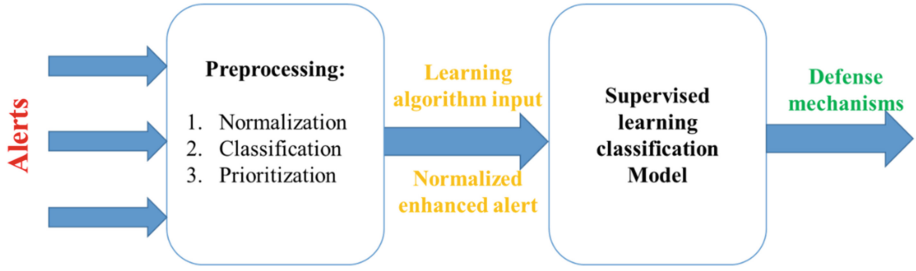


Fig. 2. Decision making system approach workflow

Preprocessing Modules. In order to create the vector containing the features, which are the input for the learning algorithm, alerts should be normalized and enhanced. This is the role of preprocessing modules. By default, three modules can be found:

- **Normalization Module:** Inputs which are mainly alerts, are normalized according to a user defined configuration. It may include an inventory of existing detection and policy enforcement tools within the secured perimeter in order to provide a normalizing algorithm for each one. This module helps us get an analyzer type, source, target in addition to other formatted data that will be examined in the other modules.
- **Classification Module:** It provides the main preprocessing functioning mechanisms. It takes keywords as an input and matches them with those that are related to AIDD classification [2]: attack vector and type of impact. This allows the identification of the attack described by keywords and standardization of attack description. To classify incoming alerts, CVE contained in the alert or other alert information are parsed for classification keywords. The pattern matching algorithm accepts as input a set of strings. The algorithm parses the data and uses the keyword matches within the repository so as to classify new attack vectors. It retrieves the associated classifications for each recognized keyword. The originality of our work consists on adding classification to response, since response is no longer attack-centric, but based on attack classes having generic patterns or behavior for each class. This classification will help respond to attacks that are faced for the first time, but whose behavior or related collected data allow classifying it into a certain category of attack. This approach allows aggregate defense mechanisms. An attack is described by Source, Target, Vector and Impact type.
- **Prioritization Module:** It provides the system with the following information: asset assessment, origins/destinations to be worried about (e.g., attacks from the Internet to the internal network are prioritized, attacks against printers deprioritized, etc.). The criticality of the target and the alert severity using CVSS metrics [19] are used to define the priority in alert handling. We use the NVD-CVE [20] database to retrieve CVSS. Alert frequency is also taken into consideration, as it alters the response to apply.

The output of these preprocessing modules is a normalized enhanced alert that will be the input for the machine learning model.

Supervised Learning Model. A supervised machine learning classifier model is used to match each normalized and enhanced alert with the appropriate defense mechanisms. It can be seen as a defense matching module. After a learning process from different generic use cases where the security policy related to the corporation is applied, the algorithm will be able to match features' vector, thus attack, with a suitable defense. Defense mechanisms are classified into different categories (detection, prevention, response – mitigation or remediation, tolerance, etc.). In our context, to ascertain the feasibility of such model, several classifiers have been tested: Naïve Bayesian, K Nearest Neighbors, Decision Tree, SVM and Neural Network.

3.3 Workflow, Specifications and Example

This system introduces the concept of responding to attacks predefined classes instead of single attack. Our solution provides appropriate response to attacks by operating in the following way: (1) normalize input alerts, (2) extract information from alert (source, target, analyzer type), (3) parse CVE in order to extract attack vector and deduce impact type, (4) find out CVSS related to CVE, (5) look for target criticality and alert frequency (5) create a features vector with collected normalized information (6) select the appropriate defense mechanism [5] and eventually (7) execute the response, though this last step is out of the scope of this work. Adding a response module that interacts with actuators (firewall, policy enforcement points, etc.) does not affect our system as it is modular and extensible. As shown in Fig. 2, the preprocessor takes raw alerts indicating attacks from different devices and generates as output a features vector. This vector is based on our proposed attack classification that we enhance with several features extracted from the alert. This vector is composed from these different parameters: (analyzer type, alert frequency, source, target, vector, impact type, alert severity and target criticality).

- Analyzer type: the analyzer that emitted the alert can provide either an active response or not.
- Alert frequency: the frequency of similar alerts of attacks using the same parameters (source target, vector, etc.).
- Source: location from where the attack is initiated (local machine or network, distant network).
- Target: system's component that is targeted by the attack (OS, network, user, application).
- Vector: technique used to launch the attack. For example, it can be a design flaw, an incorrect configuration, a spoofing, a social engineering, a malformed input, etc. The classification of attack by vector is exposed in [5].
- Impact type: final result of the attack. It can be for example an access or privilege escalation, a denial of service, an information disclosure, etc. The classification of attack by impact type is exposed in [5].
- Alert severity (CVSS): the severity of the vulnerability exploited by the attack that is indicated by the alert. It helps prioritize response.
- Target criticality: Criticality of systems component that is targeted by the attack. It is defined according to business priorities. It helps also prioritize response.

Let's take the example of an alert coming from an IDS coupled with a vulnerability scanner indicating an SQL injection attack targeting a critical Joomla web server containing the vulnerability CVE-2010-4696. Analyzer type, source and target are extracted. Parsing CVE allows to get information about vector, impact type and CVSS. Alert is enhanced by adding alert frequency and target criticality. The features vector (input of the machine learning algorithm) will be: (passive, low, distant network, application, malformed input, resource misuse, high, high). According to these parameters, the classifier model will inform the security analyst that the appropriate defense is to put the server in quarantine and patch it.

4 Feasibility: Experiment and Preliminary Results

Compose language is used in our system [6] to orchestrate different modules and functions. It is based on Spring Expression Language of Spring Framework [21]. Therefore, many expressions can be used to handle the description and the countermeasures of complex attacks such as literal expressions, Boolean and relational operators, regular expressions, class expressions, calling constructors, relational operators and user defined functions. Thus, our orchestration language gives the system a high power of expression, and brings modularity to security controls. This makes policy defining and integration easier. Furthermore, Compose integrates natively with any Remote Code Deposit which supports its APIs. This helps in the automatic deployment of new countermeasure codes.

To test the feasibility of our system, we define 432 different approved use cases related to application attacks. We take into consideration for these use cases 3 classes of defense: Referencing, Patch and Quarantine/Patch. For our simulation, we use the machine learning tool Weka [22]. Using the same data set, we make a cross validation to compute the performances of several classifiers: K-NN (Lazy IBk algorithm with K = 8, which provides the best K-NN performance), Decision Tree (J48 algorithm), Naïve Bayesian, Neural Network (Multilayer Perception algorithm) and SVM (SMO algorithm). Table 1 shows the different metrics of classifiers' performances.

Table 1. Classifiers performances comparison

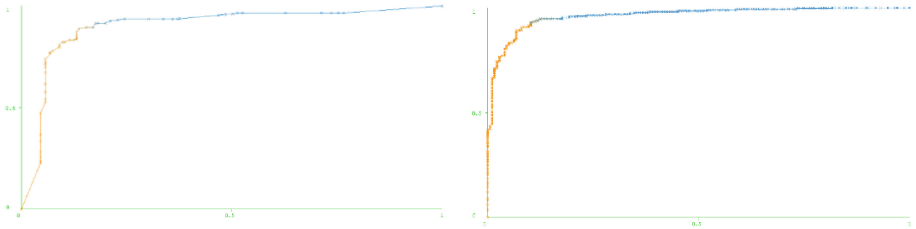
Classifier	Accuracy	Precision	Recall	F-measure
K-nearest neighbors (K = 8)	77.78%	78.3%	77.8%	76.6%
Decision tree	**86.8%**	**87%**	**86.8%**	**86.9%**
Naive Bayesian	75.46%	76.9%	75.5%	73.4%
Neural network	83.8%	83.8%	83.8%	83.7%
SVM	75%	74.7%	75%	74.4%

These different simulations show that the Decision Tree classifier provides us with the best overall performances according to all performance metrics offering an accuracy of 86.8%.

In order to enhance performances, several merging techniques have been tested: Bagging, Boosting and Random Subspace. Using the same dataset with cross validation, the best result is provided by boosting using the AdaBoost algorithm [23] (91.2% accuracy). Concerning other techniques, Bagging ameliorates a little bit Decision Tree accuracy (87.26%), however, Random Subspaces have worse accuracy (69.213%) due to overfitting.

Figure 3 shows ROC curves. The Decision Tree classifier provides a curve with more than 90% area under ROC. AdaBoost offers a curve with more than 96% area under ROC.

Thanks to this system and its machine learning-based approach, security analysts will have a tool to help them decide the response to apply. This system also helps



Decision tree **AdaBoost**

Fig. 3. ROC curves

describe not only known attacks, but also new attacks faced for the first time, thanks to the higher-level attack description.

5 Conclusion

So far, very few response systems have taken into account a machine learning approach for decision making to define attack response. In this paper, we have proposed a novel Machine Learning classifier based decision-making system that helps educate security analysts about the defense mechanisms that should be used when facing an incident. Our preliminary tests indicated a promising solution. Thus, we need to move forward with more detailed research. A key element in our system is that the collected CVE keywords are pertinent to the type of attack vector and can be used generally. Our system not only normalizes attacks description in a higher-level, but also enhances them, while providing appropriate defense mechanisms to each attack category. In addition, our system can provide a response for new attacks, as they are classified in a high-level manner before assigning related defense mechanisms.

In this paper, we focused on the architectural aspect of the solution. The next step is to specify the attack scenarios to be able to mitigate multi-step attacks and advanced persistent threats (APT). A performance study should be done to evaluate the performances in a real life SOC use case, or in a heterogeneous environment like Cloud Computing.

References

1. The Global Information Risk Report (2013). www.weforum.org/reports
2. Sherif, J.S., Ayers, R., Dearmond, T.G.: Intrusion detection: the art and the practice. Part I. Inf. Manag. Comput. Secur. **11**, 175–186 (2003)
3. Sherif, J.S., Dearmond, T.G.: Intrusion detection: systems and models. In: Eleventh IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE 2002) (2002)
4. Bromiley, M.: Incident Response Capabilities in 2016: The 2016 SANS Incident Response Survey. SANS Institute, June 2016

5. Souissi, S.: A novel response-oriented attack classification. In: CFIP-NOTERE Conference, Paris-France, July 2015
6. Souissi, S., Sliman, L., Charroux, B.: A novel security architecture based on multi-level rule expression language. In: Abraham, A., Han, S.Y., Al-Sharhan, S.A., Liu, H. (eds.) HIS 2015. AISC, vol. 420, pp. 259–269. Springer, Heidelberg (2016). doi:10.1007/978-3-319-27221-4_22
7. Snort IDS. http://www.snort.org
8. Suricata IDS. http://suricata-ids.org/
9. Paxson, V.: Bro: a system for detecting network intruders in real-time. In: 7th USENIX Security Symposium, Texas, Lawrence Berkeley National Laboratory, Berkeley (1998)
10. Ristic, I.: ModSecurity Handbook: The Complete Guide to the Popular Open Source Web Application Firewall. Feisty Duck, London (2010)
11. Simmons, C., Ellis, C., Shiva, S., Dasgupta, D., Wu, Q.: AVOIDIT: a cyber attack taxonomy, University of Memphis. In: ASIA Conference, Albany, NY (2014)
12. Simmons, C.B., Shiva, S., Simmons, L.: A qualitative analysis of an ontology based issue resolution system for cyber attack management. University of Memphis. In: Conference on Cyber Technology in Automation, Control and Intelligent Systems, China (2014)
13. Wu, Z., Ou, Y., Liu, Y.: A taxonomy of network and computer attacks based on responses. In: Proceedings of International Conference on Information Technology, Computer Engineering and Management Sciences (ICM) (2011)
14. Dasgupta, D., Gonzalez, F.A.: An intelligent decision support system for intrusion detection and response. In: Gorodetski, V.I., Skormin, V.A., Popyack, L.J. (eds.) MMM-ACNS 2001. LNCS, vol. 2052, pp. 1–14. Springer, Heidelberg (2001). doi:10.1007/3-540-45116-1_1
15. Golling, M., Koch, R., Hofstede, R.: Towards multi-layered intrusion detection in high-speed networks. In: Proceedings of 6th International Conference on Cyber Conflict, Universität der Bundeswehr München Neubiberg, Germany, University of Twente Enschede, Netherlands (2014)
16. AlienVault OSSIM. http://www.alienvault.com/products/ossim
17. Prelude SIEM. http://www.prelude-siem.com/
18. Common Vulnerabilities and Exposures CVE. http://www.cve.mitre.org
19. Common Vulnerability Scoring System CVSS. https://www.first.org/cvss
20. National Vulnerability Database NVD. https://nvd.nist.gov/
21. Srinivasan, K.: Introduction to Spring Expression Language, Spring Framework (2011)
22. Weka3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/
23. Schapire, R.E.: Explaining AdaBoost. In: Schölkopf, B., Luo, Z., Vovk, V. (eds.) Empirical Inference, pp. 37–52. Springer, Heidelberg (2010)

A New Social Media Mashup Approach

Abir Troudi(✉), Corinne Amel Zayani, Salma Jamoussi, and Ikram Amous

MIRACL Laboratory, Sfax University, Sfax, Tunisia
troudiabir@gmail.com, {corinne.zayani,ikram.amous}@isecs.rnu.tn,
jamoussi@gmail.com

Abstract. Social media present a way to discover, report and share different types of events. For this reason, they are considered as a dynamic source of information which enables individuals to stay informed of all real-world events. This specificity encourages researchers to propose methods and approaches to detect events from social media. However many researchers are faced by divers challenges. This is due to the noise of data within the social media. In this paper, we propose a new approach for data retrieval and event detection, which differs from the previous ones by using the mashup concept. The proposed approach is able on the one hand to retrieve data from various social medias which have different structures. On the other hand, our approach aim to detect events with three main dimensions such as the topic, the time and the location.

Keywords: Social media · Mashup concept · Retrieve data · Dimensions · Event detection

1 Introduction

Today, thanks to Web 2.0 and through the social media, users are able to share, criticise and discuss news and real-world events that can be associated to specific populations, countries, companies and even individuals. Social media are therefore a way to discover and bring different types of events. They can be considered as a dynamic source of information enabling individuals, corporations and government agencies to stay informed What is happening now?, Who is involved?, Where is that past?, When did that happen?. The response to these questions represent the dimensions related to events that users want to know.

Events are described in different manners [8] depending on the services and functionalities offered by the social media. In fact, social media have currently different structures therefore they can be classified, according to their features, into several categories [9] including the social networking sites(Facebook, Google Plus), the micro-blogging sites (Twitter, Tumblr), the photo sharing sites (Instagram, Flickr), the video sharing sites (YouTube, Vimeo), etc.

Since their emergence, social media have become a research domain for researchers who try to surround problems occurring. Some researchers focus on solving the problem of cold start [1,3]. These approaches aim to detect the users

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_67

interests and behaviors for recommendation. They are based on the information of users profiles and their relationships with others within the social network. Other researchers [4,5,17] are more interested in the shared data within users profiles than the relationships between them. Their purpose is to detect buzz, and news. Events are passed through a whole life cycle before being presented to the users. This cycle is generally composed by three main steps, the data retrieval step, the event detection step, and the analysis step.

In this paper, we are interested in the data retrieval step and the event detection step. The first step can be done whether retrospectively [1] or online [13] with using APIs offered by social media. In the event detection step, the approaches are classed according to the dimensions which they are focused.

The majority of works provided methods for data retrieval which are based only on one site of social media as a source of information. Generally, users can not cover all social media, they are typically active on only some of them according to their needs or interests. However, they would be informed by all details about news or real-world events propagated in any social media. However, the prohibition of exchanging information between social media prevents users from being informed by any spread news. Furthermore, users waste too much time searching to discover more details about events. To face these challenges, we propose a new approach called Mashup Data Retrieval and Event Detection. This approach differs from the other approaches as it uses the concept of mashup not only to retrieve data but also to detect the events. The proposed approach can on one the hand, retrieve data from various social media which have different structures. On the other hand, our approach aim to detect real-world events in more details using different types of data, such as text, images and videos. The detected events are characterized by three main dimensions including the topic, the time and the location which are called basic dimensions.

The remainder of the paper is structured as follows. Section 2 considers related works. In Sect. 3, we present the concept of mashup. Then, we detail our approach and discuss its architecture in Sect. 4. In Sect. 5, we present our experimental results. Finally, the conclusion and the perspectives are given in Sect. 6.

2 Related Work

Nowadays, social media have become a dynamic source of information that enables individuals to stay informed of all real-world events. For this reason, researchers aim to retrieve data from social media in order to detect events.

The Topic Detection and Tracking (TDT)[1] community defines an event as something that happens at some specific time and place. However the definition of an event in a context of social media differs according to the objective set within the step of the event detection. In fact, event detection represents a significant challenge for many researchers. Since the users share various types of

[1] TDT 2004: Annotation manual, 2004.

contents such as conversations, advertisements, opinions, events and others. It is difficult to detect only the emerging events. Moreover, the messages used in social media are very short, imprecise, often do not provide sufficient information and written in an informal style, such as abbreviations, ungrammatical construction, and spelling mistakes.

Different approaches and methods have been proposed to detect events within social media. They are classified into two types according to the data retrieval task (retrospective or Online). The event detection approaches, based on the retrospective data retrieval are based on data that occurred in the past [1,16]. These approaches retrieve data using APIs offered by the social media. From this data, they build a corpus for event detection. However, the event detection approaches, based on the online data retrieval does not use previously seen messages [7,13]. They retrieve streaming data. It is typically performed in real time.

An event can be described with various dimensions. In fact, some researchers focus on the time dimension and the social dimension [7,12]. Furthermore, the identification of interesting communities of events represents an important dimension to perform the analysis of the events [16]. These approaches are based on construction of homogeneous or heterogeneous graphs to detect communities within social media.

In addition, the increase of using visual data in social media reflects the motivation of researchers to invent other methods for the detection of events through multimedia data. In this context the work of [1], proposes an approach that is based on the detection of events from textual and visual data. The integration of the visual data in the event detection provide improved results.

Furthermore, events are strongly associated with specific geographic areas. In this context, we cite the work of [18], which proposes a system event monitoring which is used to capture events and their geographic locations. Other work intended to distinguish between the protagonists countries and the participating countries relating to events [14]. Some researchers are focused on detecting events of specific event topic such as detecting disasters or earthquake [11]. These approaches are based on specific keywords according to the chosen topic in order to identify events.

The majority of works cited above provide methods for event detection based on data retrieved from social media. They retrieve data from one site of social media which can limit the performance of the detected events. In fact, from a single source of data we can not guarantee the knowledge of all the details of happening events because some dimensions may be missing. This missing can create an ambiguity or a negative influence on the accuracy of the detected event. It is necessary to identify at least the basic dimensions related to the event such as the topic, the time, and the location.

Moreover, most of the cited approaches detect events without considering users behaviors that determine whether the detected information is related to real-world events or not. In fact, the increasing interactivity of the users with some events within social media represents a signal of the importance of these

events. For example, the number of views and likes of videos in YouTube represent the main factor which determines the interest and the importance of the video.

In addition, by using various sites of social media as sources we ensure the diversity of data types such as texts, images and videos. Each social media has its own way to describe events. In fact, Twitter for example, enables the users to post short textual messages. It uses simple text to present real-world events. However Youtube and Flickr use multimedia component to describe events. With this specificity, we can ensure that the event detection based on multi sites of social media helps to have more details about the event, and to include multimedia data for the description of the detected events.

The fact of collecting data from various sources and mixing the textual contents with the media involves the use of the Mashup concept.

3 The Mashup Concept

The Mashup concept has received a lot of attention recently. The term mashup implies easy and fast integration, frequently made possible by access to open APIs and data sources to produce results beyond the predictions of the data owners [10]. The Mashup concept is classified into two types [5]. The first is Media Mashup and the second is Web Application Mashup.

3.1 Media Mashup

The Media Mashup, means the recombination of digital media file containing any or all of texts, graphics, audios, videos, and animations to create a derivative work. It aims at placing each data type in a separate display on the same holding page. Therefore media mashup offers the possibility to view the different result sets on a single web page instead of having to visit multiple web sites. The users are able therefore to visualize various data related to the same query as well as relationships between the different data sources, such as which items are stored on the same web host. For example, [2] used Google maps and a list of top-ranked high schools to build a map with top-ranked high school features directly on the map.

3.2 Web Mashup

Web mashups are typically web applications that are developed by integrating and reusing available data and functionalities accessible over the Web. Mashups can be created by using conventional web programming technologies, dedicated mashup tools such as IBM mashup tools [6], or dashboard platforms such as Netvibes. Similarly, the Mashup concept is integrated likewise in the business domain. In fact, business researchers [15] identified an emerging opportunity for organizations to use technology to exploit information from desktops and the web, in order to react to situational business needs. Despite rapidly increasing

interest in mashups, comprehensive development tools and frameworks are powerful and flexible but suitable only for programmers, and in most cases mashing up a new application implies a significant manual programming effort.

4 Proposed Approach

Following the limits mentioned previously in our state of the art, and based on existing works, we propose a new approach in order to meet the needs of the users of social media and make them able to be informed about the news and actualities in more details.

4.1 Architecture of Proposed Approach

We propose a new approach called Mashup Data Retreival and Event Detection. We present the architecture of this approach in Fig. 1. Through this approach, we aim at retrieving data using APIs from various sites of social media with different structures such as Twitter, Facebook, YouTube, etc. Through this diversity, we focus on different types of data, such as texts, images and videos. Then, we aim at detecting real-world events in more details from the retrieved data according to three basic dimensions including, topic, time and location.

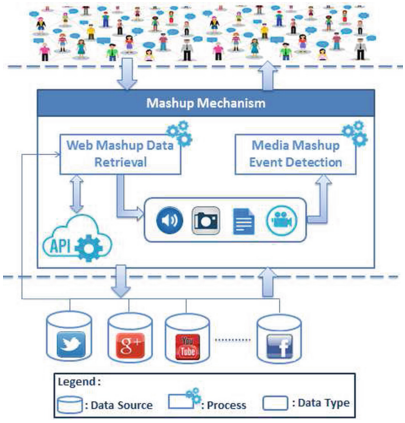


Fig. 1. Architecture of Mashup Data Retrieval and Event Detection

4.2 Web Mashup Data Retrieval

In this step, our objective is to retrieve data from social media in order to use them for event detection. To ensure more details and a variety data type we are based on web mashup which consists in using one or more web services to develop a new application. We choose three sites of social media with different structures

as source of information such as Twitter, YouTube ans Google Plus. This choice is made to ensure using different types of data as texts, images and videos. In order to cover a large scale of events, we select four general domains such as News, Technology, Sports and Entertainment. For each domain we choose four valid accounts of conventional media in each social media. For example, for the domain of News we are based on Aljezeera, France 24, BBCNews and WhiteHouse accounts. Then we retrieved profiles of users who are subscribed on these accounts using the APIs offered by the social Media. In fact, every site of social media offers for the developers an API (Application Program Interface) to help them to retrieve the data. We uses an API Rest and Twitter4j API for Twitter, Google API for Google Plus and YouTube API for YouTube. From the users' profiles retrieved, we generate xml files which represent our corpus used in the next step.

4.3 Media Mashup Event Detection

In this step, we propose a hybrid approach to detect events. Its a combination between retrospective detection and online detection. The goal behind this choice is to ensure on the one hand, the identification of events that have already occurred using the retrospective detection. On the other hand, based on an online detection, we aim to track the progress of detected events to be informed instantly by the news or actualities. The event detection by our approach relies on three main dimensions which are called basic dimensions, including the topic, the time and the location. This naming is due to the importance of these dimensions in identifying and describing an event. In fact, each happening event (topic) is necessarily related to a specific time which represents the temporal dimension and a specific place which represents the location dimension. The proposed approach is composed of three main steps. This approach takes as input xml files containing the users' profiles which are retrieved from the previous step. And generate as output, a list of detected events.

Topic Detection. The objective of this component is to identify tags representing the main topic that describes the events. In this step, we are based on Twitter social media site. In fact, Twitter dominates other social media in providing timely news, which making it the best source of breaking news. Our aim in this step is to detect Hash Tags which represent a key word beginning with "#" and used to indicate the topic of the tweets. Then among these Hash Tags, only ones that have highest degree of interactions of users, such as comments, likes or shares are selected.

Detection of Describing Components. This step has two features the first serves to determine the components that better describe the main tag detected from each social media using co-occurrence algorithm. Whereas the second consists in identifying the tags that refer to places or countries in order to determine the location of events. To identify the location of events we use a gazetteer which represents a geographic dictionary to precise the location tags.

Media Mashup. After detecting the main topic and describing components from each social media used, we apply a digital mashup to combine detected multimedia components for each event and integrate them in order to build a new description of events in more details and different data types.

5 Experimentation and Evaluation

With the increase and the diversity of social media, the users can not cover all the social media sites. They are typically active in only some of them according to their needs or interests. However, they would be informed in details about news or real-world events propagated in social media.

In order to overcome these challenges we propose our approach called Mashup Data Retrieval and Event Detection. In this section we firstly describe our extracted corpus. Then, we present our results in order to test the effectiveness of our approach. Finally we discuss the obtained results.

5.1 Corpus Description

As part of our work we are interested in three social media with different structures. The Table 1 presents a detailed description of our corpus.

Table 1. Description of retrieved corpus

Social media	Social media type	Number of users' profiles	Number of activities per profile
Twitter	Text	1000	From 1 to 3200
Google Plus	Text + Image	1000	From 1 to 150
YouTube	Video	1000	From 1 to 150

As we can see in the table above, we are based on three social media such as Twitter, Google Plus and YouTube. This choice is justified by the variety of structures of these social media. In fact, Twitter is based on texts. Whereas, Google Plus uses the combination of texts and images. For YouTube, videos represents the principal data type used.

In this paper we present an extract of our elaborated corpus containing 1000 users' profiles for each social media in a time interval starting on the 01st of May 2016 and ending on the 30th of July 2016.

As we already explained in Sect. 4.2, we are based on APIs offered by social media to retrieve data. These APIs have restrictions over the amount of data that can be captured without payment. For this reason, we can only retrieving 3200 activity from Twitter as maximum and 150 activity form Google Plus and YouTube.

A preprocessing step was performed to distinguish the trivial data from the popular ones. This step is based on the rate of the users' behavior over an activity. We eliminated all activities that have less than 1000 as a value of the rate of the users' behavior.

5.2 Results

Based on the corpus described in section above, we evaluate our proposed approach detailed in Sect. 4 in the present paper.

This evaluation was performed firstly without using the mashup concept. Due to the unsatisfying obtained results, we integrated the mashup concept in our approach. The Table 2 present the evaluation of obtained results.

Table 2. Comparison of obtained results

Approach	Precision	Recall	F-mesure
Without mashup	0.53	0.48	0.49
With mashup	0.71	0.85	0.78

The table above reports the precision and the F-measure. It should be noted that the precision represents the ratio of valid events from all extracted events, and the F-measure defined as the harmonic mean of precision and recall.

We observe that the use of mashup concept ameliorate the results of detected events. This amelioration is due to the variety of sources and data used within the event detection process.

We present in the table below some examples of detected events through our approach.

Table 3. Examples of detected events

Time interval	Topic	Description	URL media
2016-06-01 10:30:11 2016-06-10 9:58:03	Ali, muhammad, final, realize, life, dies, reality	Boxing legend Muhammad Ali dies at 74	https://www.youtube.com/watch?v=ENv9gqbaRng
2016-07-05 06:24:48 2016-07-12 02:35:59	AltonSterling, racist, usa, cop, shooting, deaths, blacklivesmatter	Alton Sterling, black victims of fatal police shootings in the USA	https://www.youtube.com/watch?v=CaAik-EvI3o
2016-06-04 05:02:20 2016-07-17 12:17:26	Istanbul, victims, airport, ankara, attack, turkey	Istanbul airport attacks	https://www.youtube.com/watch?v=tFU3Z4JhqRo
2016-06-03 19:55:50 2016-06-22 17:15:30	CopaAmerica, goal, game, scored, teamusa, win, messi	Messi vs USA - Copa America Centenario 2016	https://www.youtube.com/watchv=YNO9TdGHw5Q

As we can see in the Table 3, the detected events are described with three dimensions. In fact, each detected event described firstly with a time interval in which the event is happened. Secondly, an event is related to a specific topic. The third dimension is the URL of the media component relative to the detected event.

6 Conclusion

In this paper, we explained the challenges and problems faced by the existing works related to event detection. In fact, the use of only one site of social media as source of information limits the performance of the detected event. This causes on the one hand, the missing details that describe the events. On the other hand, the use of one social media which prevents the diversification of the data types related to the events.

To overcome these problems we proposed a new approach for data retrieve data and detect events. This approach is able on one hand to retrieve data from various social media which have different structures. On the other hand, our approach aim to detect events with main dimensions which are necessary to describe the real-world events. Our approach is based on the concept of Mashup which consists in reusing available data and functionalities in order to create another new application.

The proposed approach presents a hybrid approach which combines the retrospective and the online detection. This combination serves to identify the events that have already occurred and enables us to track the progress of the detected events and be instantly informed about all new events.

The detection of events relies on three main dimensions called basic dimensions, such as the topic, the time and the location. These dimensions are essential to identify and describe the real-world events. In future works we intend to enrich the detected events with analytic dimensions. This enrichment enables the users to know more details about the real-world event.

References

1. Alqhtani, S.M., Luo, S., Regan, B.: Fusing text and image for event detection in Twitter. CoRR abs/1503.03920, 10 (2015)
2. Archambault, L., Tsai, W.T., Crippen, K.: Exploring cyberlearning: inquiry-based mashups combining computer science with stem. In: Koehler, M., Mishra, P. (eds.) Proceedings of Society for Information Technology and Teacher Education International Conference 2011, Nashville, Tennessee, USA, pp. 3867–3874. Association for the Advancement of Computing in Education (AACE), March 2011. https://www.learntechlib.org/p/36935
3. Atefeh, F., Khreich, W.: A survey of techniques for event detection in Twitter. Comput. Intell. **31**(1), 132–164 (2015)
4. Avilés-López, E., García-Macías, J.A.: Mashing up the internet of things: a framework for smart environments. EURASIP J. Wirel. Commun. Netw. **2012**(1), 1–11 (2012)
5. Beemer, B.A., Gregg, D.G.: Mashups: a literature review and classification framework. Future Internet **1**(1), 59–87 (2009)
6. Daniel, F., Weiss, M., Matera, M.: Next in mashup development: user-created apps on the web. IT Prof. **13**(undefined), 22–29 (2011)
7. Guille, A., Favre, C.: Event detection, tracking, and visualization in Twitter a mention-anomaly-based approach. Soc. Netw. Anal. Min. **5**(1), 18:1–18:18 (2015)

8. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: a survey. ACM Comput. Surv. **47**(4), 67:1–67:38 (2015)
9. Kaplan, A.M., Haenlein, M.: Users of the world, unite! The challenges and opportunities of social media. Bus. Horiz. **53**(1), 59–68 (2010)
10. Lee, I.: Encyclopedia of E-business Development and Management in the Global Economy, 1st edn. IGI Global, Hershey (2010)
11. Lingad, J., Karimi, S., Yin, J.: Location extraction from disaster-related microblogs. In: Proceedings of the 22nd International Conference on World Wide Web, WWW 2013 Companion, pp. 1017–1020. ACM, New York (2013)
12. Mezghani, M., On-at, S., Péninou, A., Canut, M.-F., Zayani, C.A., Amous, I., Sedes, F.: A case study on the influence of the user profile enrichment on Buzz propagation in social media: experiments on *delicious*. In: Morzy, T., Valduriez, P., Bellatreche, L. (eds.) ADBIS 2015. CCIS, vol. 539, pp. 567–577. Springer, Heidelberg (2015). doi:10.1007/978-3-319-23201-0_57
13. Osborne, M., Moran, S., McCreadie, R., von Lünen, A., Sykora, M.D., Cano, A.E., Ireson, N., Macdonald, C., Ounis, I., He, Y., Jackson, T., Ciravegna, F., O'Brien, A.: Real-time detection, tracking, and monitoring of automatically discovered events in social media. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, System Demonstrations, ACL 2014, Baltimore, MD, USA, 22–27 June 2014, pp. 37–42 (2014)
14. Quezada, M., Araya, V.P., Poblete, B.: Location-aware model for news events in social media. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015, pp. 935–938 (2015)
15. Simmen, D.E., Altinel, M., Markl, V., Padmanabhan, S., Singh, A.: Damia: data mashups for intranet applications. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, pp. 1171–1182. ACM, New York (2008). http://doi.acm.org/10.1145/1376616.1376734
16. Wan, T., Liu, W., Liu, Z.: A community discovering method based on event network for topic detection. In: Advanced Communication Technology (ICACT), pp. 1242–1246. IEEE (2014)
17. Zhang, D., Hsu, C.H., Chen, M., Chen, Q., Xiong, N., Lloret, J.: Cold-start recommendation using bi-clustering and fusion for large-scale social recommender systems. IEEE Trans. Emerg. Topics Comput. **2**(2), 239–250 (2014)
18. Zhang, Y., Szabo, C., Sheng, Q.Z.: Sense and focus: towards effective location inference and event detection on Twitter. In: Wang, J., Cellary, W., Wang, D., Wang, H., Chen, S.-C., Li, T., Zhang, Y. (eds.) WISE 2015. LNCS, vol. 9418, pp. 463–477. Springer, Heidelberg (2015). doi:10.1007/978-3-319-26190-4_31

Enabling the Definition and Reuse of Multi-Domain Workflow-Based Data Analysis

Rubén Salado-Cid(✉) and José Raúl Romero

Department of Computer Science and Numerical Analysis,
University of Córdoba, Campus of Rabanales, 14071 Córdoba, Spain
{rsalado,jrromero}@uco.es

Abstract. Data analysis applications have become essential to extract significant insight from heterogeneous data sources. However, their development requires technical expertise in computer science techniques like data mining, making its broad adoption by non-experts difficult. In this context, workflows have emerged as a high-level solution to define and automate the sequence of steps involved in the data analysis process, hiding the low-level computational requirements. Existing workflow systems have some difficulties related to their complexity to adapt the provided elements and their inability to reuse workflow definitions. To address these problems, a novel framework for creating customized, ready-to-use and interoperable workflow systems is proposed and prototyped in this paper. Its multi-layer architecture has been designed on the basis of the separation of concerns and the reuse of knowledge assets. As a result, the presented approach allows reducing the time-to-market and saving development costs.

Keywords: Data analysis · Workflow management system · Process automation

1 Introduction

Currently, many companies and organizations are massively collecting data generated by their own information systems (e.g. sensitive personal information, activity logs, transaction records, geographic information) with the aim of getting useful knowledge to optimize the production model, provide better services to users, discover new consumer trends or increase revenue. In this context, computer science techniques like data mining and machine learning provide a set of methods that allow domain experts to create data analysis applications. However, these experts do not usually have the understanding required in programming languages, databases or software development, so corporations have to invest in outsourcing and hiring new well-trained professionals. This is evidenced by the emergence of a technical professional profile called *data scientist* [7].

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_68

Workflows [1] are a suitable technology to bring the creation of data analysis applications to non-technical users, enabling the high-level definition of domain-specific activities and resources necessary to this end. The workflow automation is delegated to a workflow management system (WfMS) [13], e.g. KNIME, RapidMiner, which allows its definition and execution on different domains, such as marketing, health or bioinformatics. To do so, they provide configurable elements in order to meet the requirements of multiple domains, what has caused that a great number of domain experts are interested in its implementation [8]. Nevertheless, the complexity of adaptation and the impossibility of reusing knowledge assets still calls for finding new mechanisms to provide easier, more intuitive and interoperable WfMS.

As a solution, this paper presents a novel framework to generate specific WfMS, making available new ready-to-use and interoperable workflow tools by automatically rebuilding assets for a wide range of domains. In this way, experts would be able to develop and deploy their custom tools requiring less technical skills, instead of having to adapt complex, generic WfMS to each specific case. A multi-layer architecture has been designed to accomplish this goal on the basis of the principles of the separation of concerns. Firstly, a customizable design-oriented environment facilitates the definition of workflow applications. Secondly, a flexible domain-agnostic engine is responsible for the optimal execution of workflow definitions using the available computational resources. Thirdly, a technical meta-tool allows gathering and reusing knowledge assets in order to enable the creation of custom WfMS.

The rest of the paper is organized as follows. In Sect. 2 workflows are introduced as a high-level mechanism for defining data analysis applications, enumerating some well-known WfMS and discussing current issues. Next, Sect. 3 depicts the overall software architecture designed of the proposed framework and, in Sect. 4, the current prototype is presented. Finally, conclusions are outlined in Sect. 5.

2 State of the Art

In this section, the main features and capabilities of the best known WfMS are introduced. Then, the main difficulties faced in the adoption of these systems in a wide range of domains are discussed in terms of complexity and interoperability.

2.1 Workflow-Based Approaches

Workflows are a high-level mechanism to define all the steps required for conducting a particular process. The gathered information is heterogeneous and covers multiple aspects, ranging from the specification of computing programs (e.g. command-line calls, web services), data sources (e.g. files databases), operational structures (e.g. loops, conditional statements), human tasks (e.g. write a document, make a phone call) or even nested orkflow definitions. This latter feature enables the reuse of knowledge assets and building abstract workflows.

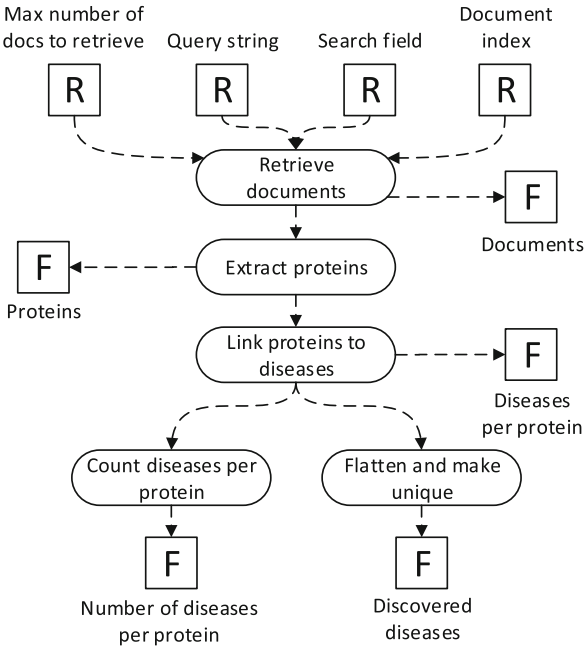


Fig. 1. Workflow in bioinformatics (extracted from [9])

The workflow automation is delegated to a WfMS that consists of software components working together to facilitate the definition of applications and control their execution by a workflow execution engine. The latter interprets and runs the individual elements that constitute the workflow. Traditionally, this technology has been extensively used to improve the performance of both automated and non-automated business processes in order to achieve strategic goals [11]. However, it has gained increasing attention in recent years to manage and execute data-intensive processes [6] like data analysis processes. In this regard, workflows can hide low-level computational requirements, allowing domain experts to concentrate on modeling the specific sequence of activities to be performed instead of focusing on how it would be executed.

Figure 1 depicts a simple, illustrative workflow in a data-intensive domain like bioinformatics. This workflow, which is used to find potential diseases from a number of terms introduced by users, is depicted by square elements that represent both data inputs and outputs (R stands for data existing in memory and F for data stored in a file), rounded rectangle elements that represent domain-specific activities and, finally, arrows that are data dependencies.

There are currently a variety of WfMS to define data analysis applications on different domains. A well-known proposal is KNIME Analytic Platform [2], an open WfMS for data scientists that provides access to a large set of configurable elements for data integration, transformation and visualization, which

allows the integration of Java code too. KNIME has been successfully applied to banking, manufacturing and life science informatics, among other domains. Similarly, RapidMiner Studio [5] is a multi-platform WfMS that includes data mining and machine learning procedures to data loading, transformation, preprocessing and visualization. The number of provided algorithms can be extended using R and Python scripts. RapidMiner has been applied by data scientists and experts in business intelligence to predictive maintenance, marketing, supply chains or political management.

Other WfMS focus on enabling the definition of computational scientific experiments. Taverna [12] is the best-known WfMS to represent and manage scientific workflows, allowing the definition of the resources necessary to accept or reject a specific hypothesis. It provides a set of elements to integrate both remote services and pieces of Java code. Taverna is applied in diverse scientific domains like bioinformatics, cheminformatics, medicine, astronomy or social science.

2.2 Issues of Current Workflow-Based Systems

As mentioned above, workflow technology provides many benefits related to the high-level definition of a particular process and the automation of their execution by a WfMS. However, there are still some challenges to face in order to extend their use in a wider range of domains. Firstly, the complexity inherent to the adaptation of an existing WfMS to meet the requirements of different domains. As a way to mitigate development costs, existing WfMS provide a set of configurable domain-agnostic elements that enable access to diverse types of resources, such as web services, data sets, computing programs or even source code. However, their configurability is limited to application domains with similar requirements, still requiring a lot of development effort and technological skills for their adaptation. Hence, domain experts demand flexible, ready-to-use elements, as well as more abstract and intuitive development methods in order to increase flexibility and reduce design time.

On the other hand, the inability to reuse knowledge assets, even in fields with similar requirements, as explained in [3]. The lack of interoperability prevents the collaboration between different WfMS to create and execute applications, since the domain-expert knowledge captured by workflows (e.g. activities, data sources, operational logic) is constrained by their particular features. It causes that each specific workflow definition can include features that could be incompatible with, or even missing in, different tools. Thus, a well-defined language comprising the main features required to define data analysis applications would offer new opportunities for increasing the flexibility, portability, easiness to learn and interoperability of the resulting workflows.

3 Framework Rationale and Design

A framework has been conceived to generate customizable WfMS and facilitate the use of workflow technology. Its architecture and main software components are presented next.

3.1 Overall Architecture

The need of providing easier and more interoperable WfMS has led to the design of a modular and adaptable framework. The main idea behind its conception is to reach a clear separation between the configuration process required to adapt WfMS to any application domain and how workflow applications should be defined and executed. As a result, its architecture (see Fig. 2) is divided into two well-differentiated layers, whose components are intended for different types of user with possibly different skills: a customizable WfMS for domain experts, supposedly non-technical users, that creates and executes data analysis applications involving ready-to-use activities and resources; and a meta-tool for domain experts, data scientists and IT professionals, supposedly with some technical expertise, that serves to collect, configure and reuse knowledge assets (e.g. workflow definitions, web services, source code).

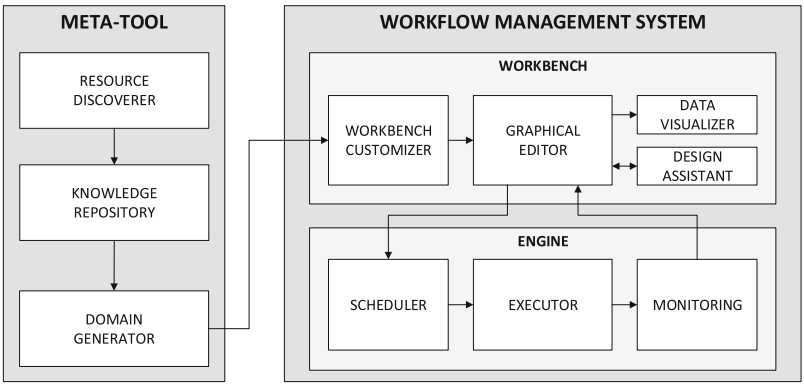


Fig. 2. The overall architecture of the proposed framework

On the one hand, the customizable WfMS is composed by two major components that work together but separately. Firstly, the workbench environment simplifies the definition of workflow applications by providing custom activities and enabling its complete configuration. Secondly, the scheduling and execution of the planned activities are delegated to a workflow execution engine. It considers the different constraints related to the availability of computational resources, data formats, security issues, parallel and distributed computing with the aim of improving the computational power of the underlying execution platform and reducing execution time. On the other hand, the meta-tool layer allows reducing the time required to deliver a new specific WfMS, gathering and reusing pieces of knowledge assets from different workflow definitions with aim of integrating them into the customizable WfMS.

The two layers and their inner components are required to be independent, mutually separated, and easily replaceable. Therefore, the formal definition of a precise workflow language for the representation of data analysis applications is

required to ensure the communication between the different elements. This way, scalability, maintainability and reusability is formally obtained by following these design criteria.

3.2 Workflow Language for Data Analysis

A domain-specific language (DSL) [4] has been formally defined to provide the concepts, relationships and constraints related to the construction and management of workflows. The language allows representing workflow-based data analysis applications for the configuration of domain-agnostic activities, the integration of resources and the use of model-driven engineering (MDE) techniques [10] in order to increase the flexibility, portability and interoperability. Moreover, it is extensible and adaptable to the particular requirements and constraints of each domain. Due to the limited space, Fig. 3 shows a partial view of the language elements:

Workflow is defined as a set of nodes (*Constructor*) and links (*Connection*), which arrange the sequence of steps followed by data analysis application. A node is composed by different connection points that define both the expected data type and the compatible connection type. It can be specialized in an activity (*Activity*), data source (*DataSource*) or control structure (*ControlStructure*). On the other hand, a link sets a directed connection between a source node and a target node.

Data source includes the elements that serve to gather and store the data managed by the workflow. Depending on the location of the data set, a data source can be defined as a record (*Record*), if data are available from memory, or a file (*File*), if data are stored into a file.

Activity provides the definition of a domain-agnostic computational task. It is a data intensive task if it transforms a large set of input data into new output data without human intervention. This type of activity is specialized as a local

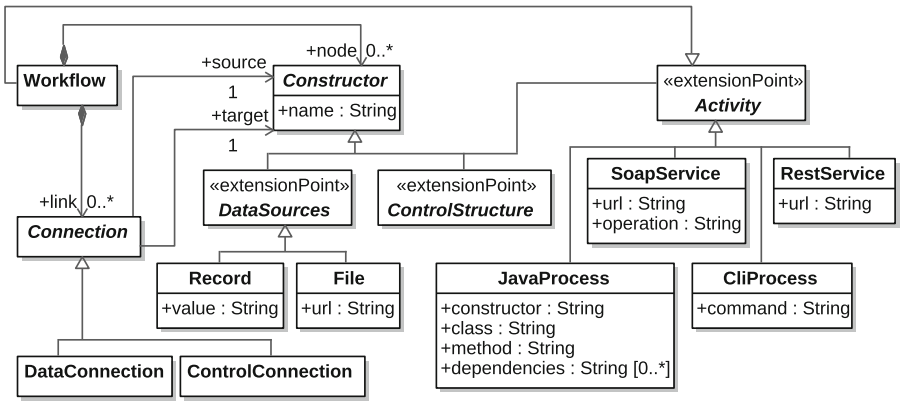


Fig. 3. A partial view of the language elements for defining data analysis workflows

computing intensive task if the process is executed in the local host, such as a Java process (*JavaProcess*), a command-line interface program (*CliProcess*) or a previously defined nested workflow (*Workflow*). On the other hand, a remote computing intensive task invokes an external service accessed using REST (*RestService*) or SOAP (*SoapService*). Finally, a computational task is also considered as a visualization task if it is intended for the graphical visualization of data.

Control structure includes elements to manage the execution flow from a programmatic perspective. Different types of control structures are considered: a conditional structure to determine the next activity to be executed according to some particular condition; a loop structure to create recursive executions; or an explicit parallelization structure to simultaneously execute two or more activities, among others.

4 Prototype

This section introduces the early development of the proposed framework performed according to the multi-layer architecture presented in Sect. 3.

4.1 Workbench Environment

The workbench environment enables the visual representation of data analysis workflows and the monitoring of their execution by domain experts. As shown in Fig. 4, this environment is composed by a set of modules, where the graphical editor is the core element. This editor provides a palette containing ready-to-use

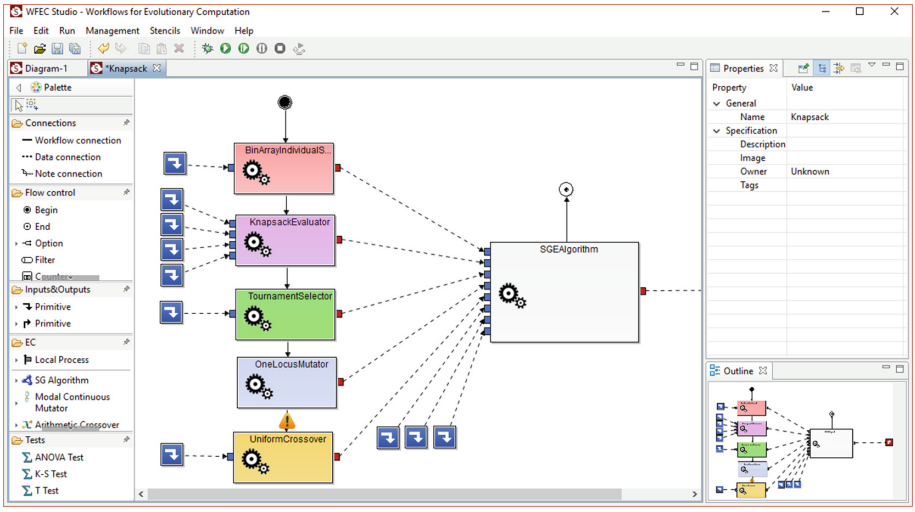


Fig. 4. A workbench environment adapted to the field of evolutionary computation

activities and elements to design the application in accordance to the workflow language concepts, a canvas to compose the corresponding data analysis application as a set of pluggable blocks, a property sheet to inspect the configuration of each element in a quick way, and an outline area to show the whole workflow representation. In addition, a design assistant checks the constraints defined by the workflow language to help domains experts to properly connect and use the activities provided by the graphical editor.

A data visualizer has been developed to graphically show the data processed inside the workflow. Depending on the data type, the information can be directly displayed by the canvas (e.g. numbers, strings) or delegated to an advanced visualizer (e.g. images, video, data sets) using different types of graphs and tables. In the latter case, the design and style are configurable with multiple options. Finally, the workbench customizer is responsible for adapting both the user interface and the palette, including texts, graphics, icons and particularized assets for an user-specific application domain.

4.2 Execution Engine

A workflow application built using the workbench is managed by the workflow execution engine, which interprets and executes the involved activities. To this end, the engine provides all the features related to the invocation of services (both local and remote), the local execution of computing programs and the management of data.

Different modules are responsible for performing the workflow execution. Firstly, the scheduler analyzes the high-level workflow definition according to the workflow language in order to generate the corresponding low-level executable version. This module coordinates and optimizes the execution taking into account the computational constraints and the workload of the executor. Then, the executor module uses this low-level executable version to run the corresponding computing programs or services. The executor is flexible to integrate external tools and platforms (e.g. for grid and distributed computing), increasing the computational capabilities and reducing the execution time. Finally, the monitoring module logs the execution to provide information related to time, available memory and outcomes in order to monitor how resources are managed and consumed, and how resulting values are generated.

4.3 Meta-Tool for the Assembly of Domain-Specific Elements

The meta-tool has been developed to facilitate the fast creation of customized WfMS based on the reuse of knowledge assets (Fig. 5), providing ready-to-use workflow technology for any type of domain in a reasonable time frame. This meta-tool is intended for users with the technical expertise required to adapt generic and configurable elements.

One of the three modules that compose the meta-tool is the resource discoverer, which allows the user to both automatically discover or manually define knowledge assets. These assets contain the necessary information to configure

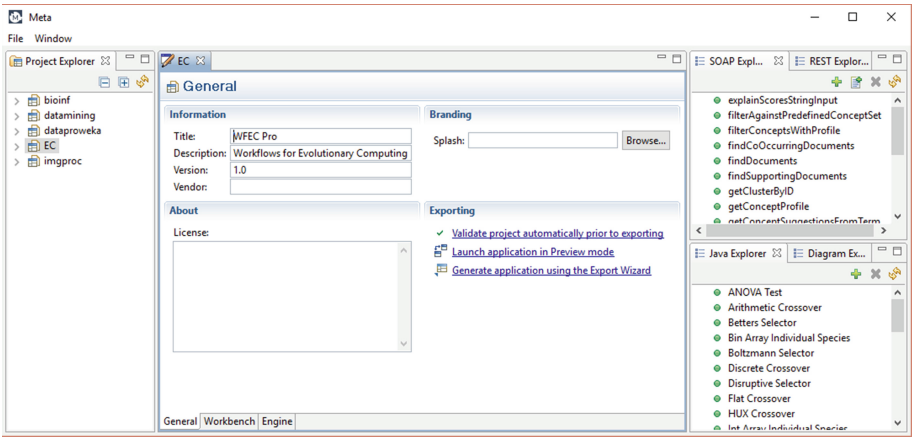


Fig. 5. Meta-tool to create custom WfMS

the domain-agnostic elements captured by the workflow specification in order to meet domain-specific requirements (e.g. to access a particular web service). Then, the reuse of these particularized knowledge assets is possible thanks to a repository that stores them all. Finally, the domain generator module is responsible for assembling all the required assets from the knowledge repository for a particular development, as well as setting the workbench parameters used to configure the user interface of the customizable WfMS. This entire process allows significantly reducing both the time to value and the configuration time spent to create workflow-based data analysis applications.

5 Concluding Remarks

Building data analysis applications by domain experts without computing skills requires more intuitive and interoperable tools to facilitate their development. In this paper, a novel framework has been presented to create interoperable, custom WfMS. This framework clearly separates the domain-specific knowledge needed to define data analysis applications from the technical background required to configure and adapt domain-agnostic elements in order to meet domain-specific requirements. To reach this goal, a multi-layer architecture has been developed based on highly decoupled and adaptable modules. A customizable workbench allows domain experts to define their workflow-based data analysis applications by integrating ready-to-use elements and resources. An execution engine is responsible for automatically running the workflow specification taking advantage of the available computational resources in a efficient and transparent way. Finally, a meta-tool for users with technical knowledge has been developed to define, store and reuse particularized knowledge assets (e.g. workflow definitions, resource descriptions) by configuring domain-agnostic elements. As a result, this meta-tool enables the creation of new custom workflow systems for

any type of domain, saving development costs and reducing the time to market. Additionally, the resulting systems are interoperable tools thanks to the formalization of a workflow language for data analysis.

In the future work we plan to validate the proposed framework with industrial users, generating WfMS adapted to well-known application domains like bioinformatics or education. Moreover, the integration of Big Data technologies, such as Apache Hadoop or Apache Spark, in order to specify and execute Big Data applications will be explored.

Acknowledgments. Work supported by the Spanish Government, project TIN2014-55252-P.

References

1. Terminology & glossary. Technical report, WFMC-TC-1011, Workflow Management Coalition (1999)
2. Berthold, M.R., et al.: KNIME: the Konstanz information miner. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin (2007)
3. Elmroth, E., Hernández, F., Tordsson, J.: Three fundamental dimensions of scientific workflow interoperability: model of computation, language, and execution environment. Future Gener. Comput. Syst. **26**(2), 245–256 (2010)
4. Fowler, M.: Domain Specific Languages, 1st edn. Addison-Wesley, Boston (2010)
5. Hofmann, M., Klinkenberg, R.: RapidMiner: Data Mining Use Cases and Business Analytics Applications. Chapman & Hall/CRC, Boca Raton (2013)
6. Liu, J., Pacitti, E., Valduriez, P., Mattoso, M.: A survey of data-intensive scientific workflow management. J. Grid Comput. **13**(4), 457–493 (2015)
7. Loukides, M.: What is Data Science?. O'Reilly Radar, Sebastopol (2010)
8. Recker, J., Rosa, M.L.: Understanding user differences in open-source workflow management system usage intentions. Inf. Syst. **37**(3), 200–212 (2012)
9. Roure, D.D., Goble, C., Bhagat, J., Cruickshank, D., Goderis, A., Michaelides, D., Newman, D.: myExperiment: defining the social virtual research environment. In: 4th IEEE International Conference on e-Science, pp. 182–189. IEEE Press (2008)
10. Schmidt, D.C.: Guest editor's introduction: model-driven engineering. Computer **39**, 25–31 (2006)
11. Weske, M., van der Aalst, W., Verbeek, H.: Advances in business process management. Data Knowl. Eng. **50**(1), 1–8 (2004)
12. Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., Bhagat, J., Belhajjame, K., Bacall, F., Hardisty, A., de la Hidalga, A.N., Balcazar Vargas, M.P., Sufi, S., Goble, C.: The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. Nucleic Acids Research **41**(W1), W557–W561 (2013)
13. Yu, J., Buyya, R.: A taxonomy of workflow management systems for grid computing. J. Grid Comput. **3**(3), 171–200 (2006)

An Improved Elman Neural Network for Daily Living Activities Recognition

Zaineb Liouane[1,2(✉)], Tayeb Lemlouma[2], Philippe Roose[3], Fréderic Weis[2], and Hassani Messaoud[1]

[1] LARATSI, Monastir University, Ibn El Jazzar, Monastir, Tunisia
liouane.zaineb@yahoo.fr
[2] IRISA, Rennes1 University, Rennes/Lannion, France
[3] LIUPPA/T2I, Pau and the Adour Countries University, Anglet, France

Abstract. One of the main issues regarding the monitoring of persons in a smart home environment is the accuracy of the daily control of the person, the health prevention and the timely prediction of abnormal situations. To tackle this problem, this work proposes the use of an improved version of the Elman Neural Network (Elman-NN). In order to minimize the error between inputs and desired outputs, we optimize some criteria of the network to gain good results. We propose to use the Differential Evolution algorithm in the learning step of the Elman-NN to evolve the error performance. Our proposed model is responsible to predict the activities of the monitored elderly and to detect any state changes. This hybridization will help to optimize the weight and the bias of the network to achieve our objective function and to obtain a good network. The experimental results reveal that the proposed model is satisfactory for elderly person's movement prediction.

Keywords: Elman neural network · Differential evolution algorithm · Prediction model · Smart home · Elderly person

1 Introduction

The number of the elderly increased considerably each year. Aged people represent a growing part of the world population [1]. Since 2012, the number of older had increased to almost 810 million [2]. This situation motivates many research studies to focus a considerable interest on this aging population in order to provide required help and assistance. Smart home environments represent an interesting and challenging way where such needed services are provided automatically (or semi automatically) by the smart system of the home. In a smart home, the occupant is continuously monitored using a set of sensors and actuators. The sensors monitor the activities of the space's occupant and timely inform the system about the services needed depending to the current situation of the person and his health state. In order to ensure an efficient monitoring that considers the current context of an elderly and predict possible changes, the system must provide a better recognition of the human activities, scenarios, health state

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_69

and possible emergency situations. Several models and methods have been used to study the data issued from various sensors existing in a smart home. For instance, Support Vector Machines (SVMs), Markov Models [3], Artificial Neural Networks, Deep Neural Networks, Bayesian Network, etc. Unfortunately, the conventional previous methods did not succeed to solve the problem of the accuracy and the rate of certainty remains limited.

The objective of this work is to overcome the problem of the prediction accuracy. We focus on three features: the identification of the adequate recognition model to be used by selecting the appropriate artificial Neural Network, the architecture of the network and the learning algorithm. In order to build a new improved model with an efficient rate of prediction, we investigate the improvement of the Elman Neural Network (Elman-NN) [4] by using a learning algorithm belongs to the evolutionary algorithms there is the differential evolution (DE) algorithm. In order to accurately predict the monitored person's activities, actions and scenarios, we propose a prediction model based on Elman-NN and DE with multiple inputs. The advantages of the DE algorithm are its simple structure, ease of use, speed and robustness. DE is one of the best genetic algorithms for solving the issues of application which deal with real valued variables. The Elman-NN, based on the DE algorithm is used during the learning step and for the prediction of activities daily living (ADL).

The remainder of the paper is organized as follows: in Sect. 2, a background of the prediction model is presented. In Sect. 3, we describe the Elman-NN and the DE algorithm and present our proposed model. Section 4 presents the data source used in this work. The effectiveness of the Elman-NN based on the DE is discussed through simulations in Sect. 5. Section 6 concludes this work.

2 Related Works

A large amount of the existing literature focuses on recognition models to monitor the activity of daily living of persons evolving in a smart home environment. Much research studied artificial networks in order to solve the problems of prediction which is important in such environment to predict abnormal and risky situations.

Callaghan *et al.* [5], used the Echo State Network (ESN) to predict future data of time series in an intelligent space. Their results show that the training step of ESN is much faster than other kind of Recurrent Neural Networks (RNN). Fang and He [6], proposed to use a Back Propagation Neural Network (BPNN) to recognize the activities of daily living in smart home. They used a feed forward strategy in the learning step and showed the impact of the number of used neurons in the error rate of accuracy. In the learning step of the Stable Neural Network (SNN), Teich *et al.* [7] proposed to use a feed forward neural network (Feed FNN) algorithm. The Feed FNN algorithm with a multiple hidden layers showed promising results. Oniga and Suto [8, 9], used an ANN to predict the body posture of the person. In [9], the authors used two layer feed-forward network and a back-propagation algorithm in the learning step. Oniga and Suto [9] concluded that for the hidden layer, 10 neurons and 5 neurons in the output layer give a good recognition rate. Xu and Mao [10], proposed the application of particle swarm

optimization algorithm to improve the Elman neural network in order to predict the wind power. The authors concluded that the proposed improved Elman-NN can overcome the problem of the local optimum and achieve a good prediction result with a high accuracy.

Chong *et al.* [11], proposed to use a new model based on Elman-NN to improve the temperature drift modeling precision of a tuning fork micro-electromechanical system (MEMS) gyroscope. The results show that the multiple input/single output for Elman-NN based on GA can guarantee a high accurate modeling.

Previous related works show that the use of several types of the ANN promises good results in different domains, but most of the previous search use a deterministic algorithm in the learning step of the ANN such as a feed-forward algorithm, a back-propagation algorithm etc. Unfortunately several search shows that there are many inconvenient in training neural network with learning deterministic algorithms (gradient type) such as: The evaluation criterion is a differentiable function. In fact, when optimizing this criterion, a change of parameters is established in the direction of the gradient evaluation criteria from this parameter. The existences of a large number of sub-optimal solutions correspond to the local optima in the network. Deterministic algorithms are unable to give up a successful global optimum and do not confirm the completeness of network.

To overcome the previous limitation we propose to use the differential evolution algorithm in the learning step, it is non deterministic algorithm and it is autonomous (not require human intervention for the learning step). It can overcome the problem of local optima and leads to a global optimum with a high probability (thanks to their stochastic) and they have the property of completeness also we investigate the improvement of a specific type of the ANN that is the Elman Neural Networks for ADL recognition.

3 The Proposed Model

3.1 Elman Neural Network (Elman-NN)

The Elman Neural Networks (Elman-NN) was proposed by Elman in 1990 [4]. Elman-NN is a dynamic recurrent network and is characterized by a multiple feed forward topology. Its architecture includes the input layer, a particular context node input layer, the hidden layer and the output layer. The main advantage of Elman-NN is that the context nodes can be used to memorize the previous data of the hidden nodes, which makes Elman-NN applicable in the fields of dynamic system identification and prediction control [12]. The structure of the Elman-NN can be presented in Fig. 1.

The network is characterized by a triplet (W, B, f). Where W is the weight, B is the bias, and f is the transfer function. Each node in these layers is linked using the weighted connection (W). Other parameters of Elman-NN are a Bias (B) providing additional adjustable parameters of the model and the transfer function (f) that calculates the output (y). The transfer function can have different forms (Log function, Tan hyperbolic function, Linear function, Radial basis function, sigmoid function, etc.).

The Elman-NN is characterized by three steps: the initialization, the learning, and the validation step. The initialization step affects the accuracy and the rapidity of the prediction since the number of the hidden layers and the number of the hidden nodes enable to increase the certainty of the model, but in the other hand, increases the

computation time. Consequently, the identification of the best structure of the network must be considered carefully. In addition to the impact of the learning step on the accuracy of the model; this step is responsible to adjust the parameter of the network (best weight and best bias). For this reason, in our work, we focus on the creation of a better structure of the network and the use an efficient learning algorithm. We propose to investigate the use of the Differential Evolution (DE) algorithm in the learning step of Elman-NN in order to increase the certainty of the model.

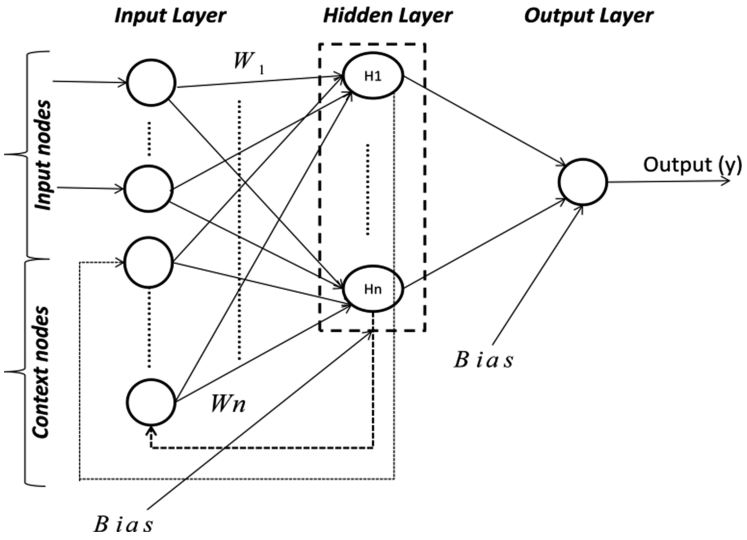


Fig. 1. Structure of Elman-Neural Network.

3.2 Differential Evolution (DE) Algorithm

The differential evolution (DE) algorithm proposed by Storn and Price [13], is a stochastic method inspired from the evolutionary algorithms. DE operates on a population of potential solutions by applying the principle of survival of the fittest to produce better approximations to a solution.

Firstly the DE algorithm produce a first/initial generation. If the optimization criterion is not satisfied, the creation of a new generation is applied. Individuals are selected according to their fitness for the production of new candidates (the offspring). Two or more candidates (called parents) are recombined to produce offspring. All offspring are transferred with a certain probability. The fitness of the offspring is then computed. The offspring are inserted into the population replacing the parents and producing a new generation. This cycle is then performed until the optimization criterion is reached. Such a single population evolutionary algorithm is powerful and performs well on a wide variety of problems. The used pseudo code of the DE algorithm is presented as follows.

Pseudo-code of DE:

Step I: generate an initial population containing *NP* individuals each of them with a dimension *D*;
Step II: for each individual, choose a target vector $y(i)$ with a dimension *D* from the Population;
Step III: randomly generate 3 vectors (r_1, r_2, r_3) from the population different to $y(i)$;
Step IV: apply the mutation by finding the noisy vector $x(i)$. Many methods can be used to calculate the noisy vector. Here, we choose the following combination that is a basic mutation method: $x(i) = r_1 + \mathrm{F}(r_2 - r_3)$;
Step V: perform the crossover between $y(i)$ and $x(i)$ in order to find the trial vector $v(i)$;

1. We choose a random number $rand_j \in [0, 1]$ which is the j^{th} value in the vector (chosen randomly among the population {1,2, ..,NP}) and a CR which is the coefficient of the crossover which belongs to [0, 1] and determined by the user or the application;
2. **If** $rand_j < \mathrm{CR}$ then $v(i) \leftarrow x(i)$ **else** $v(i) \leftarrow y(i)$;

Step VI: we calculate the costs of the trial $v(i)$ and the target vector $y(i)$;
Step VII: Once the cost is calculated, the vector that includes the lower cost replaces the population member in the initial population in case of minimization. Conversely, if the problem is maximization.

The previous code is the basic code of the DE algorithm which serves to evaluate the objective function, so in our work the objective function is to minimize the error between the input and the desired output. In the following part we discuss the use of the Elman-NN based on DE during the learning step.

3.3 Elman-Neural Network Based on DE

In this work, we improve the Elman-NN using the DE algorithm for the learning step in order to tune the parameters of Elman-NN to improve the prediction accuracy. Our resulting model combines the Elman-NN and the DE algorithm and this hybridization focuses on the improvement of the Elman-NN and to boost the certainty of the prediction. Different criteria should be optimized in Elman-NN to obtain a good result by the DE algorithm in the learning step. Such criteria concern providing a better weight/bias of the network and minimizing the error of prediction. The flow chart of Elman-NN learning step based on DE algorithm in the learning step is shown in Fig. 2.

This applied combination is responsible to optimize the weight and bias values in order to satisfy our objective function and obtain a minimum error in the prediction. The following pseudo-code (combination between the DE and Elman-NN) presents the use of the DE algorithm in the learning phase of Elman-NN.

Pseudo-code of Elman-NN based on DE:

Step I: a new population of individuals is randomly created;
Step II: we choose the target vector randomly W_{best} and B_{best} (W_{best} and B_{best} are considered as the best individuals);

Step III: randomly generate W_1, W_2 and B_1, B_2;
Step IV: create the noisy vector W_i and B_i for each individual in the population as follows: $W_i = W_{best} + F\,(W_1 - W_2)$ and $B_i = B_{best} + F\,(B_1 - B_2)$;
Step V: find the trial vector using a crossover between W_{best} and $W_{i,}$ crossover between B_{best} and B_i;
Step VI: calculates the cost of the target vector and the trial vector by Elman-NN;
Step VII: the vector that includes the lower cost replaces the population member in the initial population. Repeat the different steps until funding the best cost of prediction (with a minimum error);

Thanks to this learning step, we are able to identify the best model having the best parameters (weight and bias) of the networks which implicitly influence on our objective function that is the minimization of the error rate between the predicted and desired data. Indeed, the model obtained after the learning step provides us a minimum error of prediction. After the learning step, our model is ready to achieve the prediction used during the monitoring in a smart home environment. In our framework, we apply the prediction on the movements of monitored person with minimizing the cost of errors.

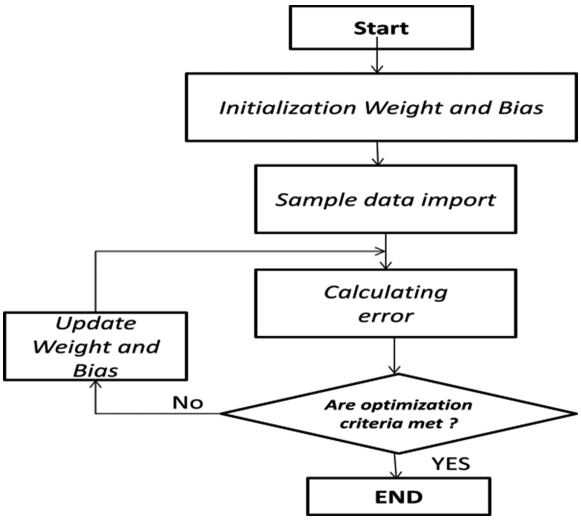


Fig. 2. The flow chart of Elman-NN based on DE algorithm

4 Data Source

In order to test efficiently our proposed model, i.e. the accuracy and the error of prediction, we propose to use a dataset collected by different sensors existing in different spaces of a smart home. The objective here is to predict the activities of daily living, the health perturbation and the abnormal activities of a monitored elderly person. In this work, the dataset used comes from a health monitoring project "e-health monitoring open data project" where data represents scenarios of an elderly person in a smart home monitored during one year.

Each input data includes three parameters: starting time – ending time – code of activities and actions. Table 1 shows the different activities listed by their code. From Table 1, we can note that for the grooming and meal preparation activities there are different codes in the same kind of activity, each code presents a detailed action.

Table 1. Activities of the monitored person classed by code

Codes	Activities
1	Eating
2	Dressing
3	Washing
{6-4-5-34}	Grooming
7	Unary-function
8	Bowel-function
9	Toileting
10	House-keeping
11	Laundry
{13-14-15-16}	Meal preparation
17	Telephone
18	Medication Use
19	Watching TV
24	Sleeping
33	Go out
35	Reading

For grooming activity *{Washing hand/face (4), Hair dry (5), Move dish (34), Make up (6)}.*

For Meal preparation activity *{Wash dish (11), Make coffee (13), Make tee (14), Make sandwich (15), Make hot food (16)}.*

It is rather difficult to directly identify and extract the movement information of the monitored person from the collected sensor data existing in the input dataset seeing that a complex and heterogeneous data. For this reason; we firstly convert the input data in the form of time series indicating the different activities of the elderly at home. A time series is a sequence of data, typically consisting of successive measurements made over a time interval. Secondly, we normalize the input data between 0 and 1 to be ready as inputs for the predictive techniques.

The value of the input data is normalized as:

$$\overline{X} = \frac{X - \min(\mathrm{X})}{\max(\mathrm{X}) - \min(\mathrm{X})} \tag{1}$$

Where X is the current values, max(X) is the largest value and min(X) is the smallest value of the sample.

5 Simulation and Results

To efficiently create a good network of the Elman-NN based on DE we choose firstly the structure of the model (input, output and hidden neurons number). We analyze the results of the Elman-NN-DE using 8 input neurons, 9 output neurons and 8 hidden neurons during 30 days. In order to evaluate the performance of the Elman-NN-DE we calculate the prediction error. For this, we evaluate the root mean square error (RMSE) given by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i)^2} \tag{2}$$

Where N represents the number of the learning observations, Y_i represents the output network, and X_i represents the input value for the i observation (with i from 1 to N). After the simulation of the Elman-NN based on DE, the observed result of the RMSE is $2.5710 * 10^{-5}$ which is a good result that shows the certainty of our model.

To confirm the performance of our proposed model in the prediction of the elderly activities in the smart home, we compare our Elman-NN based on DE algorithm model with Elman-NN based on back-propagation algorithm and Elman-NN based on Genetic algorithm. Table 2 shows that our model succeeds to provide a better result if compared with the other algorithms during the learning step. Indeed, the results show that the RMSE of the prediction using Elman-NN based on Genetic Algorithm is 0.0145 while for Elman-NN based on multi-layer back-propagation algorithm the RMSE is 2.7800. Our proposed model Elman-NN based on DE provides a RMSE of $2.5710 * 10^{-5}$ which improve significantly the prediction.

Table 2. Comparison of the RMSE between Elman-NN based on GA, Elman-NN based on multi-layer-back-propagation algorithm and Elman-NN based on DE.

Prediction models	Elman-NN based on GA	Elman-NN based on multi-layer back-propagation algorithm	Elman-NN based on DE
RMSE	0.0145	2.7800	$2.5710 * 10^{-5}$

As observed, the use of the differential evolution algorithm in the learning step decreases the root mean square error, i.e. increases the certainty of the prediction. To study the performance of our proposed model, we compared the cost of error during the learning step between Elman-NN based on GA and Elman-NN based on DE, since DE and GA belong to the same algorithm class of evolutionary/stochastic algorithms (Fig. 3).

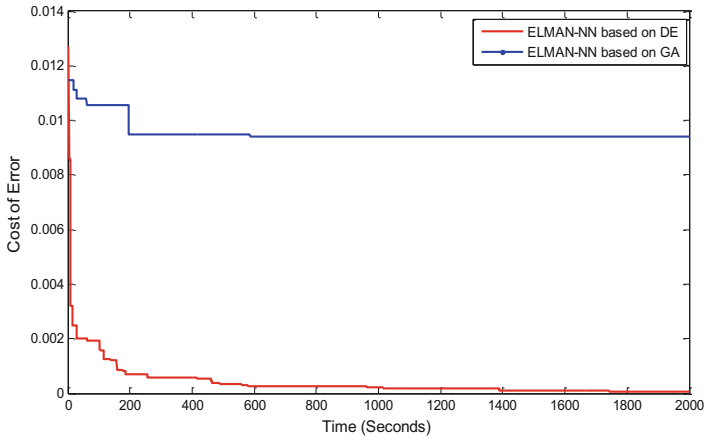


Fig. 3. Comparison of the cost of error between Elman-NN based on GA and Elman-NN based on DE.

From Fig. 3, we observe that the cost of error of the Elman-NN based on DE is very small compared to the cost of error of Elman-NN based on GA. Moreover, we can see that the error rate with DE starts to decrease rapidly against the error rate with GA which decreased slowly.

These results confirm the performance of Elman-NN based on DE in term of accuracy and rapidity during the training step. Regarding the validation step of our proposed model, we propose to test the accuracy of our prediction data separately in different times of the day: the morning, evening, afternoon and night. Figure 4 presents four curves where each one shows the real data and predicted data during different periods of the day.

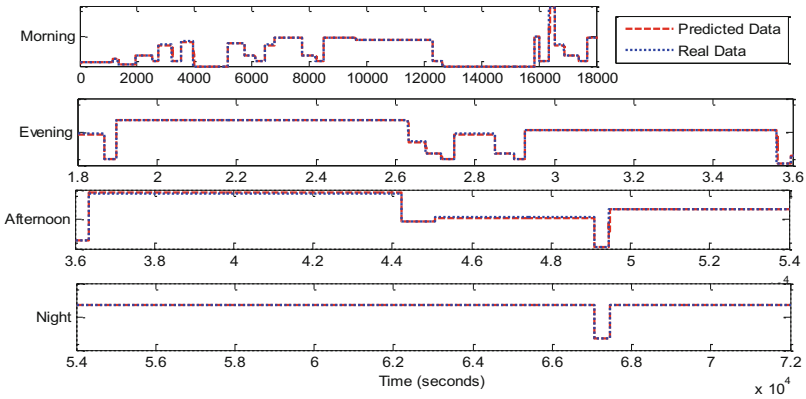


Fig. 4. The predicted and the real data in four times of the day (Morning, Evening, Afternoon, Night).

The results of the validation step show that the shape of the predicted curve is very similar to the real data curve. This result confirms the accuracy of Elman-NN based on DE at different times of the day. Consequently, our improved model succeeded to perform a continuous monitoring with a high accuracy.

6 Conclusion

In this paper, we proposed a hybrid model based on Elman Neural Network and differential evolution (DE) algorithm in order to predict efficiently and with great accuracy the activities of daily living of a monitored person.

Elman-NNs have been widely used in various areas, but the accuracy of results remained an opened issue. In order to overcome these problems, we focused this work on the optimization the Elman-NN during learning step; we investigated the use the DE algorithm in the learning step of the Elman-NN to improve the accuracy and the efficiency of the prediction activities.

The results of our simulations show that the Elman-NN based on the DE model provided a better accuracy, higher efficiency and a significant rapidity if compared to the Elman-NN based on a back-propagation algorithm or the Elman-NN based on genetic algorithm. The obtained results of using Elman-NN based on DE are promising.

Acknowledgments. The authors would like to thank all teams of the project "e-health monitoring open data project".

References

1. UNDESA: Population Division. United Nations Department of Economic and Social Affairs (2012)
2. Skouby, K.E., Kivimäki, A., Haukiputo, L., et al.: Smart cities and the ageing population. In: The 32nd Meeting of WWRF (2014)
3. Zaineb, L., Tayeb, L., Philippe, R., Fréderic, W., Hassani, M.: A Markovian-based approach for daily living activities recognition. In: International Conference on Sensor Networks (SENSORNETS) (2016)
4. Elman, J.L.: Finding structure in time. Cogn. Sci. **14**(2), 179–211 (1990)
5. Callaghan, V., et al.: Echo state network for occupancy prediction and pattern mining in intelligent environments. In: 5th International Conference on Intelligent Environments, Barcelona, p. 474. IOS Press (2009)
6. Fang, H., He, L.: BP neural network for human activity recognition in smart home. In: 2012 International Conference on Computer Science and Service System (2012)
7. Liu, Z., Song, Y., Shang, Y.: Posture recognition algorithm for the elderly based on BP neural networks. In: The 27th Chinese Control and Decision Conference (2015 CCDC), pp. 1446–1449. IEEE (2015)
8. Oniga, S., Sütő, J.: Human activity recognition using neural networks. In: 15th International Carpathian Control Conference (ICCC), pp. 403–406. IEEE (2014)
9. Oniga, S., Suto, J.: Activity recognition in adaptive assistive systems using artificial neural networks. Elektronika ir Elektrotechnika **22**(1), 68–72 (2016)

10. Xu, L., Mao, J.: Short-term wind power forecasting based on Elman neural network with particle swarm optimization. In: 2016 Chinese Control and Decision Conference (CCDC), pp. 2678–2681. IEEE (2016)
11. Chong, S., Rui, S., Jie, L.: Temperature drift modeling of MEMS gyroscope based on Genetic-Elman neural network. Mech. Syst. Signal Process. **72**, 897–905 (2016)
12. Yongchun, L.: Application of Elman neural network in short-term load forecasting. In: 2010 International Conference on Artificial Intelligence and Computational Intelligence (AICI), pp. 141–144. IEEE (2010)
13. Storn, R., Price, K.: Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. J. Glob. Optim. **11**(4), 341–359 (1997)

Towards an Approach Based on Ontology for Semantic-Temporal Modeling of Social Network Data

Rim Chiha[1,2(✉)] and Mounir Ben Ayed[1,2]

[1] Research Groups in Intelligent Machines, National School of Engineers (ENIS), University of Sfax, BP 1173, 3038 Sfax, Tunisia
[2] Computer Science and Communication Department, Faculty of Science of Sfax, Route Sokra Km 3.5, BP 1171, 3000 Sfax, Tunisia
rim.chiha@gmail.com, mounir.benayed@ieee.org

Abstract. Recently, social networks have attracted a particular interest in the web, particularly Facebook. Consequently, a large amount of social data is produced. For semantic modeling of such data, several social network ontologies are introduced. The most popular is FoaF (Friend of a Friend) ontology.

In our research, we are interested in data generated from Facebook social network. Heretofore, the semantic-temporal modeling of Facebook data is not yet studied. Thus, we propose a semantic modeling approach of social data that allows extending the FoaF ontology with Facebook data and temporal information.

Keywords: Social network · Facebook data · Semantic-temporal modeling · FoaF ontology

1 Introduction

Recently, we have witnessed the emergence of a variety of social networks such as Facebook, Twitter, Flickr, Instagram, Linkedin, and so on, through which we can connect, share and exchange diverse kind of information with one other. Indeed, the most popular social network among these is Facebook [8].

The social network data are enormous and unstructured. In the beginning, the researchers [6,15] were interested to model social data with traditional formalism such as graph theory. However, the main problem of this type of modeling is the lack of semantics between concepts. In contrast, ontological modeling of social data represents a challenge for application developers. Therefore, several semantic web standards are proposed such as FoaF (Friend of a Friend) [3], SIOC (Semantically Interlinked Online Communities) [2] SKOS (Simple Knowledge Organization Syntax) [7], etc. Nevertheless, FoaF social network ontology is considered as the most popular standard vocabulary for representing social data [8,9]. Thus, extending existing ontologies represents a new research field

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_70

in which recent approaches are proposed such as [8] that proposed to integrate in FoaF ontology Facebook user likes and friend lists. In [1], authors introduce mappings between FoaF, SIOC, SKOS and OWL-time ontologies in order to present temporal semantic description of social data.

Facebook research opens the way to several challenges such as data collection, modeling, analysis, optimization, etc. However, the data extracted from the Facebook social network suffer from a lack of semantics. Therefore, our focus in this paper is to provide a semantic model of Facebook data. Obviously, ontologies with semantic web technologies are considered as the best solutions for the semantic modeling of social data [5]. In addition, the storage of social data into ontology knowledge base makes the access and analysis of such data easy by the external systems.

Moreover, most social network data modeling approaches provide a static description of social data, the dynamic aspect is overlooked. Although, the consideration of temporal information can improve mainly the analysis process.

Based on the advantages offered by ontologies together with semantic web technologies as well as the consideration of temporal aspect, we provide a generic model that represents the Faceobok concepts while considering the semantic and temporal aspects together.

To the best of our knowledge, the extension of FoaF ontology with Facebook concepts and temporal information together is not yet examined. To fulfill this goal, we present an ontology-based model FBTO (Facebook Temporal Ontology) that represents an extension of the FoaF ontology with Facebook user activities which are visualized by his/her friends in the home page, and activities realization time.

In the remainder of this paper, we first present the different approaches supporting the semantic modeling of social data. Then, we detail our methodology for modeling FBTO. Section 4 describes the implementation and the validation of our ontology. Finally, we conclude this paper and present some future works.

2 Related Work

In the beginning, the most popular structure for representing the social networks data is graph theory [6]. In fact, the representation of social networks as graphs offers better visualization of concepts and their relationships. However, the consideration of the semantics between concepts is overlooked. For this reason, the notion of ontology appeared in order to organize the knowledge of a given domain semantically.

2.1 Existing Ontologies

In the literature, several ontologies are proposed for representing social network data.

FoaF[1] ontology [3]: FoaF is an ontology that provides a description of persons, their activities and their relationships with others that can be persons or

[1] http://www.foaf-project.org/.

objects. This ontology is formalized with RDF (Resource Description Framework)/OWL (Web Ontology Language) vocabularies that make it readable by machines. Anyone can have FoaF document that describes him or her.

SIOC[2] ontology [2]: provides a description of the interactions ensured by online communities such as blogs, forums, and mailing lists. SIOC and FoaF are considered as open linked data because of the existence of mappings between them for example SIOC: UserAccount is subclass of FoaF: OnlineAccount.

SKOS[3] ontology [7]: is a common data model for sharing and linking knowledge organization systems via the Semantic Web. Similarly, there are mappings between SIOC and SKOS vocabularies, for example the one defined by [1]: SKOS: concept SKOS: isSubjectOf SIOC: post.

2.2 Ontological Approaches for Modeling Social Data

There are many research works focus on the extension of existing ontologies such as: Shoaib and Basharat [12] are the first among those. So, they proposed ontology-based approach for ensuring the semantic profiling in social network sites. The proposed semantic model extends the existing ontologies FoaF, SIOC, Dublin Core (DC) [16] with the concepts and the relationships that are related to academia and researchers field such as relations among conferences, conference proceedings, journals, research publications and authors.

Challenger [9] proposed to enrich FoaF ontology with UniGrad ontology concepts which are related to graduate education field such as UniGrad: academician and UniGrad: university classes are considered as subclass of FoaF: organization class. Nevertheless, he used DBLP[4] ontology in order to integrate in UniGrad ontology the publications of academicians and students (DBLP: publications) that are associated to research domain (UniGrad: research area).

Kijas and Kozielski [8] aim to extend the FoaF ontology with Facebook user data that are described in Fc (FaCebook) ontology. The data concerned by this research work are user interests that can be a read book, visited place, watched movie and others likes.

Because of the dynamic structure of social network, authors in [1] were focused on this idea and consequently proposed an ontological approach for temporal semantic modeling of social network. In this approach, authors developed an ontology called SemTemp that consists in extending and aligning the ontologies: FoaF, SIOC and SKOS. So, they proposed mappings between them in order to enhance the semantic description of these ontologies. In order to support the evolution of social network, authors described mappings between FoaF and SIOC ontologies and OWL-time ontology [11].

By exploring the approaches mentioned in the literature, we found out that: (i) In these approaches, authors were interested in extending the Foaf ontology, (ii) Kijas and Kozielski [8] focus on extending FoaF with only user likes and

[2] http://sioc-project.org/.
[3] http://www.w3.org/2004/02/skos/.
[4] http://swat.cse.lehigh.edu/resources/onto/dblp.owl.

friends lists, the others activities are neglected and (iii) Only Boudebza and colleagues [1] investigate the temporal evolution of social network.

Based on these ascertainments, we found out that the extension of FoaF ontology with Facebook concepts and temporal information together is not yet examined. So, we focus in this paper on this problem.

3 Proposed Approach for Modeling FBTO

Modeling and analysis of social network data represent a new challenge. Various standards have been proposed for the semantic modeling of such data. Indeed, FoaF ontology is considered as the most popular standard.

In this paper, we describe an ontological model FBTO (Facebook Temporal Ontology) which represents an extension of Foaf with Facebook user activities and temporal information.

In the remainder of this section, we detail our approach for modeling FBTO. So, we start with defining the concepts and relations that describe the Facebook social network and that are used to extend the FoaF ontology. Then, we investigate the integration of temporal aspect on the Facebook user activities by reusing OWL-Time ontology.

3.1 FBTO Ontology Modeling

In this work, we are interested in extending the FoaF ontology with Facebook user activities that can be visualiezd by his/her friends in the home page. Therefore, we introduce the following classes:

User_FB: represents the central element of Facebook structure. A user_FB is characterized by an identifier (id_user) and name (name_user).

Account_FB: Each User_FB has a Facebook account accessible through an email account and password. Account_FB is composed of a set of Facebook pages.

Pages_FB: a Facebook page can be home page, user profile page, interesting page, event, group or application. Each Facebook page is characterized by an identifier (id_page) and name (name_page)and a description (description_page).

- The home_page: After the connection of a User_FB to his account, the home_page will be displayed. This latter represents the main page of Facebook which contains links allowing the User_FB to navigate anywhere (i.e., friends profiles, events, groups, interesting page, and so on.) and visualize the activities of his/her friends.
- The user_profile page contains activities, information, friends list and images of a user.
- Interesting_pages: pages liked by User_FB.
- Group: a page is considered as a group which is composed of groups of persons having a common interest.
- Event: a page containing information about an organized event.

Posts_FB: describe the different publications shared by a Facebook user on Facebook page. A post can be an image, a status, a video, a place, a link and an article.

– Place: a visited place mentioned by a user.
– Link: a link of external site or object posted by a user.
– Article: text shared by a Facebook user.

Each post is characterized by an identifier (id_post) and name (name_post), a description (description_post) and message (message_post). For a given post, we can distinguish different types of reactions (type_reaction). The type_reaction property can have one of the literal values: comments, likes, loves and the other emoticons (Haha, wouah, sad, grrr).

Connection_state: indicates the connection state of a Facebook user that can be on-line or off-line.

Through Table 1, we provide the relationships that connect the Facebook objects (objects properties).

Table 1. FaceBook objects properties

Subject	Predicate	Objects
Account_FB	composed_of	Pages_FB
User_FB	Has_profile	profile_page
User_FB	Has_home	Home_page
Pages_FB	Contains	Posts_FB
Image	Rdfs: subClassOf	Posts_FB
Status	Rdfs: subClassOf	Posts_FB
Video	Rdfs: subClassOf	Posts_FB
Link	Rdfs: subClassOf	Posts_FB
Article	Rdfs: subClassOf	Posts_FB
Place	Rdfs: subClassOf	Posts_FB
Home_page	Rdfs: subClassOf	Pages_FB
User_profile_page	Rdfs: subClassOf	Pages_FB
Interested_page	Rdfs: subClassOf	Pages_FB
Group	Rdfs: subClassOf	Pages_FB
Event	Rdfs: subClassOf	Pages_FB

User activities can be modeled with binary or n-ary relationships. A binary relationship connects exactly two objects, while an n-ary relationship allows to connect several objects with signal relation. In our research, we propose to use the reification process to represent the n-ary relationships.

Through Table 2, we present the binary relationships.

Table 2. The binary relationships

Subject	Predicate	Objects
User_FB	Reacts_to	Posts_Fb
User_FB	Participates	Event
User_FB	likes_page	interested_page
User_FB	Joins	Group
User_FB	Knows	User_FB
User_FB	Known_by	User_FB
User_FB	Has_connection_state	Connection_state

In Facebook social network, a user can be identified in a post shared by another user. For modeling this activity, we define the ternary relationship FBTO: Person_identification which requires two subjects i.e., two user_FB that are connected to an object Post_FB.

In some cases, a Facebook user can share posts in the profile pages of his friends or in the other pages such as group, liked page or events. This activity can be performed when the identified user or the page administrator authorizes this kind of action. To represent this activity, we introduce the relation Share_posts which connects the User_FB, Posts_FB and Pages_FB classes.

3.2 Mappings Between FBTO and FoaF Ontologies

A User_FB is considered as a sub class of FoaF: person.

Besides, an Account_FB is represented as a sub class of FoaF: OnlineAccount.

Each Facebook user has a account Facebook. Therefore, we use the relation FoaF: HoldsAccount for connecting the User_FB to his Account_FB.

In addition, the posts description should be presented in the FBTO model for analysis purposes. For this reason, we propose to use the FoaF: primaryTopic that allows relating a post to its main topic.

3.3 Mappings Between FBTO and OWL-Time Ontologies

To integrate temporal information of activities performed by a User_FB, we introduce in FBTO ontology two new classes activities_time and connection_time describing respectively the time of realization of such activity (i.e. likes page, comments, joins, etc.) and the connection time of a User_FB.

Afterwards, we provide mappings between FBTO and OWL-Time ontologies: Activities_time and conection_time are considered as sub class of OWL-time: instant.

Activities_time class describes the realization times of user activities such as:

- Reacted_at indicates the time of commenting and reacting to a Post_FB.
- Ensured_at provides the time of liking, joining and participating to Pages_FB.

– Shared_at shows the publication time of such post on a Facebook page such as user_profile, event, group.
– Identified_at introduces the identification time of a user name on a Post_FB.
– Knows_at presents the time when a user become friend with another User_FB.

Connected_at allows linking connection_state and connection_time classes in order to provide connectivity time if the User_FB is online and the last connection time if he is off-line.

The FBTO ontology is shown in Fig. 1.

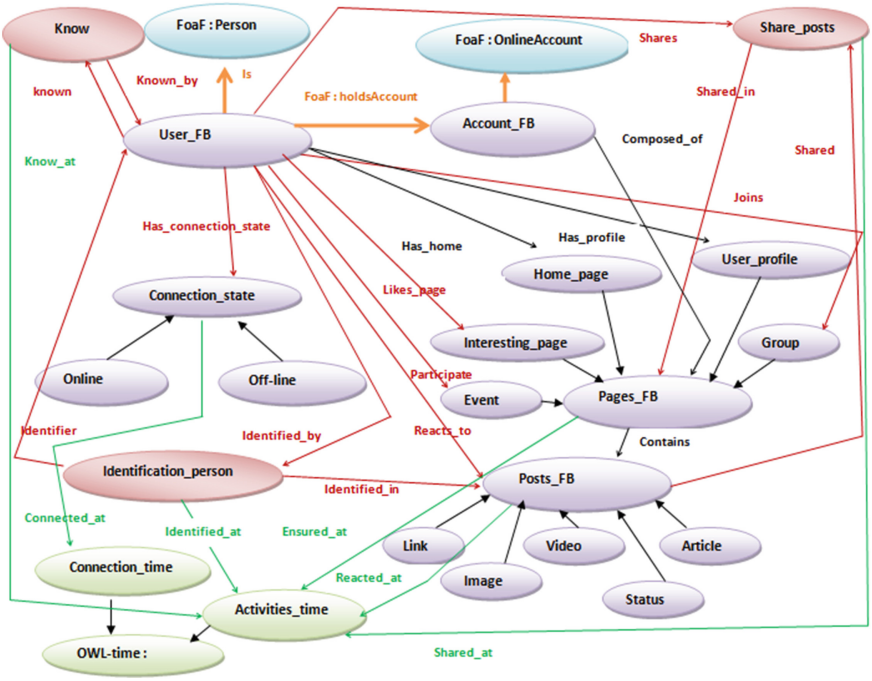


Fig. 1. FBTO model.

4 Experimentation

In this section, we describe the Facebook data extraction process as well as the loading and the validation of the FBTO ontology.

4.1 Extraction and Integration of Facebook Data into FBTO

The Facebook data collection can be ensured through two techniques: access to graph API explorer tool or HTML parsing. Today, the Facebook API becomes more restricted. Only the data associated to a logged user is accessible as well as the data of pages which have public privacy parameters. To access to data

of a random user or private page, two solutions are proposed [8]: developing Facebook application asking every user for relevant privileges or using HTML parsing technique.

To extract information about Carrefour page and their posts, we use Graph API Explorer tool (API, 2016).

To integrate the extracted data into FBTO, we use Jena API. Indeed, the loading step is ensured through mappings between sources (i.e. Facebook data) and target (i.e. the proposed knowledge base: FBTO ontology).

Table 3 shows mapping examples.

Table 3. Mappings between sources and targets

Source: Facebook data	Target: FBTO ontology
User: id	id_user
User:name	name_user
Page: id	id_page
Page: name	name_page
Page: description	description_page
Post: id	id_post
Post: name	name_post
Post: message	message_post
Post: description	description_post
Post: reactions	type_reaction
Post: created_time	shared_at

4.2 FBTO Validation

In our research, we adopt the OWL Web Ontology Language to formalize the FBTO ontology. For ensuring the syntax validation of FBTO, we use the W3C RDF validator that validates OWL document with RDF syntax. To detect logical inconsistencies, the verification is performed by reasoners such as Pellet [13], HermiT [10], Fact [14]. Therefore, we take advantage of reasoning capabilities from Hermit for ensuring the validation tests of our ontology such as the consistency tests, classification tests and query tests.

- Consistency test: ensures that ontology does not contain any contradictory facts. Each class must have at least one individual member. A class is considered inconsistent if it cannot have any instance.
- Classification test: checks whether a class is a subclass of another class or not. It computes the subclass relations between every named class to create the complete class hierarchy.
- Query test: queries are not part of our knowledge base, but are a way to identify the instances in our project, based on class properties. For querying FBTO

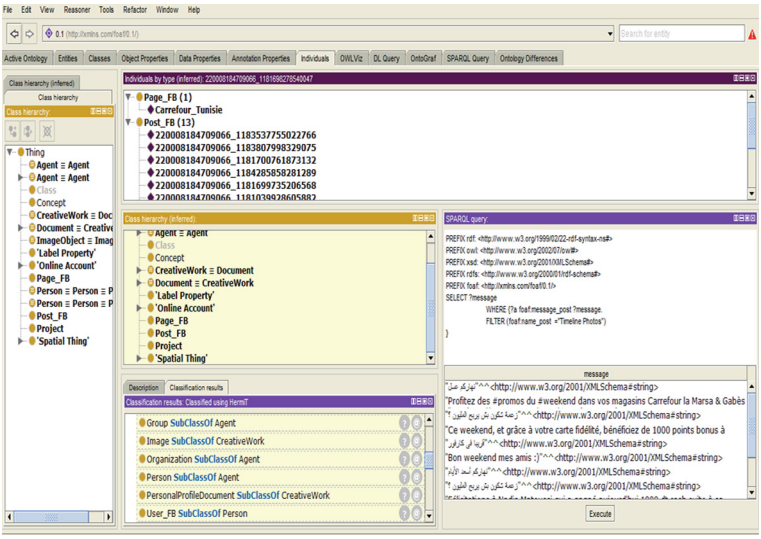


Fig. 2. FBTO validation using Hermit reasoner

knowledge base, we use SPARQL language. The SPARQL query mentioned in Fig. 2 allows to identify the messages of posts which have the name "Timeline Photos".

Figure 2 shows the FBTO validation using Hermit reasoner.

Indeed, the FBTO application on a use case will be investigated in future works in order to highlight the usefulness of this ontology and to compare between this latter and other researches describing ontological modeling of social data for further validating the proposed approach.

5 Conclusion

Over the recent years, we have noticed that the social networks have attracted the attention of millions of users. Also, they have important influences on daily life. According to [4], the social network is defined as the connection of person or group of persons by friendship, co-working or information exchange relationships. Thus, the use of social network produces a high amount of social information.

For semantic modeling of social network data, several ontologies are proposed such as FoaF, SIOC, and SKOS. In addition, different approaches are proposed in the literature for creating social ontologies. These approaches are divided into two main categories: extending existing ontologies or creating ontology from scratch. In this work, we are interested in the first category.

Our objective in this paper is to give, firstly, a survey related to the modeling of social network data. Then, we presented the proposed approach for semantic modeling of Facebook data taking into consideration the temporal aspect. The

proposed Facebook temporal ontology (FBTO) allows extending FoaF ontology with Facebook user activities and realization time of the latter.

Our model is applied on the data of Facebook social network. In future work, we will focus on the generalization of our approach to be applied on other social network data.

References

1. Boudebza, S., Azouaou, F., Nouali, O.: Ontology-based approach for temporal semantic modelling of social networks. In: 2015 3rd International Conference on Future Internet of Things and Cloud (FiCloud), pp. 736–741, August 2015
2. Breslin, J.G., Harth, A., Bojars, U., Decker, S.: Towards semantically-interlinked online communities. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 500–514. Springer, Heidelberg (2005). doi:10.1007/11431053_34
3. Brickley, D., Miller, L.: FOAF Vocabulary Specification 0.97. Namespace Document, January 2010
4. Garton, L., Haythornthwaite, C., Wellman, B.: Studying online social networks. J. Comput.-Mediated Commun. **3**(1), 75–105 (1997)
5. Golbeck, J., Rothstein, M.: Linking social networks on the web with FOAF: a semantic web case study. In: Fox, D., Gomes, C.P. (eds.) AAAI, pp. 1138–1143. AAAI Press (2008). http://dblp.uni-trier.de/db/conf/aaai/aaai2008.html#GolbeckR08
6. Harary, F., Norman, R.Z.: Graph theory as a mathematical model in social science. University of Michigan Press, Ann Arbor (1953)
7. Isaac, A., Summers, E.: SKOS simple knowledge organization system primer. World Wide Web Consortium, Working Draft WD-skos-primer-20080829, August 2008
8. Kijas, W., Kozielski, M.: Integration of Facebook online social network user profiles into a knowledgebase. In: Kozielski, S., Mrozek, D., Kasprowski, P., Małysiak-Mrozek, B., Kostrzewa, D. (eds.) BDAS 2015. CCIS, vol. 521, pp. 245–255. Springer, Heidelberg (2015). doi:10.1007/978-3-319-18422-7_22
9. Moharram, C.: The ontology and architecture for an academic social network. IJCSI Int. J. Comput. Sci. **9**(6), 22–27 (2012)
10. Motik, B., Shearer, R., Horrocks, I.: Hypertableau reasoning for description logics. J. Artif. Intell. Res. (JAIR) **36**, 165–228 (2009). doi:10.1613/jair.2811
11. Pan, F.: An Ontology of Time: Representing Complex Temporal Phenomena for the Semantic Web and Natural Language. VDM Verlag (2009)
12. Shoaib, M., Basharat, A.: Ontology based knowledge representation and semantic profiling in personalized semantic social networking framework. In: 2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), vol. 5, pp. 95–99. IEEE (2010)
13. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: a practical OWL-DL reasoner. Web Seman.: Sci. Serv. Agents World Wide Web **5**(2), 51–53 (2007). http://www.websemanticsjournal.org/index.php/ps/article/view/110
14. Tsarkov, D., Horrocks, I.: FaCT++ description logic reasoner: system description. In: Furbach, U., Shankar, N. (eds.) IJCAR 2006. LNCS (LNAI), vol. 4130, pp. 292–297. Springer, Heidelberg (2006). doi:10.1007/11814771_26
15. Wasserman, S., Faust, K.: Social network analysis: methods and applications, vol. 8. Cambridge University Press (1994)
16. Weibel, S.: The dublin core: a simple content description model for electronic resources. Bull. Am. Soc. Inf. Sci. Technol. **24**(1), 9–11 (1997)

Evaluation of the Simulated Annealing and the Discrete Artificial Bee Colony in the Weight Tardiness Problem with Taguchi Experiments Parameterization

André S. Santos[1], Ana M. Madureira[2], and Maria R. Varela[1(✉)]

[1] School of Engineering - Department of Production and Systems, University of Minho, Guimarães, Portugal
abg@isep.ipp.pt, leonilde@dps.uminho.pt
[2] School of Engineering - GECAD Research Group, Polytechnic Institute of Porto, Porto, Portugal
amd@isep.ipp.pt

Abstract. Meta-Heuristics (MH) are the most used optimization techniques to approach Complex Combinatorial Problems (COPs). Their ability to move beyond the local optimums make them an especially attractive choice to solve complex computational problems, such as most scheduling problems. However, the knowledge of what Meta-Heuristics perform better in certain problems is based on experiments. Classic MH, as the Simulated Annealing (SA) has been deeply studied, but newer MH, as the Discrete Artificial Bee Colony (DABC) still need to be examined in more detail. In this paper DABC has been compared with SA in 30 academic benchmark instances of the weighted tardiness problem ($1||\Sigma w_j T_j$). Both MH parameters were fine-tuned with Taguchi Experiments. In the computational study DABC performed better and the subsequent statistical study demonstrated that DABC is more prone to find near-optimum solutions. On the other hand SA appeared to be more efficient.

Keywords: Complex Combinatorial Problems · Simulated Annealing · Discrete Artificial Bee Colony · Taguchi Experiments · Weighted tardiness problem

1 Introduction

Complex Combinatorial Problems (COPs) due to of their exceptional complex nature have been the focus of much research, in search of techniques that could address them efficiently and effectively. Operational Scheduling Problems (OSPs) given the number of constrains and the number of variables makes the problem extremely hard to approach with exact methods. Even if some OSPs can be addressed by simple heuristic rules, MHs have become the most used technique and have achieved the best results in the resolution of these problems [1, 2].

Even with the development and study of several Meta-Heuristics, including the Simulated Annealing (SA) and the Discrete Artificial Bee Colony (DABC), there is still uncertainty on what makes certain MH more efficient, without a computational

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_71

comparative analysis, since it is impossible to determine a priori, which parameters will be more efficient or even which MH will preform better.

This paper is structured as follows: Sect. 2 overviews Operations Scheduling Problems (OSPs) and presents in detail the problem used in the computational study. Section 3 overviews Meta-Heuristics, and presents the inspiration and algorithms for the two MH implemented in the computational study: Simulated Annealing (SA) and Discrete Artificial Bee Colony (DABC). Section 4 presents the parameterization procedure for both meta-heuristics in detailed, which is conducted with Taguchi Experiments whose results are showed in this section. In Sect. 5 the computational results and the statistical study are described. Section 6 summarizes the paper results and presents the conclusions.

2 Operational Scheduling Problems

Operational Scheduling Problems (OSPs) are decision problems that distribute the operations over the available resources and disperse the operations in time, or sequence them, in each resource [3]. In other words, it is the process that determines which resources will be used in each operation and in which sequence the operations should be processed.

2.1 Single Machine Weighted Tardiness Problem

In the computational study the single machine weighted tardiness problem ($1||\Sigma w_j T_j$) will be used. It is a classical benchmark problem that sequences operations in order to minimize the tardiness, weighted by importance criterion. Given the pure sequence nature of the problem, most models of the problem use binary variables that can indicate the position of the operation in the production sequence, where $x_{jk} = 1$ indicates that operation j occupies position k on the sequence, the relative position of the operation, where $x_{jk} = 1$ indicates that operation j should be executed after operation I or the moment in which the operation should start to be executed, where $x_{jt} = 1$ indicates that operation j should start to be executed in t. One model can be examined bellow [4, 5]:

OF:

$$Min \sum_{k=1}^{n} \sum_{j=1}^{n} w_j t_k \tag{1}$$

ST:

$$\sum_{j=1}^{n} x_{jk} = 1, \quad \forall k \tag{2}$$

$$\sum_{k=1}^{n} x_{jk} = 1, \quad \forall i \tag{3}$$

$$\sum_{j=1}^{n} p_j \sum_{u=1}^{k} x_{ju} - \sum_{j=1}^{n} d_j x_{jk} \leq t_k, \quad \forall k \tag{4}$$

If t_k is the tardiness of the operation in the k position of the sequence, w_j the weight of operation j and d_j the due date of operation j, (1) is the objective function, which is the minimization of the weighted tardiness. Equation (2) defines that each operation can only be assigned to one position in the sequence and (2) that each position can only have one operation assigned to it. Equation (4) defines the tardiness of the operation in position k. It is also important to mention that the problem deals with tardiness and, as such, the t_k can be defined as $max(L_k,0)$ or tardiness.

3 Meta-Heuristics

Meta-Heuristics can be described as iterative procedures that conduct a heuristic in the exploration of the solution space [2]. Like all heuristic techniques, meta-heuristics cannot assure the optimum, but they usually produce suitable solutions within the available computational time and outperform other approximation methods.

3.1 Simulated Annealing

Simulated Annealing (SA) is one of the best-known MH, proposed by Kirkpatrick et al. [6] and Cerny [7], and inspired by the thermic treatment of metals, which are heated and cooled down repeatedly. It modifies the basic neighborhood search procedure with the introduction of a mechanism that allows SA to overcome local-optimums, which controls the likelihood of solutions, worse than the current one, being accepted. SA selects one solution per-iteration, between the solutions in the vicinity of the current solution, but, unlike neighbourhood search techniques, it also accepts solutions that are worse than the current solution. Worse solutions are accepted as the current solution with a predetermined probability that decreases with the evolution of the search procedure. At the start almost all solutions are accepted as the new solution, but with the advance of the search procedure SA becomes more selective. The most common SA algorithm can be examined in [8].

3.2 Discrete Artificial Bee Colony

Discrete Artificial Bee Colony (DABC) is one of the most recent Meta-Heuristics based on the Artificial Bee Colony (ABC), which was proposed by Karaboga [9] and Pham et al. [10], adapted for discrete problems [11]. It works through the interaction of three phases, until the interruption criterion is met: the phase of the worker bees, the phase of opportunistic bees and the phase of the scout bees. Initially it is determined a food source (x_i) for each working bee. In the phase of working bees, each one will explore a solution in the neighbourhood (v_i) of its food source. In the opportunistic bees phases, the bee's wait for the performance of each food source to determine which food source they will explore. Once a food source is abandoned, the working bee will become a scout bee and

starts to search for a new food source. The details about the DABC algorithm can be examined in [11].

4 Computational Study

SA and DABC were studied in 30 academic benchmark instances of the weighted tardiness problem ($1||\Sigma w_j T_j$) [1]. $1||\Sigma w_j T_j$ corresponds to the minimization of the weighted tardiness for n task. All the instances of the problem are available in the *ORLibrary* website [12] that also provides the optimum for each instance.

All computational trials were performed on a MacBook Pro with a 3 GHz Intel® Core i7 processor, 16 GB of RAM and Windows 10 64-bit. SA would conclude after 10,000 iterations while DABC would conclude after it explored the same number of solutions. All the parameters were define by analogy [13] and tuned with Taguchi Experiments [14] in the first instance of the problem used in the computational study not to over-fit the MH for each instance. Taguchi approach to Design of Experiments reduces the full factorial experiment with Orthogonal Arrays. For example, the full factorial experiment with 4 parameters and 3 levels would require 81 experiments, but with a L9 Orthogonal Array it requires only 9 experiments. Unlike full factorial experiments that seek the parameters that improve the mean outcome, Taguchi Experiments search for the parameters that improve the mean outcome and reduce the variance. Each parameter is evaluated in its Sign to Noise Ratio (S/N), which can be seen in [14]. One example of Taguchi Experiments in MH tuning can be seen in [15].

4.1 Simulated Annealing Parameterization

SA does not need an exhaustive parameterization procedure, since with enough time the MH performs well even with carefully determined parameters [16, 17]. However, since the computational study is limited to 10,000 iterations the parameters need to be carefully set to maximize its performance. SA parameters are the Initial Temperature (T_i), the Cooling Ratio (CR), the Epoch Length (EL), the Neighborhood Structure (NS) and the Stoppage Condition (SC), which is set to 10,000 iterations. All the other parameters will be set into three levels and fine-tuned with Taguchi Experiments.

Kirkpatrick et al. [6] states that the Initial Temperature (T_i) should accept all new solutions. Other authors refer Initial Temperature in P_i between [0.7, 0.8] [18]. Several metrics to determine the Initial Temperature are presented in [19, 20]. In order to determine this parameter, 20 random new solutions from the initial solutions with a P_i of 0.70, 0.75 and 0.80 which resulted in the rounded initial temperatures of 750, 1,000 and 1,250. Rose et al. [20] state that the Geometric Cooling Ratio should result in a slow decrease of temperature, with α between [0.80, 0.99]. One metric to calculate the α is presented in [16]. Since the smallest Initial Temperature considered was of 750 and that the procedure will stop after 10,000 iterations the Cooling Ratio should be 0.97, 0,98 and 0,99. In terms of the Epoch Length, it should be identical to the size of the problem, in this case 50, as such, the Epoch Length was set in 50, 75 or 100. In this paper three simple NS were selected: **Insert** which consists of the solutions that can be obtained by

removing one element of the solutions and inserting it into another position, **Swap** which consists of the solutions that can be obtained by interchanging two elements of the solution and **Transpose**, which consists of the solutions that are obtained by interchanging two adjacent elements of the solution. In [21, 22] there are studies on the impact of the NS in SA.

The S/N for each parameter level can be examined in Fig. 1, which resulted in an Initial Temperature of 1,250, a Cooling Ratio (α) of 0.97, an Epoch Length of 100 and Swap as the Neighborhood Structure.

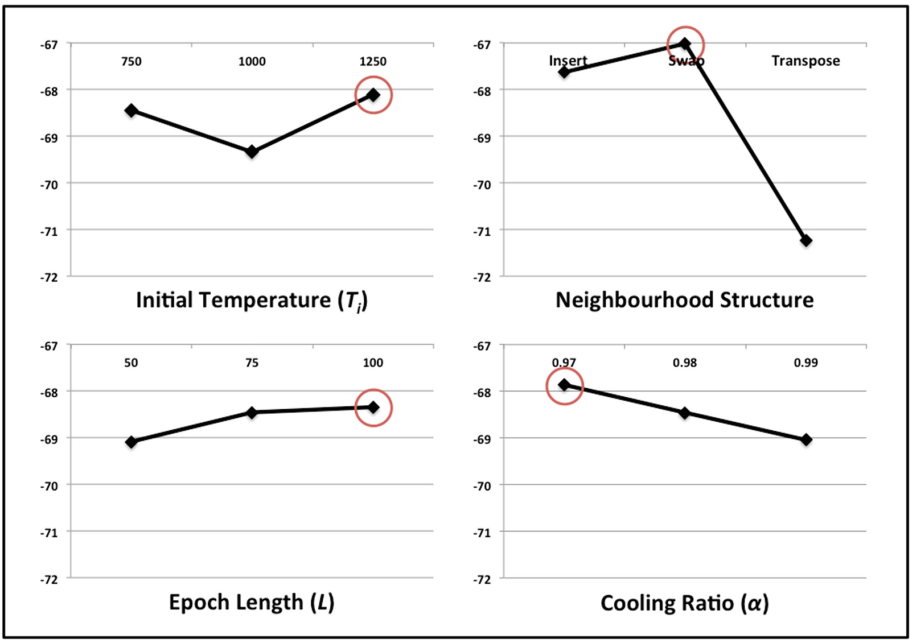


Fig. 1. Simulated annealing parameters

4.2 Discrete Artificial Bee Colony Parameterization

ABC and DABC are MHs with less parameters [23], but unlike SA, which does not require an exhaustive parameterization procedure; the DABC Meta-Heuristic requires a meticulous parameterization to perform well. In [22] it is showed how the SA scales with the available time, without a rectification of the parameters, while DABC does not. DABC parameters are the Colony Size (L), the Limit (l), the Neighborhood Structure (NS) and the Stoppage Condition. Since SA was limited to 10,000 iterations, the number of iterations of the DABC would be limited to either 10,000/10, 10,000/20 or 10,000/3 iterations, in order to ensure that both MH explored a similar number of solutions. DABC will explore, at least, one solution per-bee per-iteration and, as such, depends as much on the stoppage condition as it does on the colony size. It is important to notice that DABC will explore extra solutions since whenever a food source is abandoned another solution will be explore in the same

iteration, however the number of solutions explored by scout bees cannot be estimated. One extra parameter was considered, which was the proportion of worker/scout and opportunistic bees, which is usually fixed at 50%, however the increasing the number of worker/scout bees should increases the diversity of the MH and the increase of the opportunistic bees will increase the intensity of the MH.

Kockanat and Karagoba [24] studied the influence of the Colony Size (L) and concluded that the performance of the MH is not overly dependent on the Colony Size, which can be balanced with Stoppage Condition. Other authors refer that the L should be related to the size of the problem [25], but also concluded that it did not require detail tuning since its impact on the performance is limited. In the Taguchi Experiments, L of 10, 20 and 30 will be tested with the number of iterations previously mentioned to ensure that both MH explored the same number of solutions. For the Limit (l) number, which determines the number of iterations without improvement before a food source is abandoned, numerous authors [26–28] concluded it required to be meticulously established for larger Liu and Liu in [27] present a method to calculate the l, based on the L and the dimension of the problem. In the Taguchi Experiments the l tested will be 500, 550 and 600, determined with the Liu and Liu method [27] in a problem with 50 tasks and L of 10, 20 or 30. In the case of the Bees Proportional, which as previously mentioned is not a usual parameter, but can be similar to the set of locked scout-bees. Usually this is fixed at 50%/50%, however 40%/60%, 60%/40% and the usual 50%/50% will also be tested. Finally the Neighborhood Structure (NS) is almost a ubiquitous parameter, which can

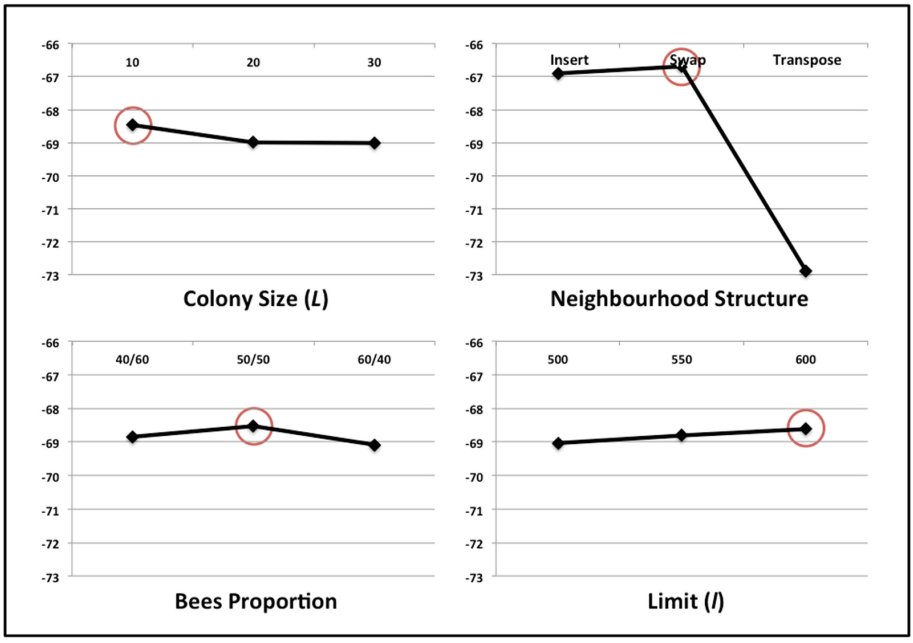


Fig. 2. Discrete Artificial Bee Colony parameters

have an impact on the performance of all MH. For DABC the same three Neighborhood NS used in SA, which are **Insert**, **Swap** and **Transpose**.

The S/N for each parameter level can be examined in Fig. 2, which resulted in a Colony Size of 10 and consequentially 10,000/10 iterations, a Limit of 600, a Bees Proportion of 100 and Swap as the NS.

5 Computational Results

SA and DABC were run 3 times for each of the 30 iterations of the problem with the solution from the best test-run being used in the statistical analyses. Overall both MH performed well, but it appears that the DABC performed better than SA, as can be seen in Fig. 3, which shows the best solution. SA found the optimum solution in 5 instances, 16.(6)%, while DABC found the optimum in 11 instances of the 30 instances, 36.(6)%. If DABC performed better than SA, it also took more computational time in almost all instances. SA took less than 0.001 s in 24 of the 30 instances and up to 0.031 s in the other 6 instances. DABC took more computational time, less than 0.001 s in 7 of the instances and up to 0.032 s in the other.

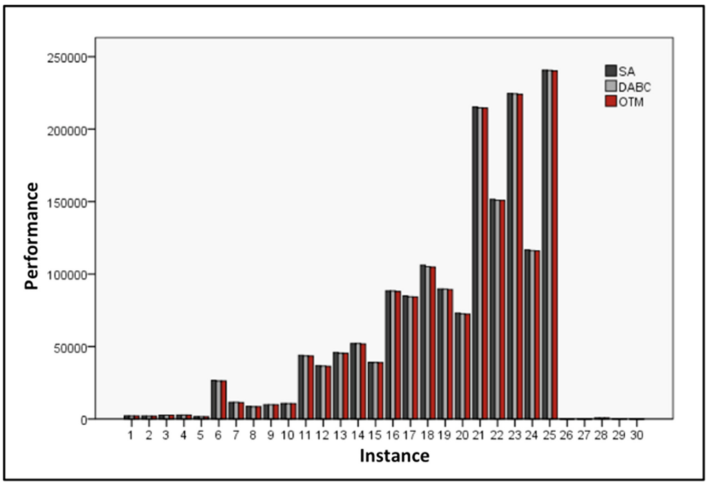


Fig. 3. SA and DABC computational results

SA achieved better results, in the computational study, than DABC, however to better evaluate this performance variance it is necessary to statistically examine the obtained results. Both MH solutions were normalized around the best-known solution for each instance, available in the ORLibrary, and measured using the mean percent deviation from those solutions, which Silberholz and Golden [28] state is the best metric to compare MH across several instances of one problem. It is preferable to use the mean percent deviation from the optimum, or in this case the best-known solution, over the absolute difference, since it allows MH to be compared across all the different instances.

For example, a 10 units difference, in whatever metric is used, cannot represent the same in an instance with an optimum solution of 10 and an optimum of 1000.

Once both MH solutions were normalized around the best-known solution, the superior performance of the DABC becomes evident. DABC obtained 28 near-optimal solutions, or solutions with deviation from the best-known solution of less than 1%, while SA obtained 16, which means that DABC found near-optimums in almost all instances, 93.(3)%, while SA found near-optimums solutions in half of the instances, 53.(3)%. SA also found 5 solutions with substantial difference from the optimum, or solutions with deviation from the best-known solution of over 5%. DABC achieved a mean deviation from the best-know solution of 0.22% while SA mean deviation was of 1.54%. DABC performance also appeared constant, with standard deviation of 0.30 and 1.91 for SA. In Fig. 4 the boxplot of the mean deviation from best-known solution, of SA and DABC can be examined.

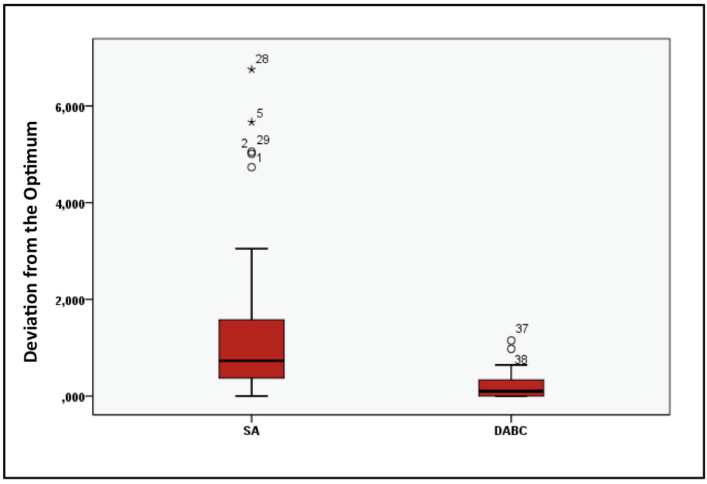


Fig. 4. Boxplot of the deviation from the best-known solution

In the computational study DABC appeared to have achieved the best results, but to determine if the meta-heuristics have the same probability of achieving near optimum solutions (>1% from the best-known solution) a statistical inference test needs to be used, in this case the Chi-Square to determine if the probability to find near-optimum solutions is independent of the MH. In Table 1 the results can be examined.

Table 1. Chi-Square results

	Value	df	Asym. sig. (2-sided)
Pearson Chi-Square	12.273	1	0.000
Continuity correction	10.313	1	0.001
Likelihood ratio	13.439	1	0.000
N. of valid cases	60		

Results from the Chi-Square test demonstrate that the likelihood of near optimum solutions is not identical for both meta-heuristics (*P-Value* 0.001). It is also possible to conclude that DABC is more likely to find a solution closer to the optimum than SA (p-value 0.001/2), which reinforces what was apparent in the computational study.

6 Conclusions

In this paper the performance of SA and DABC were studied in the weighted tardiness problem ($1||\Sigma w_j T_j$). SA is a well known meta-heuristic, but DABC, which is a discreet version of ABC, is more recent and its behavior has not been examined with the required detail. SA and DABC were studied in 30 instances available in the *ORLibrary*. Both MH parameters were tuned with Taguchi Experiments, which simplified the often-extensive parameterization procedure. SA search procedure concluded after 10,000 iterations while DABC after 10,000/10 iterations.

In the computational study DABC appeared to perform better than SA in most instances. SA found the optimum solution in 5 instances, 16.(6)%, while DABC found the optimum in 11 instances of the 30 instances, 36.(6)%. Once the MH solutions were normalized, DABC found near-optimums in 93.(3)% of the instances, while SA found near-optimums solutions in around half of the instances, 53.(3)%. In order to determine if both have the same probability of achieving near-optimums the Chi-Square was used. It demonstrated that DABC is more likely to find near-optimum solutions than SA. On the other hand, DABC took more time than SA to explore the same number of solutions. SA took less than 0.001 s in 24 of the 30 instances and up to 0.031 s in the other 6 instances. DABC took more computational time, less than 0.001 s in 7 of the instances and up to 0.032 s in the other.

Acknowledgments. This work is supported by FEDER Funds through the "Programa Operacional Factores de Competitividade - COMPETE" program and by National Funds through FCT "Fundação para a Ciência e a Tecnologia" under the project: PEst-OE/EEI/UI0760/2014, and PEst2015-2020.

References

1. Xhafa, F., Abraham, A.: Metaheuristics for Scheduling in Industrial and Manufacturing Applications. Studies in Computational Intelligence, vol. 128. Springer, New York (2008)
2. Osman, H.I., Kelly, J.P.: Meta-heuristics: an overview. In: Osman, H.I., Kelly, J.P. (eds.) Meta-Heuristics Theory and Applications, pp. 1–21. Springer, Heidelberg (1996)
3. Pinedo, M.L.: Scheduling Theory, Algorithms, and Systems, 4th edn. Springer, New York (2012)
4. Keha, A.B., Khowala, K., Fowler, J.W.: Mixed integer programming formulations for single machine scheduling problems. Comput. Ind. Eng. **56**, 357–367 (2009)
5. Khowala, K., Keha, A., Fowler, J.: A comparison of different formulations for the non-preemptive tardiness scheduling problem. In: Proceedings of the International Conference on Scheduling: Theory and Applications, pp. 643–651 (2005)

6. Kirkpatrick, S., Gelatt, C.D., Vecchi, P.M.: Optimization by simulated annealing. Science **220**, 671–680 (1983)
7. Cerny, V.: A thermodynamical approach to the travelling salesman problem: an efficient simulated annealing algorithm. J. Optim. Theory Appl. **45**, 41–51 (1985)
8. Boussaid, I., Lepagnot, J., Siarry, P.: A survey on optimization metaheuristics. Int. J. Inf. Sci. **237**, 82–117 (2013)
9. Karaboga, D.: An idea based on honey bee swarm for numerical optimization. Technical report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department (2005)
10. Pham, D.T., Ghanbarzadeh, A., Koc, E., Otri, S., Rahim, S., Zaidi, M.M.: The Bees Algorithm. Technical note. Cardiff University (2005)
11. Karaboga, D., Gorkemli, B.: A combinatorial artificial bee colony algorithm for traveling salesman problem. In: Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications (INISTA), pp. 50–53 (2011)
12. Beasley, J.E.: ORLibrary (1990). http://www.brunel.ac.uk/~mastjjb/jeb/info.html
13. Montero, E., Riff, M., Neveu, B.: A beginner's guide to tuning methods. Appl. Soft Comput. **17**, 39–51 (2014)
14. Taguchi, G.: Introduction to Quality Engineering: Designing Quality into Products and Processes. Quality Resources (1986)
15. Nadir, B., Zandieh, M., Fatemi Ghomi, S.M.T.: Scheduling job shop problems with sequence-dependent setup time. Int. J. Prod. Res. **47**(21), 5959–5976 (2009)
16. Park, M.W., Kim, Y.D.: A systematic procedure for setting parameters in simulated annealing algorithm. Comput. Oper. Res. **25**(3), 207–217 (1998)
17. Anily, S., Federgruen, A.: Simulated annealing methods with general acceptance probabilities. J. Appl. Probab. **24**, 657–667 (1987)
18. Talbi, E.G.: Meta-Heuristics: From Design to Implementation. Wiley, Hoboken (2009)
19. Eglese, R.W.: Simulated annealing: a tool for operational research. Eur. J. Oper. Res. **46**, 271–281 (1990)
20. Rose, J., Klebsch, W., Wolf, J.: Temperature measurement and equilibrium dynamics of simulated annealing placement. IEEE Trans. Comput. Aided Des. **9**, 253–259 (1990)
21. Chek, K.M., Goldberg, J.B., Askin, G.: A note on the effect neighborhood structure in simulated annealing. Comput. Oper. Res. **18**, 537–547 (1991)
22. Santos, A.S., Madureira, A.M., Varela, M.L.: Study on the impact of the NS in the performance of meta-heuristics in the TSP. In: IEEE International Conference on Systems, Man, and Cybernetics, pp. 1110–1115 (2016)
23. Yan, G., Li, C.: An effective refinement artificial bee colony optimization algorithm based on chaotic search and application for PID control tuning. J. Comput. Inf. Syst. **7**(9), 3309–3316 (2011)
24. Kockanat, S., Karaboga, N.: Parameter tuning of artificial bee colony algorithm for Gaussian noise elimination on digital images. In: Proceedings of the International Symposium on Innovation in Intelligent Systems and Applications (INISTA), pp. 1–4 (2013)
25. Akay, B., Karaboga, D.: Parameter tuning for the artificial bee colony algorithm. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS (LNAI), vol. 5796, pp. 608–619. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04441-0_53
26. Kiran, M.S., Gunduz, M.: The analysis of peculiar control parameters of artificial bee colony algorithm on the numerical optimization problems. J. Comput. Commun. **2**, 127–136 (2014)
27. Liu, Y.F., Liu, S.Y.: A hybrid Discrete Artificial Bee Colony algorithm for permutation flowshop scheduling problem. Appl. Soft Comput. **13**, 1459–1463 (2013)
28. Silberholz, J., Golden, B.: Comparison of metaheuristics. In: Gendreau, M., Potvin, J. (eds.) Handbook of Metaheuristics, vol. 146, pp. 625–640. Springer, Heidelberg (2010)

Industrial Plant Layout Analyzing Based on SNA

M.L.R. Varela[1(✉)], V.K. Manupati[2], K. Manoj[2], G.D. Putnik[1], A. Araújo[1], and A.M. Madureira[3]

[1] Department of Production and Systems, School of Engineering, University of Minho, Guimarães, Portugal
{leonilde, putnikgd}@dps.uminho.pt, dricafaraujo@hotmail.com
[2] Division of Manufacturing, School of Mechanical Engineering, VIT University, Vellore, Tamil Nadu, India
manupativijay@gmail.com, kakarla.manoj2013@vit.ac.in
[3] Department of Computer Science, Polytechnic Institute of Porto, GECAD Research Group, Porto, Portugal
amd@isep.ipp.pt

Abstract. Social network analysis (SNA) is a widely studied research topics, which has been increasingly being applied for solving different kind of problems, including industrial manufacturing ones. This paper focuses on the application of SNA on an industrial plant layout problem. The study aims at analyzing the importance of using SNA techniques to analyze important relations between entities in a manufacturing environment, such as jobs and resources in the context of industrial plant layout analysis. The study carried out enabled to obtain relevant results for the identification of relations among these entities for supporting to establish an appropriate plant layout for producing the jobs.

Keywords: Manufacturing systems · Plant layout · Social network analysis · Case study

1 Introduction

Social network analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. The nodes in the network are the people and groups while the links show relationships or flows between the nodes [1].

The growth of interest in the techniques of social network analysis has been considerable since the 1970s and has been especially marked in the last decades [1–4]. Recent growth has been sparked partly by the increasing emphasis on the importance of "networking" in practical management guides and partly by the proliferation of "social networking" websites such as Facebook and Twitter. This has encouraged many to explore the advantages of using social network analysis [1].

In this paper is intended to analyze data through SNA, about a study for analysing the relationship between a set of jobs that incorporate a set of tasks, which have to be

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_72

performed on a set of resources, with the aim of preparing a plant layout, by checking the influence that may exist between these entities through typical measures of SNA techniques, such as centrality, closeness and betweeness, among other descriptive statistical analysis associated that might be of relevance to be able to propose an appropriate layout for arranging the resources, according to a given production process for the jobs.

This paper is organized as follows: besides this Sect. 1 about this introduction, a brief overview about Social Network Analysis (SNA) is presented next, in Sect. 2. Section 3 presents a brief literature review about some more or less closely related work about the application of SNA techniques. Section 4 describes the case study carried out in this work for supporting the establishment of an industrial plant layout, along with the social network analysis and some important measures of network structure. Finally, in Sect. 6 are presented some conclusions and directions for future work, which are also discussed.

2 Social Network Analysis

Research on social networks has grown significantly over the last few years [5]. A social network consists of a finite set of actors and the ties between them [6]. The three basic elements in social networks are actors, ties and graphs [7]. Actors are network members that can be distinct individuals or collective units. Ties, which can be formal or informal, link actors within a network. Graphs are visual representations of networks, displaying the actors as nodes and the ties as lines [8].

Social network analysis (SNA) is the study of a social structure [9]. SNA describes a group of quantitative methods for analyzing the ties among social entities and their implications [10]. An important aspect in social network analysis is to identify key players in a network [11].

2.1 SNA Measures

The main measures considered in SNA are cohesion measures and centrality measures. Cohesion measures describe the interconnectedness of actors in a network [7]. The main measure of cohesion is the density of the network, which corresponds to the total number of ties divided by the total possible number of ties.

To calculate the density of the network Eq. 1 was used.

$$D = T/PT \tag{1}$$

Where, D, T, and PT refer to Density, Ties, and Possible Ties, respectively.

Centrality measures identify the most prominent actors, i.e. those extensively involved in relationships with other network members [11]. The most commonly used centrality measures are: degree, betweenness and closeness [12]. The centrality measures of degree, betweenness and closeness were calculated in this work through the UCINET tool.

Degree centrality is the number of actors with whom a particular actor is directly related. Betweenness centrality and closeness centrality are related to the distance (neighbourhood) between actors in a network. Betweenness centrality is the number of times an actor connects pairs of other actors [7]. Closeness centrality presents distances between actors and describes how closely actors are connected to the entire network population [13].

3 Literature Review

According to the authors in [14], Social Network Analysis (SNA) provides tools to examine relationships between people. Text Mining (TM) allows capturing the text they produce in Web 2.0 applications, for example, however it neglects their social structure.

In their paper [14] the authors apply an approach to combine the two methods named "content-based SNA". Using the R mailing lists, R-help and R-devel, they show how this combination can be used to describe people's interests and to find out if authors who have similar interests actually communicate. As stated by the authors, they found that the expected positive relationship between sharing interests and communicating gets stronger as the centrality scores of authors in the communication networks increases [14]. Moreover, they refer that the paper shows how content-based SNA can be used to find people's interests in mailing list networks.

Additionally, by comparing communication graphs and networks showing who has similar interests, a relationship between the correlation of these two and node centrality could be found. Accordingly, the authors conclude that the expected relationship between sharing interests and communicating exists only for very active authors while less active authors do not answer everyone who has similar interests. Thus, they refer that the communication efficiency can be regarded to be high for very active mailing list authors while it is moderate for middle-active authors. The paper also suggests using only the subjects to find the relationship between communicating and sharing interests because the content contains more noise [14].

Another interesting contribution is presented in [15], were a case study examines infrastructure planning in the Swiss water sector. According the authors, water supply and wastewater infrastructures are planned far into the future, usually on the basis of projections of past boundary conditions, but they affect many actors, including the population, and are expensive. Therefore, their objective consisted on investigating fragmentation in water infrastructure planning, to understand how actors from different decision levels and sectors are represented, and which interests they follow [15]. The network analysis they did obtain confirmed their hypothesis of strong fragmentation, as they sated that they found little collaboration between the water supply and wastewater sector (confirming horizontal fragmentation), and few ties between local, cantonal, and national actors (confirming vertical fragmentation). Moreover, according the authors infrastructure planning was clearly dominated by engineers and local authorities, and little importance was given to longer-term strategic objectives and integrated catchment planning, which was perceived as more important in a second analysis carried out by the authors, that went beyond typical questions of stakeholder analysis. In their study,

the authors concluded that linking a stakeholder analysis, comprising rarely asked questions, with a rigorous social network analysis is very fruitful and did enable to generate complementary results. Moreover, this combination gave them deeper insights into the socio-political-engineering world of water infrastructure planning, which according their opinion is of vital importance to the general welfare [15].

As stated in [16] coordination increasingly occurs through networks of informal relations rather than channels tightly prescribed by formal reporting structures or detailed work processes. However, while organizations are moving to network forms through joint ventures, alliances, and other collaborative relationships, executives generally pay little attention to assessing and supporting informal networks within their own organizations. Moreover, the authors refer that social network analysis is a valuable means of facilitating collaboration in strategically important groups such as top leadership networks, strategic business units, new product development teams, communities of practice, joint ventures, and mergers. Moreover, according to them, by making informal networks visible, social network analysis helps managers systematically to assess and support strategically important collaboration [16].

Another interesting example of application of SNA is provided in [17], were is presented a mixed evaluation method that combines traditional sources of data with computer logs, and integrates quantitative statistics, qualitative data analysis and social network analysis in an overall interpretative approach. The authors propose the use of several computer tools to assist in this process, integrated with generic software for qualitative analysis. The authors applied their evaluation method and tools incrementally and validated them in the context of an educational and research project that had been going on during three years. The use of their proposed method was illustrated on their paper through an example consisting on the evaluation of a particular category within their project. Moreover, their proposed method and tools aimed at providing an answer to the need for innovative techniques for studying new forms of interaction emerging in Computer-Supported Collaborative Learning (CSCL); for increasing the efficiency of the traditionally demanding qualitative methods. The authors concluded that their methods can be used by teachers in curriculum-based experiences; and at the definition of a set of guidelines for bridging different data sources and analysis perspectives [17].

In paper [18], the authors provide another kind of application of SNA, for supply chain researchers with an overview of social network analysis, covering both specific concepts (such as structural holes or betweenness centrality) and the generic explanatory mechanisms that network theorists often invoke to relate network variables to outcomes of interest. As stated by the authors, one reason for discussing mechanisms is to facilitate appropriate translation and context-specific modification of concepts rather than blind copying. Therefore, they have also taken care to apply network concepts to both "hard" types of ties (e.g., materials and money flows) and "soft" types of ties (e.g., friendships and sharing of information), as according to them, both are crucial (and mutually embedded) in the supply chain context. Moreover, the authors did also aim to point to areas in other fields that they thought would be particularly suitable for supply chain management (SCM) to draw network concepts from, such as sociology, ecology, input–output research and even the study of romantic networks. According to their statement, they believe that the portability of many network

concepts is high to provide a potential for unifying many fields, and a consequence of this for SCM it may enable to decrease the distance between SCM and other branches of the management science.

As we can realise through the overview presented in this section about applications of SNA, this kind of approaches and underlying techniques can be applied to many different domains, in general, and also to some more specific areas, for instance, in the context of industrial management, and in this paper the aim is its application for establishing an industrial plant layout. In this paper we are working with a computational complex problem, which therefore requires heuristic solutions as the one proposed in this paper, based on SNA analysis.

Moreover, due to the computational complexity of this kind of problems there was also a need to restrict the dimension of our case study entry data, as will be briefly described on the next section.

4 Case Study

In this study a set of 25 jobs (jobs, J1… J25), including a set of tasks, varying from 1 up to 5, have been considered to be processed on manufacturing resources, among a set of 5 available (R1,…, R5), to analyze, through the application of a SNA technique, measures of centrality, closeness and betweeness, among other relevant descriptive statistical analysis for supporting the establishment of an appropriate plant layout for producing the jobs.

In a social network analysis first a network has to be modelled. Therefore, it was created a matrix with all ties identified between the jobs and the corresponding manufacturing resources, for producing the jobs. This data is presented in the affiliation matrix expressed in Table 1 below, were 1 is assigned in cases when a given manufacturing resource processes an operation on a given job and 0 for the opposite situation. The matrix was uploaded in the software UCINET, which was the software tool used for the SNA technique execution.

After that, using the same software, the social network graph was created (Fig. 1). In this figure are represented all the ties that occurred between the different entities (jobs and manufacturing resources) for accomplishing the underlying tasks. With this social network graph it is possible to realize how many manufacturing resources the jobs interact with, and also to observe the entities that have more intense activity in this production scenario.

5 Experimentation Method Based on the Social Network Analysis

In this paper and experimentation method based on Social Network Analysis (SNAM) is proposed for enabling to describe how the manufacturing execution data can be extracted and viewed as a network with nodes. Consequently, through our considered case study the data has been extracted, and which has been analysed through various SNA tools. Finally, we did carry out a process for identifying different characteristics of the obtained

Table 1. Jobs affiliation matrix

	R1	R2	R3	R4	R5
J1	0	1	0	0	0
J2	0	0	0	1	0
J3	1	1	0	0	0
J4	0	0	1	0	0
J5	0	1	0	0	0
J6	0	1	0	1	0
J7	0	0	1	1	0
J8	1	0	0	0	1
J9	0	0	1	0	0
J10	1	1	1	1	1
J11	1	1	1	1	1
J12	0	0	0	0	0
J13	0	1	1	0	1
J14	0	0	0	0	0
J15	0	1	1	1	1
J16	1	1	0	1	1
J17	1	1	0	1	0
J18	0	1	1	0	1
J19	0	1	0	0	0
J20	0	1	1	1	0
J21	1	1	0	0	1
J22	1	0	1	1	1
J23	0	0	0	0	1
J24	0	1	1	0	1
J25	0	1	0	1	0

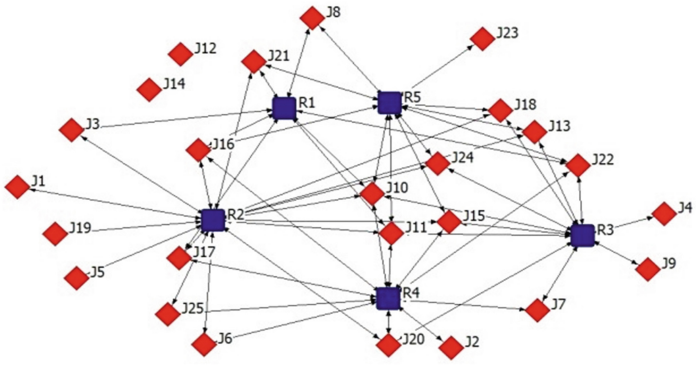


Fig. 1. Social network graph of the production scenario.

network in detail. The SNAM is categorized into two steps: (a) network modelling, and (b) network analysis as it is mentioned in the following Sects. 5.1 and 5.2.

5.1 Network Modelling

A network consists on a set of nodes connected through ties, which indicate interaction. In this section is briefly described how the manufacturing system execution data can be represented as a network. In this step we have fed the input data which is in the form of affiliation matrix into the UCINET software package. Using Net draw we have represented the matrix in the form of a collaboration network. The network is a highly expressive and meaningful way of data representation by enabling to show a variation in the representation of the resources and jobs in terms of shape, size, and colour. The considered resources are represented with a blue coloured square shape and the resources are shown as a red diamond. The arrows connecting the jobs and resources clarify that a particular resource is required to complete a given job. This collaborative network can thus be effectively used to understand which resource is highly influential and the inter relations among them. With the data that has been collected from the case study and their corresponding details through the use of three centralities, a visual interpretation is possible, as can been seen in Table 2. The relationship between the attributes (jobs and resources) is represented in Fig. 1.

5.2 Network Analysis

The main objective of network analysis is to breakdown and comprehend the complex information of the structure into collaboration networks for the extraction of potential synergies. In order to obtain the information of the structure, crucial features about descriptive statistics such as degree centrality, betweeness centrality, and closeness centrality about the network have to be considered to examine the complexity, interdependencies and interrelationships involved on it.

In this research, we have considered the three most popular centrality measures such as Freeman's degree, closeness and Freeman's betweenness centrality about each attribute. The centrality is used to find how influential a node is in the network and also the interrelations among them for its complete analysis. In the Table 2 below, values of the three centralities about the collaboration networks for the above referred scenarios regarding 5 resources and 25 jobs is presented.

The degree centrality measures influence on the node from and to its closest neighbour with complexity of O(n) to linearly scale the nodes in the network, where n is the number of nodes. The jobs and resources with higher degree centrality represent strongly connected ones, whereas the jobs and resources with lower degree centrality exhibit very less connections. Thus, we have identified the key resources which are having higher degree centrality and can act as hubs and also serve as the central elements of the industrial plant.

The other two centrality measures betweenneess and closeness give us information about the shortest path involved among the various attributes of the network.

Table 2. Centrality measures for the resources

	Degree centrality	Betweeness centrality	Closeness centrality
R2	16	142.896	27.619
R3	11	76.180	25.217
R4	11	64.360	25.217
R5	11	60.580	25.217
R1	8	24.977	23.967
J10	5	20.686	26.606
J11	5	20.686	26.606
J15	4	13.928	26.126
J16	4	10.613	25.217
J22	4	11.284	24.786
J18	3	7.478	25.217
J24	3	7.478	25.217
J13	3	7.478	25.217
J20	3	7.600	24.786
J17	3	5.292	24.370
J21	3	4.730	23.967
J3	2	1.706	23.200
J25	2	2.057	23.577
J6	2	2.057	23.577
J7	2	2.095	22.481
J8	2	0.831	21.481
J5	1	0	22.137
J1	1	0	22.137
J19	1	0	22.137
J2	1	0	20.567
J9	1	0	20.567
J23	1	0	20.567
J4	0	0	20.567
J14	0	0	0
J12	0	0	0

The betweenness centrality has the higher level of control on the information floating between different nodes in the network while the closeness centrality is a measure of how closely the nodes are connected with each other. This data can be used to analyze the order in which the resources of the plant need to be arranged and how close each resource need to be from each other.

From the above analysis we have gained complete understanding of the various relations among the resources to complete the assigned jobs effectively. The resources that are key machines as hubs of the industrial plant used for completing maximum number of jobs are identified by degree centrality. Also, the arrangement of the resources in the plant layout based on their interrelations is also understood through the

betweenness and closeness centrality measures. In this way the complete industrial layout can be designed through the obtained statistical data.

6 Conclusion

This paper focused on the application of SNA on an industrial plant layout problem. The study aimed at analyzing the importance of using SNA techniques to examine important relations between entities in a manufacturing environment, such as jobs and resources in the context of industrial plant layout analysis. The study carried out enabled to obtain relevant results for the identification of relations among these entities for supporting to establish an appropriate plant layout for producing the jobs.

Future work is planned for continuing to explore the application of SNA techniques to this kind of industrial plant layout problem, for instance, by exploring this problem in the context of extended manufacturing environments, namely including a big data analysis, and also to compare the application of SNA techniques with some other existing methods for industrial plants layouts establishment and analysis.

Acknowledgments. This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – The Foundation for Science and Technology within the Project Scope: UID/CEC/00319/2013.

References

1. Schott, J.: Social Network Analysis. Sage, Newbury Park (2013). www.orgnet.com/sna.html
2. Hollenbeck, J.R., Jameson, B.B.: Human capital, social capital, and social network analysis: implications for strategic human resource management. Acad. Manage. Perspect. **29**(3), 370–385 (2015)
3. Thongphubate, T., Piekkoontod, T.: Social network analysis on mangrove ecosystem management of Welu Basin, Thailand. Songklanakarin J. Sci. Technol. **38**(3), 243–248 (2016)
4. Fischer, E.N.: Serving more than one master: a social network analysis of section 8 of the clayton act. J. Corporation Law **41**(1), 313–341 (2015)
5. Borgatti, S.: Identifying sets of key players in a social network. Comput. Math. Organ. Theor. **12**(1), 21–34 (2006)
6. Borgatti, S., Halgin, D.: On network theory. Organ. Sci. **22**(5), 1168–1181 (2011)
7. Hawe, P., Webster, C., Shiell, A.: A glossary of terms for navigating the field of social network analysis. J. Epidemiol. Commun. Health **58**(12), 971–975 (2004)
8. Barabási, A.: Linked: The New Science of Networks. Perseus Books, Cambridge (2002)
9. Wellman, B., Berkowitz, S.: Social Structures: A Network Approach. Cambridge University Press, New York (1988)
10. Wasserman, S.: Social Network Analysis: Methods and Applications. Cambridge University Press, New York (1994)
11. Freeman, L.: Centrality in social networks conceptual clarification. Soc. Netw. **1**(3), 215–239 (1979)

12. Opsahl, T., Agneessens, F., Skvoretz, J.: Node centrality in weighted networks: generalizing degree and shortest paths. Soc. Netw. **32**(3), 245–251 (2010)
13. Hanneman, R., Riddle, M.: Introduction to social network methods. University of California, Riverside (2005). http://faculty.ucr.edu/∼hanneman/nettext/
14. Bohn, A., Feinerer, I., Hornik, K., Mair, P.: Content-based social network analysis of mailing lists. R J. **3**(1), 11–18 (2011)
15. Lienert, J., Schnetzer, F., Ingold, K.: Stakeholder analysis combined with social network analysis provides fine-grained insights into water infrastructure planning processes. J. Environ. Manage. **125**, 134–148 (2013)
16. Cross, R., Borgatti, S.P., Parker, A.: Making invisible work visible: using social network analysis to support strategic collaboration. Calif. Manage. Rev. **44**(2), 25–46 (2002). doi: 10.2307/41166121
17. Martíneza, A., Dimitriadisb, Y., Rubiac, B., Gómezb, E., Fuentea, P.: Combining qualitative evaluation and social network analysis for the study of classroom social interactions. Comput. Educ. **41**(4), 353–368 (2003). Elsevier
18. Borgatti, S.P., Li, X.: On social network analysis in a supply chain context. J. Supply Chain Manage. **45**(2), 5–22 (2009). Wiley Online Library. doi:10.1111/j.1745-493X.2009.03166.x

A Genetic Neural Network Approach for Unusual Behavior Prediction in Smart Home

Zaineb Liouane[1,2(✉)], Tayeb Lemlouma[2], Philippe Roose[3], Fréderic Weis[2], and Hassani Messaoud[1]

[1] LARATSI, Monastir University, Ibn El Jazzar, Monastir, Tunisia
liouane.zaineb@yahoo.fr
[2] IRISA, Rennes1 University, Rennes/Lannion, France
[3] LIUPPA/T2I, Pau and the Adour Countries University, Anglet, France

Abstract. Detect efficiently the activities of daily living of elderly people at home in order to provide a secure life and to intervene in the necessary time is an important problem we propose here an improved artificial neural network model. As we need an efficient prediction model, we propose a recurrent output neural network model (RO-NN) combined with a genetic algorithm (GA) which surely monitors and predicts the state of the concerned elderly person. Furthermore, we propose a prediction algorithm "Unusual Behavior Algorithm (UBA)" dedicated to detect the unusual activities and hold us account in the dangerous state.

Keywords: Unusual behavior algorithm · Recurrent output neural network · Genetic algorithm · Elderly people · Smart home

1 Introduction

The growing number of the aged person is remarkable in the recent years [1]. Consequently, the concept of smart home becomes very promising in the aim of providing solutions for elderly that respect their autonomy and their intimate life. Also, it is know that the behavior of the elderly is very sensitive then it is necessary to monitor and supervise the daily activities of the person in order to detect any unusual behavior. Functional tasks in the daily lives of older people are divided into two parts, ADL's and IADL's [2–4]. Activities of daily living (ADL) are the basic tasks of everyday life, such as eating, dressing, grooming, washing and unary function. The instrumental activities of daily living (IADL's) are the activities that people can do since they are awake, such as dressing, homework, phone use, etc. In this paper, we focus our thoughts to monitor the activities of daily living of seniors evolving in smart home environment. We propose an new algorithm, *Unusual Behavior Algorithm* (UBA), in order to detect any unusual and dangerous state in order to maintain the behavior of the person and to provide prevention. The proposed algorithm is based on the routine of daily activities, the routine duration of activities and the number of routine of elapsed activity. Thus, to monitor persons and to predict efficiently with great accuracy their behavior state, it is necessary to define an efficient prediction model. Various types of Artificial Neural Networks

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_73

(ANN) were used in the literature [5, 6], such as Multi-Layer Perceptron (MLP), Echo State Networks (ESN), Radial Basis Function networks (RBFN), Boltzmann Machine, etc. ANNs are used to apply case-based reasoning applications for intelligent homes. Some applications are elder care, health care and emergency solutions. The result of the recent researches shows that neural networks seem effective in different applications, for this reason, in this paper, we propose to improve a specific type of the ANN in the aim of providing a good prediction model. We propose a recurrent output neural network model (RO-NN) to predict the usual behavior of the monitored elderly. We also propose to integrate a genetic algorithm (GA) in the learning step in order to evolve the rate of accuracy. So, the output data of the model who are the basic data of UBA to detect the abnormal state. Our work is based on real data, which are the activities provided by a Functional Autonomy Measurement System (SMAF) profile P_1 of one monitored.

This paper is structured as follows: the next section presents the related works. In Sect. 3, we describe our proposed model, the RO-NN based on GA. Section 4, is devoted to the proposed algorithm (UBA). Section 5 presents the results of the implementation and the simulation of the RONN-GA. Finally, Sect. 6 concludes this work.

2 Related Works

In the context of smart homes literature, several works concentrate about the recognition activities models and methods in order to improve the quality of the monitoring. Most researches focus on the Artificial Neural Network (ANN) in order to evolve the recognition accuracy rate.

To monitor disabled people in their life, Hussein et al. [7] propose to use two types of Neural Network "the Feed-Forward Neural Networks and the Recurrent Neural Networks". The neural network testing results of [7], show the importance of the NN for prediction activities in a smart home. Teich et al. [8] build a *Stable Neural Network* (SNN) based on a feed forward neural network in learning step. The results show that a supervised feed forward neural network with multiple hidden layers gives a good result after a short training period. Fran et al. [9] use a back propagation neural network with a feed forward strategy in the learning process to recognize the Activities of Daily Living (ADL) in a smart home. Thus, [9], reveals that the bigger number of neurons in the hidden layer is, the more accurate the prediction is, however a large number of neurons may result in a large learning time. These authors deduce that the Back Propagation Algorithm (BP) neural network achieves good recognition accuracy. Stefan Oniga et al. [10, 11], design a complex assistive system based on ANN in order to improve the quality of human life, In [10], the authors concentrate on the recognition of the body posture, arm posture and usual activities. Zhigang Liu et al. [12], propose a mechanism for estimation of elderly posture based on BP neural network, these authors conclude that the BP algorithm has a high accuracy rate and robustness in the elderly posture recognition.

View the obtained results used by the ANN models in our work we investigate to propose a recurrent output neural network based on Genetic algorithm in the learning step to predict efficiently the usual behavior of our elderly in the aim to detect any behavior change using our proposed algorithm.

3 RO-NN Based on GA Model

3.1 Structure of the RO-NN

The architecture of the neural network typically consists of an input, a hidden layer and an output layer. Each node in those layers is linked using a weighted connection called weight (W) and a Bias (B) providing additional adjustable parameters of the model and the transfer function (f) that can have different forms (Log function, Tan hyperbolic function, Linear function, Radial basis function, sigmoid function, etc.) calculates the output (y). Therefore, we notice that a neuron network is characterized by a triplet (Weight W, Bias B and Transfer function f). In this paper, we propose a recurrent output neural network (RO-NN) model consists of one input layer, one hidden layer and one output layer, the simplicity of this architecture directly improves the speed of our model. To improve the prediction accuracy, we also include the recurrence notion, so that, the intervention of the previous output in each iteration in the input layer increase the rate of the accuracy.

The input data consists of the activities of daily living, a real data sets (more detail in the Sect. 4.1).

The output (the predicted data) depend on the lagged in inputs and in the following outputs

$$y(t) = f\,(x(t), x(t-1), \ldots x(t-N), y(t-1), \ldots y(t-M)) \tag{1}$$

y(t) denotes the predicted output expected at time *t*; *x(t)* is the real input; *y(t−1)* is the previous output; *f* is the nonlinear function, *N, M* is the input and the output memory orders. After the structure initialization step, RO-NN is characterized furthermore by two steps the learning and the validation. Learning step is necessary to perform the actual learning. It consists in adjusting weight and bias. It enables determining the best parameters of the model and performing a good prediction model. View the importance of the learning step of the RO-NN model and in the aim of evaluating the accuracy performance we propose to integrate a genetic algorithm (GA) in the learning step in order to perform a best learning parameters (W and B) [13].

3.2 Learning Algorithm of the RO-NN

Genetic Algorithms consist in heuristic search and optimization techniques that mimic the process of natural evolution [14, 15]. Genetic Algorithms implement the optimization strategies by simulating evolution of species through natural selection. GA work on a population of individuals instead of individual. At the beginning of the computation a number of individuals (the population) are randomly initialized. The Genetic algorithm model operates like natural processes such as selection, recombination and mutation. The first/initial generation is produced. After the data coding, the genetic-algorithm operators are applied [16, 17].

Selection Operator: Individuals are selected according to their fitness for the production of offspring. The primary objective of the selection operator is to emphasize the good solutions and eliminate the bad solutions in a population while keeping the population size constant.
Crossover Operator: Parents are recombined to produce offspring. Crossover is an operation that exchanges part chromosomes between a pair of parent individuals with a relatively large probability (crossover probability) and produces two new individuals.
Mutation Operator: All offspring will be transferred with a certain probability (mutation probability) this operator allows to introduce the novelty of the population.

However, there is no guarantee that the new offspring obtained will produce good solutions. The Genetic-algorithm operators continue until an optimal solution is found. If the optimization criterion is not satisfied, the creation of a new generation starts.

3.3 Concepts of the RO-NN Based on GA Model

This paper proposes to combine GA with RO-NN to form an improved prediction model.

The algorithm is formed by three steps:

Step 1: Determine RO-NN topology. Determine the number of input nodes, define the number of output neurons based on output parameters, and then define the optimal number of hidden nodes. Finally, we can get the individual length of GA.

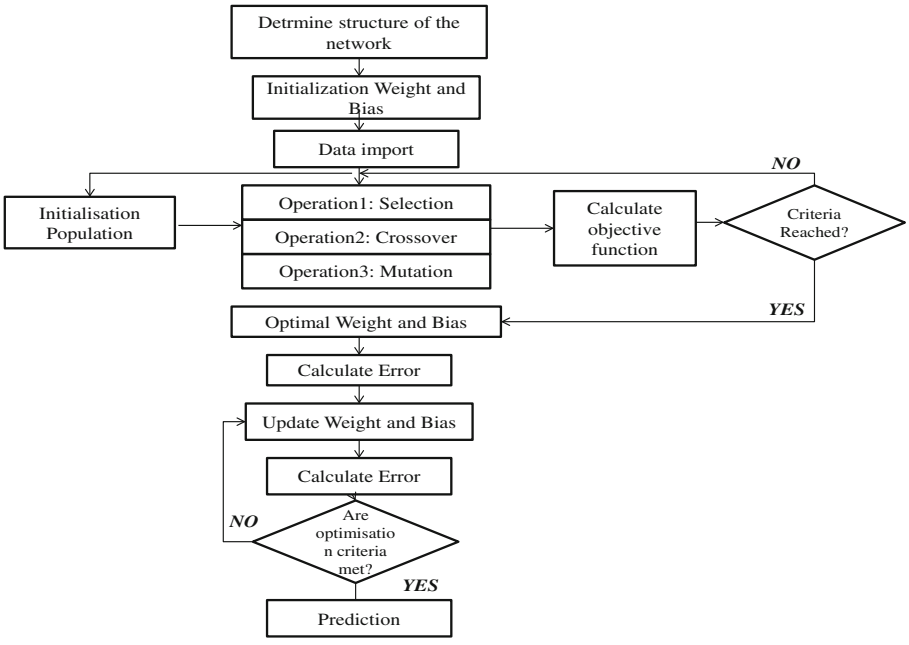


Fig. 1. The flow chart of RO-NN based on GA in the learning step.

Step 2: Utilize GA to optimize RO-NN weights and bias. Generate a population randomly; whose individual represents network weights and Bias. Then compute the fitness value through fitness function and find the best individual by selection, crossover and mutation operations.
Step 3: Use improved RO-NN to predict. Initialized with the best individual, RO-NN weights and Bias can be local optimized again during training. The optimized RO-NN can obtain an accurate prediction and an excellent efficiency.

The flowchart of combinational RONN–GA model is described in Fig. 1.

4 Unusual Behavior Algorithms

In this section we consider the abnormal behavior detection, thus the proposed algorithm based on the output data of the RO-NN who is the predicted data. So, from the predicted data, we will try to detect and identify the abnormal state of the elderly.

4.1 The Concerned Data

To ensure a secure life at home we must know all the routine activities of the occupant in order to be able to remotely recognize (without disturbing) unusual activities. In our work we focus on to monitor the activities of daily living of the elderly people to detect any presence of the unusual behavior. Activities of daily living ADLs are a basic self care tasks. They include eating, dressing, grooming, washing and unary function. In this paper, we use a real data collected from a health monitoring project" e-health monitoring open data project". These data concern the elderly in his smart home during one year. This database includes three parameters "starting time, ending time and activities" based on SMAF profile P1. In our work we are interested in the activities of daily living ADLs. Table 1 shows the ADLs with a specific code for each activity.

Table 1. The activities of SMAF profile P_1.

ADL	Activity	Code
Eating	Eating	1
Dressing	Dressing	2
Washing	Take a shower	3
Grooming	Washing hand/face	4
	Hair dray	5
	Move dish	34
	Make up	6
Unary function	It takes same toileting action	7

Figure 2 shows the usual scenario of the elderly during a week, each activity identified by a code is given in Table 1. The unusual scenarios of daily living is represented by any presence of changes in the daily living routine. So any perturbation in the routine activities indicates that our patient has a potential problem in those activities. Figure 2

shows four week scenarios (Week1, Week2, Week3, and Week4) of the monitored elderly. As we can notice, the structure of each week scenario is maybe different. However, this is not sufficient to identify with any certainty how many unusual states are in this period.

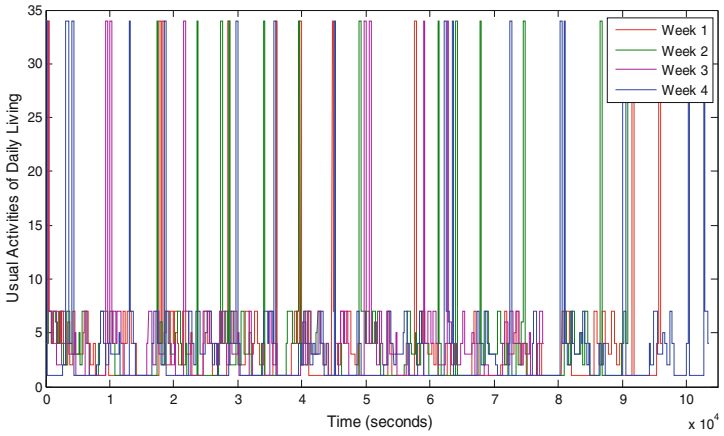


Fig. 2. The different scenarios of the activities by different week.

Sometimes the perturbation of the usual activities indicates a perturbed state of the person all depends on the disturbances of the activity duration. So in our work we propose to take care on the routine duration of each activity. Figure 3 shows the routine number of each elapsed activity by week during a month in the aim to know approximately the routine number of each activity.

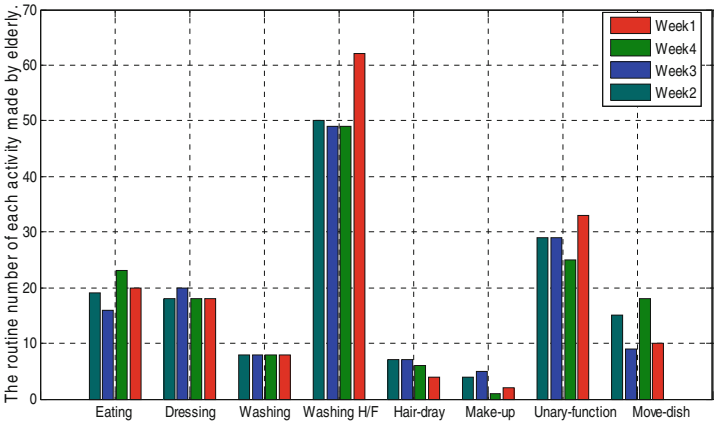


Fig. 3. The routine number of elapsed activity by week during a month.

Another crucial factor, it is very important in the abnormal activities detection that is the routine duration of each activity, so the routine duration of the activity encourage

us to really know the behavior state of our elderly. Then it is necessary to study the routine duration of activities (ADL) in different period of time in the aim to having important information about our seniors behavior which will help us thereafter in our proposed algorithm.

Figure 4 shows the routine duration of some activities (Eating, Dressing, Washing Hand/Face and Unary-function) during 4 weeks in the aim of detecting almost the maximum and the minimum of the routine duration. From Fig. 4 we can deduce each minimum limit and maximum limit of the duration from each curve. For example the curve one shows the routine duration of the eating activity during 4 weeks, the same about the dressing, washing H/F, and Unary-function. The following equations show how to compute the threshold of the min/max routine duration.

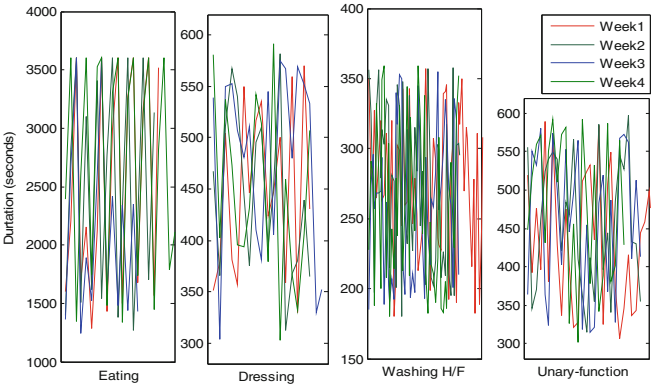


Fig. 4. The routine duration of activities (eating, dressing, washing hand/face and unary-function) in different weeks.

Max routine duration activity threshold (Max RDA):

$$\frac{\sum_{i=1}^{n} min(durtion\ activity)}{n} + logical\ adjustment \tag{2}$$

Min routine duration activity threshold (Min RDA):

$$\frac{\sum_{i=1}^{n} max(durtion\ activity)}{n} - logical\ adjustment \tag{3}$$

The logical adjustment is a parameter responsible to determinate the min and max duration thresholds, It is fixed according to the state of the person (health state, age, behavior, etc.).

4.2 Concepts of the Unusual Behavior Algorithm

The goal of our proposed algorithm untitled "Unusual Behavior Algorithm (UBA)" is to protects and monitors the life of elderly people. Firstly, our algorithm takes into

consideration any unusual activity based on the min/max routine duration during a week, so each detection of unusual duration of activity is saved. Then for each week we control the result of the unusual duration number. Secondly, we can evaluate that our person has a normal state or a behavioral disturbance from the unusual duration number. For this purpose, we must propose a parameter that allows us to fix the state of the person, so the unusual duration threshold number is defined depending on the person's behavior. In the following we present the pseudo code of our algorithm.

Pseudo code of UBA:

```
Var
  Duration activity (DA) = End time activity-Start time activity;
  Unusual Duration Activity Number = UDAN;
  Unusual Duration Threshold Number = UDTN;
While time< 604800 (week duration/sec)
    Switch (ADL){
       Case{Actvity(n) }
        If  DA(n) < MinRDA || DA(n)> MaxRDA
                UDAN(n)= UDAN(n)+1;
         End
     End
  End
if time= 604800 (week duration/sec)
      If UDAN (n)> UDTN
        unusual activity state =1
      else
        unusual activity state =0
End
  End
```

5 Implementation and Results

To evaluate the performance of the RO-NN based on GA, we must calculate the prediction error. The error function indicates how the prediction of our network is close to the target values and, therefore, what adjustment should be applied to the weight and bias in the learning algorithm in each iteration. Therefore, we evaluate *the root mean square error* (RMSE) given by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i)^2} \tag{4}$$

Where, N is the number of learning observations, Y_i represents the predicted data (network output), and X_i represents the real data (network input) of the i^{th} observation.

In order to have a better result, we determine at the beginning the best structure of our model (input, output and hidden neuron number). Firstly, we study the impact of the hidden and output neuron number on the certainty rate and execution time to select a

better structure. Table 2 shows the different structure of the RO-NN-GA and their impact on the RMSE and the learning elapsed time.

Table 2. Comparison of the structure (number of hidden and output neuron) of the RO-NN-GA model in terms of the error and the elapsed learning time.

Number of hidden and output neuron of the RO-NN-GA				
	Hidden neuron 16 Output neuron 8	Hidden neuron 10 Output neuron 4	Hidden neuron 8 Output neuron 4	Hidden neuron 6 Output neuron 4
RMSE	0.00250	0.00155	0.00100	0.0009
Elapsed learning time (seconds)	3025,2098	2971,2806	2803,5581	2690,7309

So, from Table 2 we notice that the RO-NN based on GA has a good result from the structure that comprise a 6 hidden neuron in the hidden layer and 4 output neuron about the output layer in terms of the RMSE and the elapsed learning time. Consequently, we execute the results of the RO-NN based on GA using one input neuron, 4 output neuron and 6 hidden neurons, during a month. To confirm the performance of the model we must compare the RMSE obtained by RO-NN based on the GA with the NARX model based on multi-layer back-propagation algorithm and the Elman model based on multi-layer back-propagation algorithm Table 3 summarizes all results showing better performance of our proposition.

Table 3. The RMSE of the predicted results of the RO-NN-GA model, NARX based on multi-layer back-propagation algorithm, and Elman based on multi-layer back-propagation algorithm.

Model	NARX based on multi-layer back-propagation algorithm	Elman based on multi-layer back-propagation algorithm	RO-NN-GA
RMSE	0.0517	2.7800	0.0009

Then we can conclude that the RO-NN based on GA is more efficient and precise than the other model since it is a stochastic algorithm. So, we notice that the concept of the recurrence and the use of the genetic algorithm in the learning step increase the certainty of the model furthermore. We can conclude that we can detect the unusual behavior of the elderly with a great accuracy. When the neural network is trained, it is necessary to go to the validation step to see the performance of our model and assess the error rate, for this reason we must evaluate our prediction results. We compare the predicted data with the real data. Figure 5 shows the actual and the predicted activities of the person according to the time.

Figure 5 shows the similarity of the predicted duration curve and the real one which shows the accuracy of our model throughout the time. So, we can say that the RO-NN based on GA can predict the usual behavior which helps us to detect any behavior change with a great certainty rate using UBA from the predicted duration.

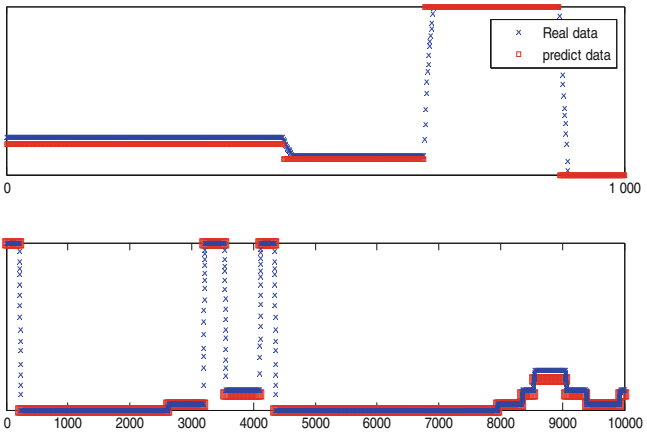


Fig. 5. The results of the prediction data predicted with RO-NN-GA compared with the real data in different interval time.

6 Conclusions

The efficient monitoring of activities of daily living regarding monitored persons in smart homes and the detection of the unusual scenarios represent on of the important issues that have been studied in the recent years. The identification of any unusual behavior or situation is very critic. In this paper, we proposed a prediction unusual behavior algorithm (UBA). The UBA is based on the routine of daily activities, the routine duration of activities and the number of routine of elapsed activity. From the UBA we can effectively identify the unusual activities, then, evaluate the behavior state of our elderly. Subsequently in this work, a recurrent output neural network RO-NN was proposed based on a genetic algorithm (GA) during the learning step. The objective was to increase the certainty of the prediction. The provided results based on the RMSE confirm that our proposed model RO-NN based on GA gives better results if compared to other models proposed previously in the literature. So we can conclude that RO-NN-GA and UBA can help the supervisor to monitor the elderly and timely intervene in the emergency perturbation state with a great confidence.

Acknowledgments. The authors would like to thank all teams of the project "e-health monitoring open data project".

References

1. Skouby, K.E., Kivimäki, A., Haukiputo, L.: Smart cities and the ageing population. In: The 32nd Meeting of WWRF (2014)
2. Covinsky, K.E., Palmer, R.M., Fortinsky, R.H.: Loss of independence in activities of daily living in older adults hospitalized with medical illnesses: increased vulnerability with age. J. Am. Geriatr. Soc. **51**(4), 451–458 (2003)

3. Katz, S., Downs, T.D., Cash, H.R.: Progress in development of the index of ADL. Gerontologist **10**(1), 20–30 (1970). Part 1
4. Zaineb, L., Tayeb, L., Philippe, R., Fréderic, W., Hassani, M.: A Markovian-based approach for daily living activities recognition. In: International Conference on Sensor Networks (SENSORNETS) (2016)
5. Monjezi, M., Hasanipanah, M., Khandelwal, M.: Evaluation and prediction of blast-induced ground vibration at Shur River Dam, Iran, by artificial neural network. Neural Comput. Appl. **22**(7–8), 1637–1643 (2013)
6. Esfe, M.H., Saedodin, S., Bahiraei, M.: Thermal conductivity modeling of MgO/EG nanofluids using experimental data and artificial neural network. J. Thermal Anal. Calorim. **118**(1), 287–294 (2014)
7. Hussein, A., Adda, M., Atieh, M.: Smart home design for disabled people based on neural networks. Procedia Comput. Sci. **37**, 117–126 (2014)
8. Teich, T., Roessler, F., Kretz, D.: Design of a prototype neural network for smart homes and energy efficiency. Procedia Eng. **69**, 603–608 (2014)
9. Fang, H., He, L.: BP neural network for human activity recognition in smart home. In:2012 International Conference on Computer Science and Service System (CSSS), pp. 1034–1037. IEEE (2012)
10. Oniga, S., Sütő, J. : Human activity recognition using neural networks. In: 2014 15th International Carpathian Control Conference (ICCC), pp. 403–406. IEEE (2014)
11. Oniga, S., Suto, J.: Activity recognition in adaptive assistive systems using artificial neural networks. Elektronika ir Elektrotechnika **22**(1), 68–72 (2016)
12. Liu, Z., Song, Y., Shang, Y.: Posture recognition algorithm for the elderly based on BP neural networks. In: The 27th Chinese Control and Decision Conference (2015 CCDC), pp. 1446–1449. IEEE (2015)
13. De, J., Kenneth, A., Spears, W.M., Gordon, D.F.: Using genetic algorithms for concept learning. In: Grefenstette, J.J. (ed.) Genetic Algorithms for Machine Learning, pp. 5–32. Springer, New York (1993). doi:10.1007/978-1-4615-2740-4_2
14. Jiang, C., Fugen, S.: Forecasting chaotic time series of exchange rate based on nonlinear autoregressive model. In: 2010 2nd International Conference on Advanced Computer Control (ICACC), pp. 238–241. IEEE (2010)
15. Whitley, D.: A genetic algorithm tutorial. Stat. Comput. **4**(2), 65–85 (1994)
16. Deb, K., Pratap, A., Agarwal, S.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002)
17. Kazarlis, S.A., Bakirtzis, A.G., Petridis, V.: A genetic algorithm solution to the unit commitment problem. IEEE Trans. Power Syst. **11**(1), 83–92 (1996)

New Adaptive Resource Allocation Scheme in LTE-Advanced

Radhia Khdhir[1(✉)], Kais Mnif[2], Aymen Belghith[3], and Lotfi Kamoun[1]

[1] ENIS, LETI Laboratory, University of Sfax, Sfax, Tunisia
rkhdhir@gmail.com, lotfikamoun2@gmail.com
[2] ENETCOM, University of Sfax, Sfax, Tunisia
kais.mnif@gmail.com
[3] University of Sfax, Road of Aeroport Km 0.5, 3029 Sfax, Tunisia
aymen.belghith@gmail.com

Abstract. To cope with the increasing demand of multimedia applications, Long Term Evolution Advanced (LTE-A) networks were introduced. Indeed, this promising technology supports a high transmission rate, a wide range of voice (VoIP) as well as video services and data. Moreover, LTE-A guarantees lower latency and higher spectral efficiency compared with Long Term Evolution (LTE). However, it is necessary to efficiently benefit from scarce network resources in order to achieve high performance. Therefore, we propose, in this paper, a new scheduling algorithm for LTE-A uplink systems entitled Adaptive Resource Allocation Process (ARAP) that supports uplink multiuser transmissions. The main contribution of this work is to allocate the Physical Resources Blocks (PRBs) to Real Time (RT) and non-Real Time (NRT) users using various criteria. Our proposed algorithm aims to at providing a fair distribution of PRBs, maximizing the system throughput and reducing queuing delay of the RT packets. Simulation and analysis showed that ARAP has a significant positive impact on delay and throughput performance.

Keywords: QoS · Scheduling · LTE-A · ARAP · Throughput

1 Introduction

The Long Term Evolution Advanced (LTE-A) is the enhanced version of Long Term Evolution (LTE) compatible with all versions of the LTE systems. Indeed, an LTE user can easily communicate in the LTE-A network. The International Telecommunication Union (ITU) defined the requirements of the 4G system, called International Mobile Telecommunications-Advanced (IMT-A). Indeed, the 4G system supports higher system capacity compared with LTE. The LTE-A supports a downlink data rate of 1 Gbps and an uplink peak data rate of 500 Mbps [1]. This network specifies that a User Equipment (UE) supports bandwidth up to 100 MHz. To support this high capacity, many techniques were introduced for LTE-A. For instance, we can cite the Carrier Aggregation (CA), Coordinated Multiple Point (CoMP), MIMO-Advanced technology and relay nodes [2]. In CA, multiple LTE Component Carriers (CCs) that

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_74

belong to various frequency bands can be aggregated. So, a higher throughput can be achieved using this technique. An overview of technical issues regarding CA for LTE-A systems was presented in [3].

Call Admission Control (CAC), scheduling and resource allocation in LTE-A are implemented at the evolving NodeB (eNodeB). Indeed, this station is responsible for controlling the frequency and time domain resource in both uplink and downlink transmissions.

The resource in LTE-A is defined by a couple of frequency and time domains named as Physical Resource Blocks (PRB). The management of these PRBs is handled by the scheduling algorithm. The later selects a UE to be scheduled in the next Transmission Time Interval (TTI) based on its QoS requirements. Then, for the resource allocation step, the eNodeB needs some channel quality information perceived by each UE. This channel quality information is obtained by Sending Sounding Reference Signal (SRS) and Channel Quality Indicator (CQI) values. These values are computed by the eNodeB of each PRB for each UE.

In the LTE-A standards, any guideline of scheduling and allocation algorithm has to be defined [4]. Later on, different service flow criteria were used in literature to design and test scheduler. In this paper, we propose a scheduling algorithm for LTE-A. To design and simulate our scheduling algorithm, we select 'Vienna LTE-A' simulator [5, 6]. This simulator contains delay budgets, channel conditions and other LTE-A services parameters.

The rest of this paper is organized as follows: First, Sect. 2 discusses some of the relevant research works. In Sect. 3, we describe the principle of our proposed scheduling scheme called Adaptive Resource Allocation Process (ARAP). Then, Sect. 4 presents the performance evaluation of ARAP in terms of Packet Drop Rate (PDR), throughput and fairness. Finally, Sect. 5 concludes the paper and discusses the future research challenges.

2 Related Work

2.1 Review of Existing Scheduling Schemes

In this section, give an overview of the existing LTE-A scheduling algorithms.

In [7], the authors focused on the CC method selection for macro-cell LTE-A UEs. The UEs aggregate different CCs having various frequency bands with different radio propagation characteristics. To achieve higher performance, the authors proposed load balancing scheduling schemes for the inter-band CA processes.

In [2, 8], two classical packet scheduling schemes were studied. These schemes are Proportional Fairness (PF) and maximum CQI (MAX C/I) exploiting multi-user diversity. The PF allocates PRBs to UEs based of their instantaneous data rate and acquired time average throughput. This scheme takes into account fairness among users. The MAX C/I aims at maximizing the total system throughput. Then, it allocates PRBs to users having the highest channel qualities.

In [9], Zhao et al. proposed a modified version of PF called Similar PF (SPF). This scheduling algorithm, based on the classical PF scheme in crossing CC joint

scheduling, calculates the user weight on each CC and adopts the PF scheduling to calculate the scheduling weight at each TTI.

In [10], Y.-L. Chung et al. developed a new scheduler. Their work aims at designing an Efficient Packet Scheduling Algorithm (EPSA) in LTE-A networks. The proposed scheme takes into consideration the QoS requirements, such as delay and throughput performance. The QoS requirements are studied in both types of traffics: RT and NRT. The scheduler guarantees portion of reserved PRBs for RT packets without neglecting the NRT packets.

In [11], the authors modified the scheduler proposed in [10]. The proposed scheme reserves portion of PRBs to Guaranteed Bit Rate (GBR) and the rest of PRBs for Non GBR (NGBR) and selects a specially period to transmit NGBR packets.

A new scheduling algorithm was introduced in [12]. The authors presented a quantized water-filling packet scheduling algorithm for high data rate users in LTE-A with CA systems. In this work, several QoS performances were provided, especially to enhance the delay performance.

Another algorithm, developed for LTE-A system was presented in [13]. In this work, the traffics are classified into various service classes flows. They attribute different scheduling priorities to each class. After the classification step, the authors followed the Hebbian learning process to assign PRBs to different class of service an adaptively.

The authors in [14] proposed a new scheduling algorithm named Adaptive Hybrid Scheduling Algorithm (AHSA). This algorithm is based on adapting Kwan [15] and PF schedulers in the LTE-A network. The authors analyzed the effect of this algorithm in terms of users' fairness and throughput on the cell performance of PF and Kwan schedulers. They showed that the AHSA scheduler overcomes the uplink delay effect on the scheduler performance.

2.2 Discussion of Reviewed Works

From the overview of the afore-mentioned works, we conclude that:

- Some authors did not consider the QoS requirements of different applications and multiclass traffics as it is the case in [2, 7, 8].
- Some schedulers did not take into account fairness among users, which is the case of schedulers elaborated in [10, 11].

Hence, there is a need for a scheduler that supports both RT and NRT traffics and handles all services flow by considering the priority of each class of service and fairness among users. To tackle these objectives, we designed a new scheduler called Adaptive Resource Allocation Process (ARAP). Mainly, we used a threshold value named ∂ to efficiently manage the RT and NRT queues. Based on ARAP, flows are served based on the RT and NRT queues status and their priority. Our proposed scheduler handles the differentiation between RT and NRT traffics classes. It aims at improving the throughput as well as the fairness. The principle of our proposal as well as its performance analysis will be discussed in the next sections.

3 Proposed Scheme

The objective of our resource allocation algorithm is to optimize the use of the resources, maximize the system throughput and ensure the fairness of resources sharing.

The packets, coming to the network from mixed traffic, are classified into two queues (one for RT packets and one for NRT packets). This classification is useful because the latency of each packet depends on the type of its traffic, by respecting the delay budget (the upper delay bound). These two queues will be served on the basis of ARAP.

Let $tpck_{RT}$ be the RT packet delay and $D_{max(RT)}$ be the RT packet delay budget. Evidently, each RT packet delay must not exceed its delay budget (i.e: $tpck_{RT} < D_{max(RT)}$). If this condition is not respected, the packet will be removed from the RT queue. The same strategy is applied for the NRT traffic. Let $tpck_{NRT}$ be the NRT packet delay and $D_{max(NRT)}$ be the delay budget. Then, condition $tpck_{NRT} < D_{max(NRT)}$ will be respected. $LENG_{RT}$ is the length of RT packets in the queue and THR_{RT} represents the RT queue threshold size. RT packets, which are buffered in the RT queue, are delivered every TTI. However, NRT packets are delivered whenever the condition $LENG_{RT} < THR_{RT}$ is satisfied.

To improve the QoS for RT traffics, we define a new parameter, called ∂. It represents the portion of PRBs reserved for the RT traffic. This parameter is given as follows (Eq. (1)):

$$\partial_{t+1} = \begin{cases} \mathbf{min}(\partial_t + \mathcal{B}, \partial_{max}) & \textit{if } \boldsymbol{LENG_{RT} \geq THR_{RT}} \\ \mathbf{max}\,(\partial_{min}, \partial_t - \mathcal{B}) & \textit{if } \boldsymbol{LENG_{RT} < THR_{RT}} \end{cases} \quad 0 < \mathcal{B} < 1 \tag{1}$$

The ratio of the reserved PRB for RT traffic, in time t + 1, depends on $LENG_{RT}$. Indeed, ARAP increments or decrements the reserved resources for RT traffic based on the RT queue occupation ratio (in previous time t). However, parameter $\mathcal{B}$ is the increment/decrement of the reserved PRBs for the RT traffics. We suppose that parameter $\mathcal{B}$ is a constant value varying between 0 and 1. ∂_{min} is proposed to guarantee a minimum of reserved PRBs for RT class. However, ∂_{max} value is introduced to ensure a minimum number of PRBs for the NRT class in order to avoid starvation. If the $LENG_{RT}$ is equal or higher than THR_{RT}, there will be a significant number of RT packets that are not served. Subsequently, we must increase the number of the reserved PRBs for RT traffics to guarantee a certain level of QoS for these RT packets. However, if the $LENG_{RT}$ is lower than THR_{RT}, then the reserved PRBs for the RT traffic can decrease with $\mathcal{B}$ value to serve the NRT packets waiting in NRT queue. Therefore, our scheme allows decreasing the queuing delay of the RT packets and minimum PDR for RT and NRT traffics by increasing the PRBs allocated to NRT traffic when $\boldsymbol{LENG_{RT}}$ is lower than the threshold. Thereafter, ARAP algorithm serves the RT packets without totally neglecting the NRT packets.

The served users are selected PF metric as defined by (2):

$$\text{metric } (i^*) = \frac{r_i}{R_i} \tag{2}$$

Where r_i is the instantaneous throughput of user i and R_i is the average throughput of user i.

Figure 1 illustrates the main ARAP steps and the steps of ARAP scheme are as follows:

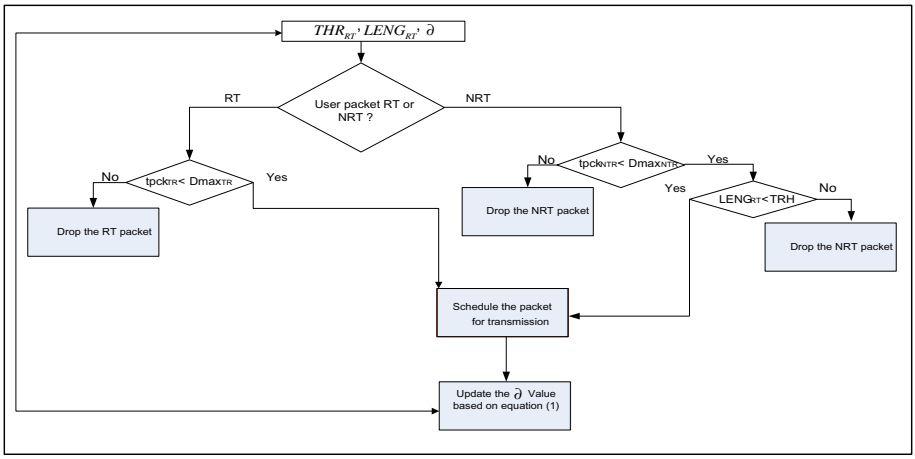


Fig. 1. ARAP design

- Step 1: Deliver RT packets buffered in the RT queue if $tpck_{RT} < D_{max(RT)}$ in the current scheduling s.
- Step 2: Deliver NRT packets buffered in the NRT queue if $tpck_{NRT} < D_{max(NRT)}$ and $LENG_{RT} < THR_{RT}$ in $s + 1$ iteration.
- Step 3: Update the ∂ according to Eq. (1)
- Step 4: Drop the RT and NRT packets if their delay constraint does not respect $D_{max(RT)}$ and $D_{max(NRT)}$, respectively
- Step 5: Repeat the first four steps until all PRBs will be allocated or the RT and NRT queues will become empty.

4 Performance Evaluation

In this section, we present the simulation model and the experimental results obtained by applying the algorithm proposed in Sect. 3.

4.1 Simulation Model

In LTE-A, the evolved NodeB (eNodeB) is responsible for performing the tasks of resource allocation and the Packet Scheduling (PS). The latter is considered as the most important step of Radio Resource Management (RRM) [4]. It contains Time Domain (TD) and Frequency Domain (FD) scheduling algorithms. At each TTI, in the UL direction, the Channel State Information (CSI) is measured directly from the Sounding Reference Signal (SRS) signals issued by the UE. The eNodeB considers this information to decide which UE will be served in the next TTI.

Our scheme is simulated using 'Vienna LTE-A' simulator which supports both networks: LTE Release 8/9 and LTE-A. In the latter considers simultaneously two CCs. We define a number of UEs that varies between 10 and 70. The users' positions are uniformly distributed at the beginning of simulation. The random-walk model is considered as the mobility model.

Requests arrive at eNodeB as Poisson processes with parameter λ. Then, service time is measured by an exponential distribution with mean $1/\mu$. The total system bandwidth, where two adjacent CCs are considered, is equal to 10 MHz. The simulation duration is 1000 TTIs. We compare our proposed scheme with RR, BCQI and ASHA. More details on the simulation parameters are given in Table 1.

Table 1. Simulation parameters

Parameters	Value
System bandwidth	10 MHz
Subcarrier spacing	15 kHz
Number of subcarriers per PRB	12
Number of available PRBs	50
Transmission time interval (TTI)	1 ms
Total number of used subcarriers	600
Carrier frequency	2.5 GHz
Frame duration	10 ms
Number of users	10–70
Simulation Time	1000 TTIs
Link adaptation ACM Modulation	BPSK, QPSK,16-QAM, 64-QAM
$\partial_{min}, \partial_{max}$	0.2, 0.8
Scheduling algorithms	RR, BCQI, AHSA and ARAP

4.2 Simulation Results

In this section, we evaluate the performance of ARAP, RR, BCQI, and AHSA in terms of PDR, number of served users, fairness, and throughput.

4.2.1 Packet Drop Rate (PDR)

The PDR of traffic flow *f* can be measured as shown in Eq. (3) [16].

$$PDR_f = \frac{(\text{Npkt_tx})_f \times (\mathbf{1} - \boldsymbol{PER})_f}{total_\text{Npkt_tx}_\cdot} \tag{3}$$

where Npkt_tx$_\cdot$ is the number of the transmitted packets of the traffic flow f and total_Npkt_tx$_\cdot$ is the total number of transmitted packets of all traffics. PER is the average packet error probability when the traffic flow f transmitted.

The PDR of RT and NRT traffics is presented in Figs. 2 and 3, respectively. As shown in the former, our proposed scheme provides the minimum PDR which can be explained by the fact that ARAP handles the packets buffered in each queues (RT and NRT) adaptively. For the RT traffic, when the number of packets buffered exceeds threshold THR_{RT}, then more PRBs will be allocated to the RT traffic. Therefore, the number of RT dropped packets is minimized. We note that the PDR of ARAP is slightly higher than that of AHSA when the number of users is between 40 and 50. Indeed, the number of PRBs allocated for RT traffic is dynamic. However, AHSA scheduling strategies are generally based on PF and Kawn methods, which makes the number of PRBs, used by RT traffic, relatively fixed.

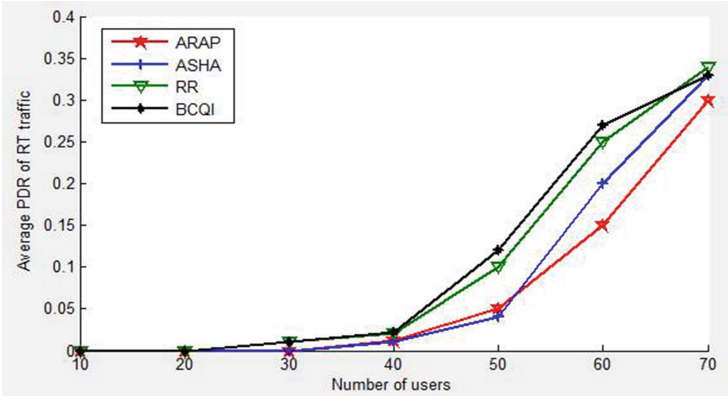


Fig. 2. Average PDR of RT traffic

If the number of RT packets buffered in RT queue is less than threshold THR_{RT}, ARAP decrements the number of PRBs allocated to the RT traffic and assigns more PRBs to NRT traffic. Thus, the NRT packets, dropped by ARAP, is lower than that of the other schedulers (see Fig. 4). We note that the PDR of the NRT traffics is higher than that of the RT traffics because ARAP serves the RT traffics at each TTI *t*. However, the NRT traffics will be transmitted only when the condition ($\boldsymbol{LENG_{RT} < TRH_{RT}}$) is satisfied.

4.2.2 Number of Served Users

The number of served users as a function of the total number of users is shown in Fig. 4. From this figure, we clearly observe that ARAP serves an interesting number of users because it can schedule more users by giving the needful PRBs for each one.

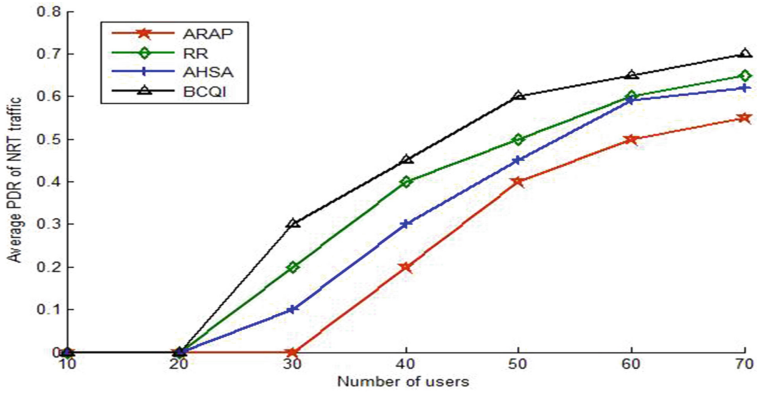


Fig. 3. Average PDR of NRT traffic

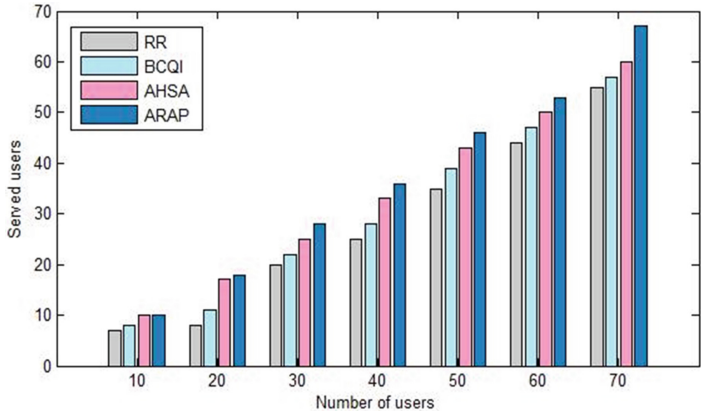


Fig. 4. Number of served users

This allows accepting greater number of users and maximizing the total number of the used PRBs. Indeed, the PRBs is effectively used thanks to the concept of resource allocation algorithm which adjusts appropriately the allocation of resource.

4.2.3 System Throughput

The system throughput is measured as the total number of bits successfully transmitted per second [17].

Figure 5 shows the average system throughput of RR, BCQI, AHSA and ARAP algorithms as a function of the number of UEs. Evidently, BCQI provides the best throughput because it always favors the UEs having the most efficient MCSs. We also observe that ARAP outperforms AHSA and RR schedulers. Indeed, ARAP can serve much more users compared to other algorithms (see Fig. 4). which requires harness the maximum of the available resources blocks which the overall throughput. Moreover, we can classify ARAP scheme as a QoS-aware scheduler. This category of schedulers

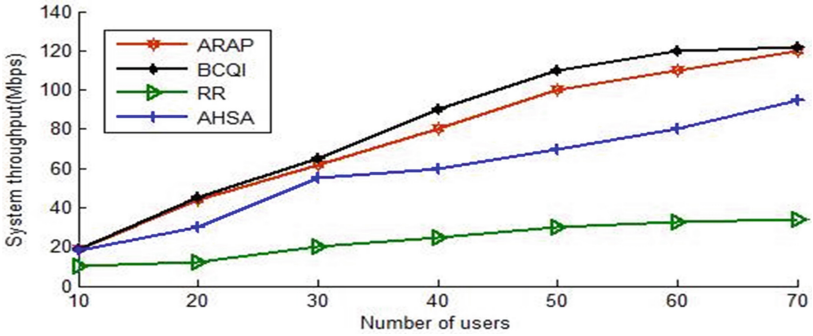


Fig. 5. System throughput

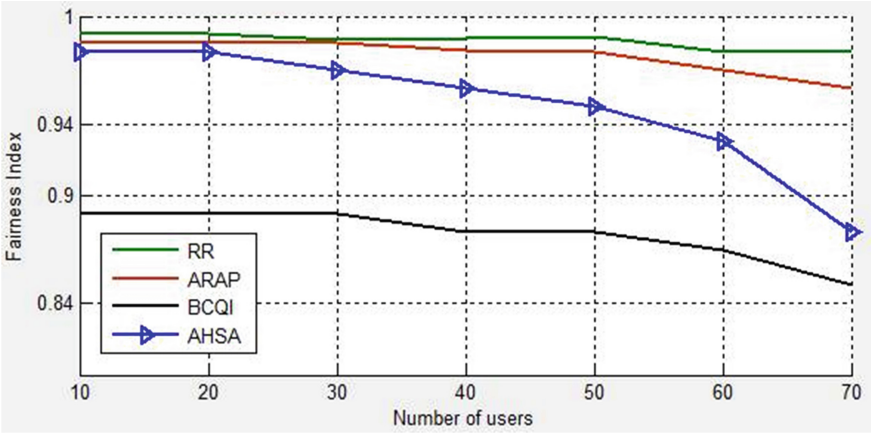


Fig. 6. Fairness index

distinguishes various UEs and allocates more PRBs to higher priority traffics. In this case, ARAP assigns more PRBs for RT traffics if the number of packets, buffered in RT queue, exceeds the threshold THR_{RT}. Briefly, results shown in Fig. 6 prove that ARAP approach has an important positive impact on global system throughput behavior.

4.2.4 Fairness Index

The fairness of the approaches was evaluated using the Jain's fairness index. The definition of this index is stated in [4, 18]. This fairness index is calculated as follows:

$$\boldsymbol{F}\left(\mathbb{C}^1, \mathbb{C}^2, \mathbb{C}^3, \ldots \cup \mathbb{C}^n\right) = \frac{\left(\sum_{\mathbf{j=1}}^{\mathbf{n}} \mathbb{C}^{\boldsymbol{i}}\right)^{\mathbf{2}}}{\boldsymbol{n} \times \sum_{\mathbf{j=1}}^{\mathbf{n}} \left(\mathbb{C}^{\boldsymbol{i}}\right)^{\mathbf{2}}} \tag{4}$$

where n represents the total number of UEs and $\mathbb{C}^i$ denotes the number of resources assigned to user i. Jain's fairness index returns a value between 0 and 1. The latter

corresponds to the best fairness in the system. Figure 6 shows the fairness results obtained by applying RR, BCQI, AHSA and MURPA. We notice that when the number of users increases, it is expected that the fairness index decreases as more UEs compete for the same number of PRBs. The maximum value of Jain's fairness index is obtained when the RR scheduler is used. This result is logical because RR assigns almost the same number PRBs for all UEs. Moreover, we observe that the ARAP and RR curves are close which can be explained by the fact that ARAP adjusts its scheduling decision depending on the previous queue status (depending on threshold THR_{RT} value). Then, ARAP schedules the RT traffic without neglecting the NRT traffic.

5 Conclusion

In this paper, we proposed a new scheduler, named ARAP, to joint together the benefits of the classification services. Indeed, we used a threshold value ∂, to efficiently manage the RT and NRT traffics. Based on ARAP, flows are served based on the RT and NRT queues status and on their priority. Indeed, the RT flows can be served at each TTI. However, the NRT flows can be served only if the condition $LENG_{RT} < THR_{RT}$ is satisfied. Simulation results proved that ARAP serves flows having lower delays and better throughput compared to RR and AHSA schedulers. Particularly, ARAP has a significant positive impact on PDR and throughput performance. Moreover, our scheduling algorithms outperforms RR and AHSA in term of the number of the served users and fairness.

As future work, we plan to extend the proposed ARAP when entering the Remote Radio Head (RRH) in the system. In this case, the LTE-Advanced with several RRHs may use an efficient scheduler to avoid packet losses and flow accumulation in queues.

References

1. 3GPP TR 136 913 v. 10: LTE; Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-Advanced), pp. 1–16 (2011)
2. Tran, T.-T., Shin, Y., Shin, O.-S.: Overview of enabling technologies for 3GPP LTE-Advanced. EURASIP J. Wirel. Commun. Network. **1**, 1–12 (2012)
3. Blanco, M.: Carrier aggregation: fundamentals and deployments. Agilent Technol., 1–34 (2014)
4. GPP TS 36.321 v. 9: LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol, pp. 1–49 (2010)
5. Taranetz, M., Blazek, T., Kropfreiter, T., Müller, M.K., Schwarz, S., Rupp, M.: Runtime precoding : enabling multipoint transmission in LTE-Advanced system-level simulations. IEEE Access **3**, 725–736 (2015)
6. http://www.nt.tuwien.ac.at/forum
7. Tian, H., Gao, S., Zhu, J., Chen, L.: Improved component carrier selection method for non-continuous carrier aggregation in LTE advanced systems. In: Vehicular Technology Conference, pp. 1–5 (2011)

8. Sessia, S., Touffik, I., Baker, M.: The UMTS Long Term Evolution Forum Theory to Practice, pp. 1–794. Wiley, New York (2009)
9. Ji-hong, Z., Hui, L., Hua, Q.: A SPF-PF crossing component carrier joint scheduling algorithm. In: Advanced Communication Technology (ICACT), pp. 173–177. IEEE (2012)
10. Chung, Y.-L., Tsai, Z., Jang, L.: An efficient downlink packet scheduling algorithm in LTE-Advanced systems with carrier aggregation. In: Consumer Communications and Networking Conference (CCNC), pp. 632–636 (2011)
11. Mnif, K., Masmoudi, A., Kamoun, L.: Adaptive efficient downlink packet scheduling algorithm in LTE-Advanced system. In: International Symposium on Networks, Computers and Communications, pp. 1–5 (2014)
12. Chung, Y.-L., Tsai, Z.: A quantized water-filling packet scheduling scheme for downlink transmissions in LTE-Advanced systems with carrier aggregation. In: Software, Telecommunications and Computer Networks (SoftCOM), pp. 275–279 (2010)
13. Kausar, R., Chen, Y., Chai, K.-K.: Adaptive time domain scheduling algorithm for OFDMA based LTE-Advanced networks. In: Wireless Mobile Computing Networking Communications, pp. 476–482 (2011)
14. Barakat, B., Arshad, K.: An adaptive hybrid scheduling algorithm for LTE-Advanced. In: International Conference on Telecommunications (ICT), pp. 91–95 (2015)
15. Kwan, R., Leung, C., Zhang, J.: Multiuser scheduling on the downlink of an LTE cellular system. Res. Lett. Commun. **3**, 1–4 (2008)
16. El-atty, S.M.A., Lizos, K.: Enhanced uplink scheduling for continuous connectivity in high speed packet access systems. In: High Capacity Optical Networks and Enabling Technologies (HONET), pp. 446–453 (2012)
17. Abu-Ali, N., Taha, A.-E.-M., Salah, M., Hassanein, H.: Uplink scheduling in LTE and LTE-Advanced: tutorial, survey and evaluation framework. Commun. Surv. Tutorials **16**, 1–27 (2013)
18. Jain, R., Chiu, D.-M., Hawe, W.R.: A quantitative measure of fairness and discrimination for resource allocation in shared computer system. ACM Trans. Comput. Syst., 1–38 (1984)

An Approach for Measuring Flexibility of Business Processes Based on Distances Between Models and Their Variants

Asma Mejri[1(✉)], Sonia Ayachi Ghannouchi[1,2], and Ricardo Martinho[3,4]

[1] Laboratory RIADI-GDL, ENSI, 2010 Manouba, Tunisia
mejri.assma@gmail.com
[2] High Institute on Management of Sousse, Sousse, Tunisia
[3] School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal
[4] Center for Health Technology and Services Research (CINTESIS), Maia, Portugal

Abstract. The strong and fast competitions in which organizations are involved lead them to frequently change and adapt their business processes. In fact, business process flexibility has been one of the richest topics in business process management for the last two decades, as it gathers all efforts to allow for changes in certain parts of the process, while keeping others stable. Nevertheless, such changes may lead, in turn, to a large number of process variants derived from the same process model, but differing in structure. In this paper, we propose a quantitative approach to measure the flexibility of business processes regarding their different perspectives: functional, operational, behavioral, informational and organizational. We define therefore the perspective-based distances, which is based on computing change operations. The resulting measures can help organizations to better understand how much flexibility do their business processes enclose, taking into account the model variants they've been using. To exemplify our approach, we used a real-world emergency care process as a case study.

Keywords: Change · Business process · Model · Flexibility · Measuring · Quantitative approach · Perspective · Emergency care process

1 Introduction

Business Process Management (BPM) is a management approach that includes the creation, development, maintenance, and optimization of business processes (BPs) in organizations. A BP is defined as "a collection of ordered tasks that need to be carried out to achieve a certain business goal" [1]. Due to dynamic changing environments and their open and competitive context, BP flexibility has become a major challenge in BPM research over the past 20 years. In evolving environments, the essential characteristic of BP flexibility is the balance between BPs and their supporting systems [2].

Although BP flexibility is abundantly mentioned in literature, it is difficult to describe properly in quantifiable and measurable terms what the flexibility of a BP entails [3]. Therefore, this paper addresses the BP flexibility issue focusing on its measurement,

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_75

taking into account the different perspectives and considering the flexible features of a BP. We also consider that changes can occur at the process model or instance levels, or both. Such changes will allow obtaining different modeling variants for a certain BP. Distances between the main BP model and its variants can be measured. In this context, the main contributions, related to measuring distances between BP models, have been brought by [4]. These contributions mainly consider measuring distance between two process models in terms of the number of high-level change operations (e.g., to insert, delete or move activities). Besides, these contributions considered only the variability of BP activities. In this setting, the variability of other process artifacts (e.g. data and resources) has yet to be considered. To overcome this limitation, we focus not only on the functional perspective, but also on a measurement approach that incorporates features for capturing changes affecting the operational, informational, organizational and behavioral perspectives, as well as elements related to the approaches that enhance flexibility.

The main contribution of this paper is to provide a quantitative approach to measure the flexibility of BPs. This quantitative approach is based on how much changes do users need to perform in a process model by comparing it against a collection of its process variants, as well as the history (number) of instances registered for each variant. We exemplify our approach using an emergency care process with four model variants, each of these having a variable number of instances occurred. We combine these data to reach a flexibility measure based not only on a static evaluation of process models, but also taking into account their runtime occurrences, and the number of change operations users need to perform under a certain Business Process Management System (BPMS).

This paper is organized as follows. Section 2 introduces the background of the paper, and Sect. 3 presents the necessary concepts of our methodology for measuring BP flexibility. Section 4 is dedicated to the presentation of the flexible emergency care process, from the healthcare domain. More precisely, this section introduces the emergency care process and illustrates the changes faced by the emergency department. It also contains the flexibility needs of this process. Section 3.4 evaluates the algorithm that we have developed (the simulation). Finally, Sect. 5 recaps our contribution and gives some perspectives for future works.

2 Background

We provide in this section a brief introduction to flexibility in BPM and flexibility measurement.

2.1 Flexibility in BPM

BPM has gained the attention of organizations mainly since the early 1990s. It includes methods, techniques, and tools to support the design, enactment, management, and analysis of operational BPs [5]. In particular, the topic of BP flexibility has become a main focus of attention from both commercial and research institutions, since understanding of requirements for BPs [6], which are able to adapt their behavior to changes in an external

environment, became crucial for supporting organizations operating in a dynamic context. Therefore, flexibility has been then one of the major research topics in the area of BPM for the past two decades. In fact, flexibility is an essential property for the maintenance of fit between BPs and their supporting systems in changing environments [2].

As mentioned earlier, flexibility has been influenced by the evolution of BPM. Indeed, the overriding concern is to adapt BPs to predicted and unpredicted real-world changing identified scenarios. Therefore, it allows reducing the gap between process models and instances digitally registered among business process management systems (BPMSs), and what is happening in the real world, during the execution of a business process.

The BP literature has proposed a multitude of business process modeling approaches or paradigms, each in response to a different business process type with a unique set of requirements [7]. Several approaches have addressed the issue of flexibility and several flexibility taxonomies have been proposed, including the ones by Schonenberg et al. [8], by Regev et al. [9], by Nurcan [5] and by Reichert et al. [10].

2.2 Flexibility Measurement

In literature, there are some explicit criteria that can be exploited in order to measure BP flexibility.

In [11], authors presented a framework for the purpose of comparing of BPs under three dimensions: time (i.e. the process should adapt to change more quickly); cost (i.e. the process should adapt to change with less cost) and ease (i.e. the process should adapt to change with maximum ease). In the same context, other definitions of these same dimensions are proposed in [3], including new definitions: time (i.e. less items), cost (i.e. having the items that have to be changed in one place) and ease (i.e. being able to make the translation from the new requirements for process change to executable workflow in less stages).

Flexibility is defined as "the ability to react to changes" in [12]. Typically, these changes depend on the various parts of the BP. For instance, it concerns the ability of resources to execute different tasks within a BP setting.

Flexibility can also refer to the degree of variation that a process allows. It can be discussed in relation to the event logs that the process produces [12]. In [13], the authors proposed the following set of flexibility performance measures: mix flexibility, labor flexibility, routing flexibility, volume flexibility and process modification flexibility. Moreover, they have specified that flexibility can be identified for individual resources, for individual tasks, for the whole BP.

Another way of approaching the performance dimension of flexibility is to distinguish between run time and build time flexibility. In [12–14], authors separate two types of measurement purposes for process models. The first type is the measurement of the process model design (the "static" or "build time" property), considering the possibility to change the BP structure. The second type consists of measuring results of a process model execution (the "dynamic" or "run time" property). It allows users to handle changes and variations while executing a specific BP.

In this paper, we present a new methodology for measuring flexibility. Besides considering flexibility either at the build time and run-time, we propose to measure it, taking into account the needs to adapt of a certain BP, as well as its collection of process variants. It is moreover increasingly important to distinguish the flexibility of a BP from the other dimensions.

3 Methodology for Measuring Flexibility

In this section, we propose a methodology to measure flexibility, which is based firstly on a set provided definitions. We then define the distances related to the different perspectives. Using this distances, we finally present our algorithm for measuring flexibility in BPs.

3.1 BP Flexibility and Its Different Perspectives

Changes regard the various elements (that compose the BP that can be changed, as presented in the taxonomy of [9] (subjects of change). The subjects of change can be characterized by associating them with different perspectives.

The functional perspective describes activities to perform during process execution. A process performs activities which can be atomic or composite (e.g. sub-processes). Activities are logical units of work [15]. In the behavioral process perspective, an atomic activity can have preconditions and post-conditions. The coordination between activities (conditional control pattern) is then specified by control connectors. The operational perspective defines elementary operations performed into the atomic activities. Activities can execute one or several operations. The informational perspective concerns the production and use of information. An activity consumes and/or produces informational resources. An informational resource can range between system data, application data, or process data. The organizational perspective describes relationships between roles and actors by giving these authorizations to perform atomic activities. An atomic activity is performed by a role, which can be played by several actors. On the basis of these perspectives, a BP encompasses a number of elements that are activities, operations, informational resources, roles, conditions and actors. Generally, the functional perspective is compulsory to be given by the process definition. Regarding the other elements, they can be associated to other perspectives (behavioral, operational, organizational and informational).

On the basis of these perspectives, a flexible BP encompasses a number of elements that are activities, operations, informational resources, roles, control connectors and actors. In order to best take into account the flexibility in BP, we added consider a set of decision elements which are related to the BP modelling paradigms (such as the constraint-based, the rule-based and the adaptive process management) dealing with flexibility in BPM.

3.2 Distances Between Models and Their Variants

A process change is performed by enforcing a sequence of change operations to a given process model. These change operations structurally modify the initial process model. Each accomplishment of a change operation therefore leads to a new process model [16]. In their work [17], the authors consider the following definitions.

Definition 1 (Process change and pocess variant). Let P denote the set of possible process models and C be the set of possible process changes; let $S, S' \in P$ be two process models; let $\Delta \in C$ be a process change; and let $\sigma = [\Delta_1, \Delta_2, \ldots, \Delta_n]$ be a sequence of changes performed on initial model S. Then:

- $S\ [\Delta\rangle\ S'$ is applicable to S and S' is the sound process model resulting from the application of Δ to S.
- $S\ [\sigma\rangle\ S'$ if $\exists\ S_1, S_2, \ldots, S_{n+1} \in P$ with $S = S'$, $S' = S_{n+1}$, and $S\ [\Delta_i\rangle\ S'_{i+1}$ for $i \in \{1, 2, \ldots, n\}$. We denote S' as process variant of S'.

Definition 2 (Change log). Let P denote the set of possible process models. Let $S \in P$ be a process model and let $S_i \in P$ for $i \in \{1, 2, \ldots, n\}$ be its variants.

A change log cL is defined as a set $\{\sigma_i \mid S\ [\sigma_i\rangle\ S'_{i+1}\}$ (σ_i denotes a sequence of change operations transforming S into S_i).

In [18], authors presented how to measure process distance based on change operations which indicate how costly it is to transform process model S into S'. This distance (Definition 4) considers only the functional perspective. Generally, it is possible to have more than one minimal sequence of change operations to transform S into S'.

Definition 3 (Functional distance). Let $S, S' \in P$ be two process models. Let $\sigma_{fun} = [\Delta_1, \Delta_2, \ldots, \Delta_n] \in C$ be a sequence of change operations performed on the operations of the initial model S, transforming it into S'.

Functional distance $d_{fun}(S, S')$ between S and S' corresponds to the minimal number of high-level change operations performed on the set of activities needed to transform S into S'; i.e., we define

$$d_{fun}(S, S') = \min\left\{ |\sigma_{fun}| \,\middle|\, \sigma_{fun} \in C \wedge S\ [\sigma_{fun}\rangle\ S' \right\}$$

In our work, we focus on considering, not only the functional distance, but also the operational, informational, organizational, behavioral and declarative distance:

- Functional distance: deals with the change of the activity set.
- Operational distance: deals with the changes of the operations.
- Informational distance: deals with the change of the data or document or forms.
- Organizational distance: deals with change of the actors.
- Behavioral distance: deals with change of the conditions.
- Declarative distance: deals with changes related to the rules or constraints or cases.

Therefore, similar definitions are given concerning respectively these perspectives. For instance, Definition 4 provides a definition of the informational distance.

Definition 4 (Informational distance). Let S, S′ ∈ P be two process models. Let $\sigma_{inf} = [\Delta_1, \Delta_2, \ldots, \Delta_n] \in C$ be a sequence of change operations performed on the operations of the initial model S, transforming it into S′.

Informational distance $d_{inf}(S, S')$ between S and S′ corresponds to the minimal number of high-level change operations performed on data needed to transform S into S′; i.e., we define

$$d_{inf}(S, S') = \min\{|\sigma_{inf}| \; | \sigma_{inf} \in C \wedge S[\sigma\rangle S'\}$$

To measure the complete transformation from S to S′, we formally define total process distance. Varied transformation types have to be considered in our work and they all impact on our total process distance measure.

Definition 5 (Total process distance). Let S, S′ ∈ P be two process models. The total process distance d(S; S′) between S and S′ corresponds to the sum of the functional, operational, behavioral, organizational, informational and declarative distances that corresponds to the sum of all the operations performed on the initial model S, transforming it into S′; i.e., we define:

$$d(S, S') = (d_{fun} + d_{op} + d_{beh} + d_{org} + d_{inf} + d_{dec})(S, S')$$

We should define the weighted flexibility between a process model S and its variants that represents how close a process model and its variants are.

Definition 6 (Weighted flexibility). Let S ∈ P be a process model. Let M be a set of process variants $S_i \in P$ for i∈{1, 2, …, n}, with w_i, representing the number of process instances that were executed on the basis of S_i. The weighted flexibility W(S; M) between S and M can be computed as follows:

$$W(S, M) = \frac{\sum_{i=1}^{n} d_{(S,S_i)} w_i}{\sum_{i=1}^{n} w_i}$$

3.3 Change Operations

In this paper, we consider that a process change is denoted as a sequence of change operations that were applied to the respective process variants. We assume that these operations can involve a large variety of BP elements. In [1], change operations are: insert activity, delete activity and move activity. In insert and delete operations, the set of activities in a process model are affected by changes. Whereas, the move operation modifies both the activity positions and the structure of the BP model [18]. These evoked change operations were discussed in relation to the ADEPT change framework. However, they are adequately generic in the way that they can be applied in other process

meta-models [1]. In this setting, in our work, the changes in the process include changes concerning the five perspectives specified in the taxonomy of Regev et al. In fact, we focus on altering the process set of activities and their order relations, its set of roles, of information/data, of conditions and operations. We focus also on altering its declarative elements. The declarative elements could be added, deleted or changed. It is important to mention that these change operations can be performed either in run time or build (design) time. In this work, these change operation could be introduced during execution (i.e. concern mainly executable models), or could be during …

3.4 Algorithm

We describe our flexibility measurement approach as a procedure listed in Algorithm 1. We compute first the distances for each process variant and then the total process distance. We then compute finally the weighted flexibility of the process.

Algorithm1 (Flexibility measurement algorithm)

Input: A BP model and a process variant collection
Output: weighted flexibility

- For each process variant compute
  (1) the functional, operational, behavioral, organizational, informational and decision distances
  (2) the total process distance
  (3) the number of process instances
- Compute the weighted flexibility

4 Application Example

In this section we apply our flexibility measurement approach on a real-world emergency care process and discuss the obtained results.

4.1 Case Study: An Emergency Care BP

In this section, we are going to give first a description of the emergency care process. We will present then the main changes that can occur. As a case study to illustrate our flexibility measurement approach we consider an emergency care (EC) process which initiates with the patient registration and ends with her/his discharge of the emergency department. The EC process was analyzed and modeled at first in [19], using the BPMN language. We have validated this process within the scope of this work, using observations and interviews with the main actors of the process. In addition, we arrived to a deeper understanding of the process and its different activities. We have also added the "sorting" activity and derived the final process model.

In [20], the author studied four health crises faced by the emergency department of the Hospital Farhat Hached Sousse between 2007 and 2010. These crises provided significant changes to the EC process, and four main variants of this process could be associated with each one, including: train accident, smoke poisoning, collective food poisoning and AH1N1 flu pandemic.

4.2 Application Example for AristaFlow BPM Suite BPMS

In the next subsections, we present the results we obtained for the EC process modeled in ADEPT2, in terms of measuring flexibility. In summary, we generate four variants related to the different changes for each model. The first step consists of specifying the change log when using the ADEPT2 notation (ARISTAFLOW BPM suite) for each possible change that can occur in the emergency care process. A large number of process instances, being in different states, may run on a particular BP model. Furthermore the four variants are weighted. Within our work, we define the weight wi of a process variant S_i as the number of process instances that were executed on the basis of Si. In our example, 16 instances were executed on the basis of process variant S1, 14 instances ran on S2, 156 instances ran on S3 while 3081 instances ran on S4.

We assume that we take the process model S as a process model to measure flexibility for the AristaFlow BPMS. As presented in Table 1, the total process distance between S and each of the four process variants is given by d(S, S1) = 3, d(S, S2) = 4, d(S, S3) = 5, d(S, S4) = 8.

Table 1. Flexibility measurement for the AristaFlow BPM suite

Distances	Variants			
	S_1	S2	S3	S4
Functional distance	3	2	2	4
Operational distance	0	0	0	0
Behavioral distance	0	0	0	2
Informational distance	0	1	3	1
Organizational distance	0	1	0	1
Declarative distance	0	0	0	0
Total process distance	3	4	5	8
Number of instances	16	14	156	3081
Weighted flexibility	**7.81**			

Considering as well the different weights, we obtain a weighting flexibility of 7.81. This means that we need to perform on average 7.81 high-level change operations to configure a process variant (and related instances respectively out of the process model, using the AristaFlow BPMS.

4.3 Discussion of the Application Example

The application is an illustration of the distance measures on a real-life EC process for which four variants were defined, each as a consequence of some disaster situation. We can analyze the results presented in Table 1, to highlight how important were the changes applied for each perspective (functional, operational, organizational, behavioral and informational), by comparing them to what degree we have performed change operations in each perspective. We have performed in total (for the three different variants) 11 change operations that concern the functional perspective, 5 change operations related to the informational perspective and 2 change operations that concern the organizational perspective. No change operations related to the operational perspective were performed. Besides, changes concerning the behavioral perspective are 2 for the ADEPT2.

Besides, the changes that were performed using the ADEPT2 BPMS, could be performed during run-time and/or build time. We can also conclude that disasters with large numbers of victims get a higher weight, because the weighted flexibility is weighted taking into account the number of instances of a process variant. It is important to mention that victims of AH1N1 flu pandemic are the greatest number of individuals who had an impact on the weighted flexibility. In addition, clearly the less you have to change, the more flexible the BP model is. The ideal is to have a BP so flexible that you could perform the minimal number of change operations in order to comply with a certain process variant.

5 Conclusions and Future Work

In this paper we propose an approach to measure flexibility in BPs, taking into account not only the process model and its variants, but also the number of instances executed. In this context, we presented a method to compute the distance between two process models. We measured the distance between a BP model and a variant, counting the number of change operations required to transform the process model into the respective variants when a change is occurred, considering the different perspectives of a BP. We then introduced an algorithm to evaluate the distance between a process model and its variants. Based on the proposed algorithm, a prototype will be implemented allowing the automatic computing of the weighted flexibility of the considered models according to each modeling perspective. Moreover, we have applied an execution of our real-world emergency care process.

Emergency care process and general healthcare processes are commonly known as having flexibility needs. It will be useful to apply other processes also known to be flexible, such as software development processes based on agile principles. In addition, we are going to study, in future work, the impact of the atomic changes of each perspective on the total process distance, and the impact of the changes on one perspective in the perspective-bound distances. In fact, six different distances between BPs are defined, in this paper, and finally summarized in the definition of total process distance. However, changes in one perspective can be performed independently from changes in another perspective. For instance, adding activities (i.e. functional perspective) might lead to

changes in role allocation (i.e. organizational perspective), changes in operations (i.e. operational perspective), etc. While this is not a problem per se for each of the perspective-based process distance definitions, such relationships may have their impact on the total process distance.

References

1. Gerth, C. (ed.): Business Process Models. LNCS, vol. 7849. Springer, Heidelberg (2013). doi: 10.1007/978-3-642-38604-6. Ph.D. thesis, University of Paderborn
2. Gong, Y.: Engineering flexible and agile services: a reference architecture for administrative processes. TU Delft, Delft University of Technology (2012)
3. van Eijndhoven, T., Iacob, M.E., Ponisio, M.L.: Achieving business process flexibility with business rules. In: Proceedings of the 2008 12th International IEEE Enterprise Distributed Object Computing Conference, EDOC 2008, pp. 95–104. IEEE Computer Society, Washington, DC (2008)
4. Li, C., Reichert, M., Wombacher, A.: On measuring process model similarity based on high-level change operations. In: Li, Q., Spaccapietra, S., Yu, E., Olivé, A. (eds.) ER 2008. LNCS, vol. 5231, pp. 248–264. Springer, Heidelberg (2008). doi:10.1007/978-3-540-87877-3_19
5. Nurcan, S.: A survey on the flexibility requirements related to business processes and modeling artifacts. In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences, pp. 378–378. IEEE (2008)
6. Mulyar, N., van der Aalst, W.M., Russell, N.: Process flexibility patterns. Technische Universiteit Eindhoven (2008)
7. Handl, J., Knowles, J.: Evolutionary multiobjective clustering. In: Yao, X., et al. (eds.) PPSN 2004. LNCS, vol. 3242, pp. 1081–1091. Springer, Heidelberg (2004). doi: 10.1007/978-3-540-30217-9_109
8. Schonenberg, H., Mans, R., Russell, N., Mulyar, N., van der Aalst, W.M.: Towards a taxonomy of process flexibility. CAiSE Forum **344**, 81–84 (2008)
9. Regev, G., Soffer, P., Schmidt, R.: Taxonomy of flexibility in business processes. In: BPMDS (2006)
10. Reichert, M., Weber, B.: Enabling Flexibility in Process-Aware Information Systems: Challenges, Methods. Technologies. Springer, Heidelberg (2012)
11. Kasi, V., Tang, X.: Design attributes and performance outcomes: a framework for comparing business processes. In: Proceedings of the Eighth Annual Conference of the Southern Association of Information Systems (SAIS). pp. 226–232 (2005)
12. Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A.: Fundamentals of Business Process Management. Springer, Heidelberg (2013)
13. Jansen-Vullers, M., Loosschilder, M., Kleingeld, P., Reijers, H.: Performance measures to evaluate the impact of best practices. In: Proceedings of Workshops and Doctoral Consortium of the 19th International Conference on Advanced Information Systems Engineering (BPMDS Workshop), vol. 1, pp. 359–368. Tapir Academic Press, Trondheim (2007)
14. Melcher, J.: Process Measurement in Business Process Management: Theoretical Framework and Analysis of Several Aspects. KIT Scientific Publishing, Karlsruhe (2014)
15. van der Aalst, W.M., Berens, P.: Beyond workflow management: product-driven case handling. In: Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work. pp. 42–51. ACM (2001)
16. Li, C., Reichert, M., Wombacher, A.: A heuristic approach for discovering reference models by mining process model variants (2009)

17. Li, C., Reichert, M., Wombacher, A.: Discovering reference models by mining process variants using a heuristic approach. In: Dayal, U., Eder, J., Koehler, J., Reijers, Hajo, A. (eds.) BPM 2009. LNCS, vol. 5701, pp. 344–362. Springer, Heidelberg (2009). doi: 10.1007/978-3-642-03848-8_23
18. Li, C.: Mining process model variants: challenges, techniques, examples. University of Twente (2010)
19. Ayachi Ghannouchi, S., Ghannouchi, S.E.: Une expérience de bpr dans un hôpital tunisien. Systèmes d'information & management **13**(1), 89–116 (2008)
20. Ouni, A.B.: Plan blanc et gestion des crises sanitaires aux urgences. Ph.D. thesis, Faculté de médecine Sousse, Tunisie (2012)

Specification of an Architecture for Self-organizing Scheduling Systems

Ana Madureira(✉), Ivo Pereira, and Bruno Cunha

GECAD - Knowledge Engineering and Decision Support Research Center,
Institute of Engineering – Polytechnic of Porto (ISEP/IPP), Porto, Portugal
{amd,iaspe,bmaca}@isep.ipp.pt

Abstract. This paper presents the specification of an architecture for self-organizing scheduling systems. The proposed architecture uses learning by observing the experts and interpretation of scheduling experience. The design of intelligent systems that learn with experts is a very hard and challenging domain because current systems are becoming more and more complex and subject to rapid changes. In this work, different areas as Intelligent and Adaptive Human-Machine Interfaces, Metacognition and Learning from Observation, Self-managed Systems, amongst others, are joint together resulting in a global fully integrated architecture for self-organizing scheduling systems.

Keywords: Scheduling · Self-organizing scheduling systems · Learning from observation · User modeling

1 Introduction

Manufacturing planning and control address decisions on the acquisition, utilization and allocation of scarce manufacturing resources to satisfy customer needs and specifications in an efficient and effective way [1].

The scheduling problem can be seen as a decision making process where a diversity of characteristics and constraints related with tasks and manufacturing system resources can affect scheduling decisions. Resources include production facilities, labor and materials. Constraints include the availability of resources, due dates for the products, etc.

A challenge is emerging in the design of scheduling support systems for manufacturing environments where dynamic adaptation and optimization become increasingly important incorporating expert's knowledge. Despite the numerous advances in Human-Computer Interaction and Multi-Agent Systems [2–4], most development systems still require that users have substantial knowledge of procedural-programming techniques as well as the specific computer system at hand.

The motivation of this work arises from the need of developing approaches that should be able to adaptively control, coordinate and optimize, addressing the several challenges presented on complex real-world problems. By using advanced algorithms, modeling and services, it is intended that deciders make the right decision when and where it is crucial to the business. It is considered that the next generation of Decision

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_76

Support Systems (DSS) will rely on collaborative intelligent Multi-Agent Systems for playing a fundamental role in actively searching and finding relevant information on behalf of their users in complex and open environments, such as the Web.

This paper describes the architecture of an Adaptive Decision Support System for Interactive Scheduling with Metacognition and User Modeling Experience (ADSyS). The architecture presented in this paper intends to overcome some of the limitations of DSS for Manufacturing Scheduling, by improving plan generation and execution from experience based on learning considering: (i) intelligent proactive interaction between the user and the system, which should support the users to access, manage, share and exchange information; (ii) active learning ability to learn from users and to infer their intentions. To achieve this goal, a user modeling was performed; (iii) adaptive optimization methods incorporating knowledge by user observation and reacting to perturbations from external environment.

The remaining sections are organized as follows: Sect. 2 gives context of background and related work on this topic; Sect. 3 presents the specification of the proposed architecture, describing each module; Sect. 4 presents some conclusions and future work.

2 Background and Related Work

Scheduling is an important aspect of automation in manufacturing organizations. Most of scheduling domains are characterized by a great amount of uncertainty that leads to significant system dynamics [1]. Such dynamic scheduling is receiving increasing attention amongst both researchers and practitioners [5–7]. However, scheduling has still difficulties in real world situations and, hence, human intervention is required to maintain real-time adaptation and optimization.

At this point, for the specification of an architecture for self-organizing scheduling systems, we identify as relevant the following areas: Intelligent and Adaptive Human-Machine Interfaces, Metacognition and Learning from Observation, Self-managed Systems, amongst others.

2.1 Intelligent and Adaptive Human-Machine Interfaces

User interactive computer systems consist of two major elements: the computational (or functional) part and the communicational part (user interface) [8]. The first one is invisible to users and is the part that actually "does the job", while the latter is responsible for communicating with humans. It processes their actions, triggers the appropriate system functions and provides users with informative feedback.

The role of user interfaces has not always been considered a major aspect to take into account when designing and developing interactive systems. Studies in the field of Human-Computer Interaction were incipient or nonexistent and consequently the development of the user interface obeyed purely to constructional criteria, which, in general, do not match the needs of who will actually use the system: the user.

However, the quality of the user interface is crucial to the success of computer applications [8]: for users, the interface is the system, i.e. constitutes the application's

face; communication with the system is at least as important as the computation; nowadays, the costs associated with the use of interactive computer systems are due less to hardware and software purchasing, and more to person ware.

In conclusion, the user should not have to adapt to the interface. On the contrary, the interface must be designed in a way that its use is easy and natural to the user. Moreover, the constructional principles that determine the development of the interface software should not interfere in any way with the interface design, which should be governed exclusively by behavioral criteria. The architecture proposed in this work takes this into account give a contribution to the literature.

2.2 Learning from Observation and Metacognition

Learning is the process of adapting behavior in response to events in the environment. Without learning, a system is limited by the ability of its designer to foresee all situations that might occur.

Learning from Observation [9] is a common and natural method of human learning. Learning from Observation refers to computer-based system learning from observations obtained from user demonstrations of a task. Some systems go further than imitation strategies and attempt to interpret the intent of users. The concept of inferring user intent is relatively new and few systems have incorporated the idea of Learning from Observation during the design process. The main motivation for incorporating Learning from Observation in dynamic scheduling is to find patterns that could improve the overall scheduling performance.

Learning from Observation needs the support of an efficient Human-Computer Interaction [8]. Mixed-initiative refer to methods that explicitly support an efficient, natural interleaving of contributions by users and automated services aimed at converging on solutions to problems [10]. Mixed-initiative approaches promise to enhance human-computer interaction by allowing computers to behave more like associates, able to work with users to develop a shared understanding of goals and to contribute to problem solving in the most appropriate way.

In this work we attempt to imitate this species behavior and incorporate Learning from Observation and Metacognition [11] into a Multi-Agent Scheduling System in order to improve the overall scheduling performance. The analogy with human vision will be incorporated in the system through an intelligent and interactive User Interface. We consider that human and computer intelligence must be combined to effectively solve the scheduling problem.

2.3 Self-managed Systems

Self–managed systems have the ability to manage them-selves and to dynamically adapt to change in accordance with evolving or dynamic business policies and objectives, allowing the addition and removal of resources/tasks with-out service disruption. This field of research has received much attention in Autonomic Computing paradigm [12].

The state of art of Autonomic Computing is to create self-managing computer systems that can configure, heal, optimize and protect themselves without the need for human

intervention. Autonomic Computing is based on four self-* properties, in which research efforts may be categorized: Self-Configuring, Self-Healing, Self-Optimizing and Self-Protecting. This concept arises as the foundation for an efficient computing environment where efficiency is measured in terms of utilization, flexibility and scalability [12].

3 ADSyS System Architecture

ADSyS is a Scheduling System in which communities of agents model a real manufacturing system subject to perturbations. Agents must be able to learn and manage their internal behavior and their relationships with other agents, by collaboration in accordance with business targets.

The proposed architecture considered in this paper consists of four main modules: the Integrated Intelligent Human-Machine Interface, the User Modeling module, the

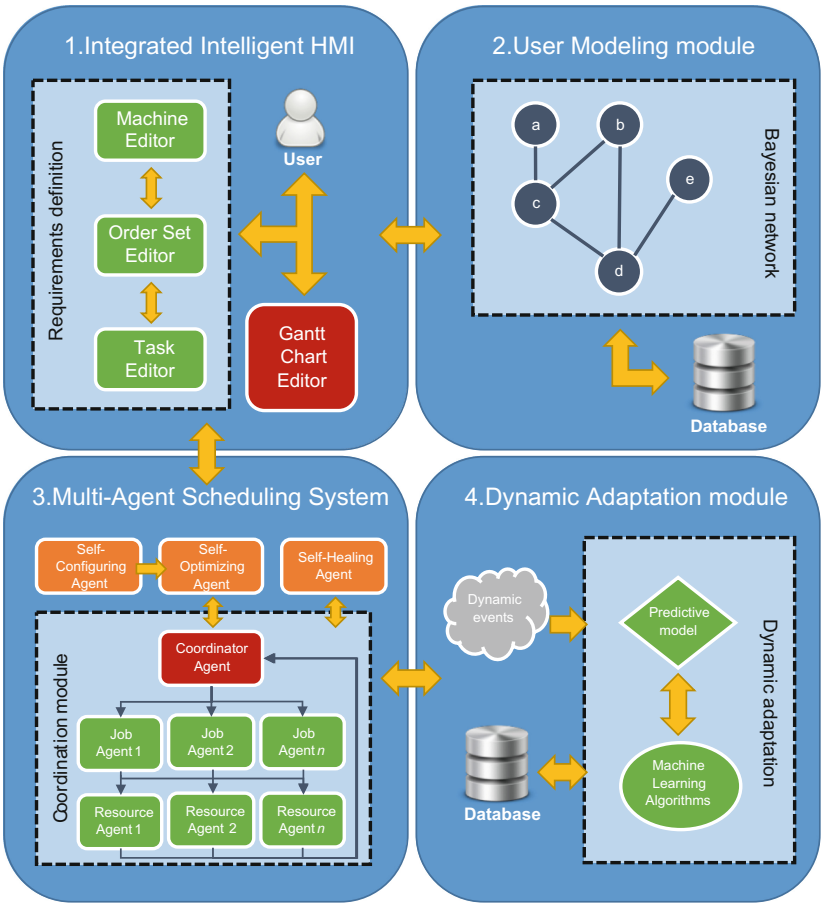


Fig. 1. ADSyS global architecture.

Multi-Agent Scheduling System, and the Dynamic Adaptation module. The system's global architecture is illustrated in Fig. 1. The information flow between each system component is represented by arrows.

3.1 Integrated Intelligent Human-Machine Interface

The Integrated Intelligent Human-Machine Interface module allows the interaction between the user and the scheduling system. The user can define the requirements for the scheduling problem, by inputting data information into the system. The results visualization in Gantt charts and reports summary gives the user the possibility of interaction with Gantt charts, allowing the insertion of changes to scheduling plans with the performance improvement purpose and dynamic events incorporation.

The proposed interaction model supports Problem Definition, Optimization Techniques parameterization and Results Visualization. It includes the multi-view functionality, which allows the user to view/interact with the different interaction objects simultaneously (Fig. 2). This functionality represents an asset because in addition of being a technological innovation, it provides a mechanism to compare and validate different interaction objects. Since any interaction objects can be viewed on the different layouts, this provides the ADSyS system a high degree of flexibility and information.

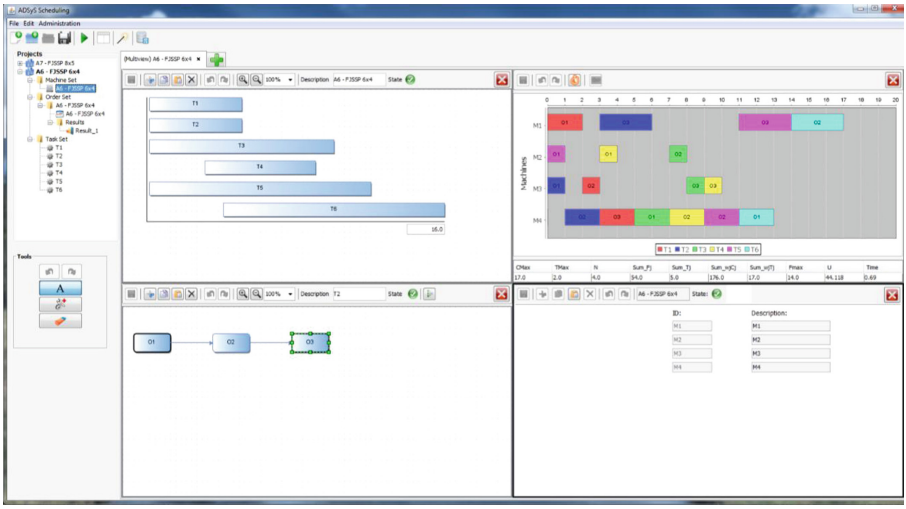


Fig. 2. Integrated intelligent human-machine interface: multi-view functionality.

Additionally, four interaction modules are considered: Task Editor, Machines Editor, Order Set Editor, and Gantt Chart Editor. They are responsible for the input data and requirements definition, i.e., for the specification of the scheduling problem and the visualization of scheduling results.

The Machine Editor purpose is to delineate the machines that are available during the operation definition. A machine is where the operations are processed. The number

of machines is not restricted and they can be inserted and not utilized later, in the final plan. Machines are characterized by their identification and a description. The user is able to create, remove or modify any number of machines using this component. When the user finishes editing, the system provides feedback about some information that might be missing (e.g. missing description on one machine).

The Task Editor allows the user to create a new task or edit an existing one. An operation is the basic unit of a task, since a task is defined as a sequence of operations. It is provided the possibility for the user to view and edit a task routing. A routing consists of operations and precedence. The user defines the sequence in which operations must be performed. This sequence defines the order in which the operations should be executed. An operation may have another operation or a set of operations as precedent. Using this editor, the user is able to insert or remove operations and to define their sequence, amongst other options (e.g. copy previous operations and graph auto-alignment).

The Order Set Editor allows to create a new order set or edit an existing one. A task is the basic unit for the definition of an order set. The Order Set Editor function is to define and edit the problem, taking advantage of a graphical interactive tool.

The results of the scheduling module are provided to the user via a Gantt chart. The results are included in a tool that allows the user to see and edit the operations in the scheduling plan. The interaction gives the user the option to tune the plan reflecting his needs. The user can alter the plan by performing horizontal operation movements. The lateral transfers imply a modification of the position of an operation in the given machine, allowing the delaying or anticipation of an operation. When a movement is performed, the system does two validations: operation overlapping – two operations are not able to be executed at the same time by one machine - and task constraints – an operation is only executed after the task release date and once the corresponding (if any) preceding operation is finished.

More details about the human-machine interface can be found in [13].

3.2 User Modeling Module

The proposed User Modeling module is responsible for modeling users with the objective of not only increasing the learning rate of new users but also improving the productivity of the more experienced ones. It consists in multiple components that cooperate in order to characterize and analyze user behavior, using the information gathered to provide a customized system experience, using proper helps and tips according to the type of user.

Although its foundation is laid on the collaboration of several smaller components, the most relevant constituents of the User Modelling module are the user classifier and the information capture unit. The information unit is responsible for detecting, capturing and understanding every interaction that the user has with the system. Some examples of the mentioned usage are the removal of planned tasks or going back and forth between scheduling operations (meaning a high degree of indecision). The user classifier is the mathematical structure responsible for attributing the user to one of three levels of expertise: beginner, intermediate or expert. The process is quit straightforward: the user

captured information is fed to the user classifier and, in return, the level of expertise is the obtained output.

More details about the proposed User Modeling module (including how to prepare the user classifier) can be found in [14].

3.3 Multi-agent Scheduling System

In the proposed Multi-Agent Scheduling System (Fig. 3) there are agents that represent tasks/jobs (Task Agents) and agents representing machines/resources (Resource Agents) in an industrial environment.

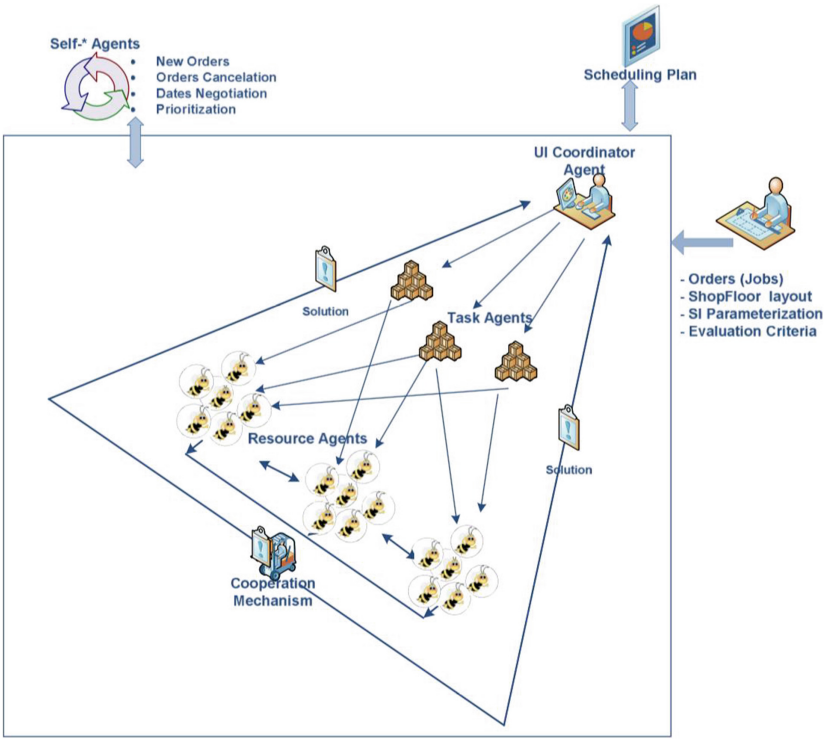


Fig. 3. ADSyS multi-agent scheduling system.

The proposed model also considers a Coordinator Agent responsible for coordinating individual solutions obtained by each Resource Agent in order to obtain a global scheduling for the original problem. The ADSyS Multi-Agent System architecture is similar to AutodynAgents system [16], but the agents were enriched with additional behaviors, including learning from observation and experience.

The Coordinator Agent is responsible for a permanent and fluid communication with the user. Besides being connected to the human-machine interface, this agent dynamically generates Task Agents, according to the number of tasks that comprise the

scheduling problem. In the end, collects the solutions of each Resource Agent and validate these solutions applying a repair mechanism and a coordination mechanism (cooperation or negotiation) if necessary.

Task Agents, also known as Job Agents, are responsible for processing the necessary information about each task, i.e. preprocessing the scheduling problem data. This is the generation of the first and last processing times and allocation of machine operations to the respective Resource Agents.

Resource Agents are responsible for scheduling operations that require processing on the machine that they control. They implement intelligent optimization methods (Metaheuristics) such as Genetic Algorithms, Particle Swarm Optimization, Ant Colony Optimization, among others, in order to find the best combination of operations and then coordinates through cooperation/negotiation to improve the plans.

The model presented in Fig. 3 is able of self-management through the implementation of self-* agents: Self-Configuring Agent, Self-Optimizing Agent, and Self-Healing Agent.

The Self-Configuration Agent is responsible for monitoring the system to detect changes that affect the scheduling, allowing the system to dynamically adapt himself. It is responsible for coordinating the Dynamic Adaptation module.

The Self-Optimizing Agent is responsible for tuning the parameters of Metaheuristics, according to the problem the schedule. This agent detects the initial problem characteristics (number of tasks/machines, their precedences, operations and attributes, etc.) or a change detected by the Self-Configuration Agent, and automatically chooses the Metaheuristic to use, creating the self-parameterization. If there is a dynamic event, the parameters are changed at runtime. This tuning of the parameters is done through learning and experience, since it uses an hybrid module with Racing techniques and Case-based Reasoning (CBR). Each time a new problem (case) arises, the CBR uses past experience to define the Metaheuristic and parameters to use. If a new case was resolved, it is saved for future use [15].

The Self-Healing Agent gives the system the ability to detect deviations from normal conditions and proactively take steps to normalize, avoiding service disruptions. Since agents can fail for various reasons, this agent provides one or more backup states to secure storage for the reactivation of those who were lost or ingrown and have significant results, allowing the system to restart a previous point (and not from the start). With this agent, the system becomes stable, able to deal with any problems that may occur.

More details about this Multi-agent Scheduling System can be found in [16].

3.4 Dynamic Adaptation Module

To enhance the scheduling of manufacturing orders in dynamic environments, a self-organized integration mechanism (IM) module based on supervised learning techniques was used in order to allow scheduling systems to decide autonomously which IM should be used to incorporate new orders in the current scheduling plan.

The Dynamic Adaptation module includes mechanisms for incorporating urgent tasks in the current scheduling plan. This module follows a supervised learning approach. In this way the system is able to proactively decide which IM to use.

The main aim of this machine learning approach is to create a classification model, where the input resources are mapped from the input variables to the output variable by a supervised classification algorithm. Figure 4 illustrates the proposed machine learning approach.

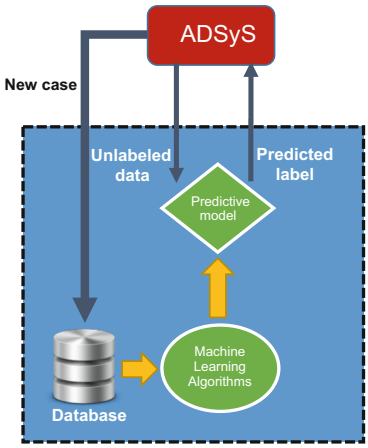


Fig. 4. ADSyS dynamic adaptation module

This approach has three major components: a database, a machine learning algorithm and a predictive model. Any intelligent system requires a database/knowledge base to be trained and tested. The database contains the training data previously collected. These training data is then used as input to the training of machine learning algorithm that will produce the classifier. This classifier has the responsibility to provide the best IM to use in each problem to solve. A problem may be characterized by data relating to the current scheduling level and also information about the dynamic task.

More details about the proposed Dynamic Adaptation module can be found in [17].

4 Conclusions and Future Work

The specification of an architecture for self-organizing systems applied to the resolution of the scheduling problem was presented in this paper. The scheduling problem can be seen as a decision making process where a diversity of characteristics and constraints related with tasks and manufacturing system resources can affect scheduling decisions, and it is still considered a critical optimization problem.

The proposed architecture was applied in the ADSyS project and uses learning by observing the expert's and interpretation of scheduling experience. In this work, different areas as Intelligent and Adaptive Human-Machine Interfaces, Metacognition and Learning from Observation, Self-managed Systems, amongst others, were joint together resulting in a global fully integrated architecture for self-organizing scheduling systems. Four main modules were described: an Integrated Intelligent Human-Machine Interface,

a User Modeling module, a Multi-Agent Scheduling System, and a Dynamic Adaptation module.

Future work include application of this architecture to real-world problems, with possible partnerships with industry.

Acknowledgments. This work is supported by FEDER Funds through the "Programa Operacional Factores de Competitividade - COMPETE" program and by National Funds through FCT "Fundação para a Ciência e a Tecnologia" under the project: FCOMP-01-0124-FEDER-PEst-OE/EEI/UI0760/2014.

References

1. Pinedo, M.: Scheduling: Theory, Algorithms, and Systems, 4th edn. Springer, Heidelberg (2012)
2. Boeres, C., Lima, A., Vinod, E., Rebello, F.: Hybrid task scheduling: integrating static and dynamic heuristics. In: 15th Symposium on Computer Architecture and High Performance Computing, vol. 199 (2003)
3. Monostori, L., Vncza, J., Kumara, S.: Agent based systems for manufacturing. CIRP Ann. Manuf. Technol. **55**, 697–720 (2006)
4. Madureira, A., Pereira, I., Pereira, P., Abraham, A.: Negotiation mechanism for self-organized scheduling system with collective intelligence. Neurocomputing **132**, 97–110 (2014). Elsevier
5. Aytug, H., Lawley, M.A., McKay, K., Uzsoy, R.: Executing production schedules in the face of uncertainties: a review and some future directions. Eur. J. Oper. Res. **16**(1), 86–100 (2005)
6. Blazewicz, J., Ecker, K.H., Trystram, D.: Recent advances in scheduling in computer and manufacturing systems. Eur. J. Oper. Res. **164**(3), 639–668 (2005)
7. Xhafa, F., Abraham, A. (eds.): Metaheuristics for Scheduling in Industrial and Manufacturing Applications. Studies in Computational Intelligence, vol. 128. Springer, Heidelberg (2008)
8. Nielsen, J.: Usability Engineering. AP Professional, Boston (2003)
9. Michalski, R.S., Carbonell, J., Mitchell, T.: Machine Learning: An Artificial Intelligence Approach. Springer, New York (2013)
10. Ferguson, G., Allen, J.: Mixed-initiative dialogue systems for collaborative problem-solving. AI Mag. **28**(2), 23–32 (2007). Special Issue on Mixed-Initiative Assistants
11. Larson, C.B.: Metacognition New Research Developments. Nova Science Publishers Inc., New York (2009). ISBN: 9781606927809
12. Kephart, J., Chess, D.: The vision of autonomic computing. Computer **36**, 41–50 (2003)
13. Piairo, J., Madureira, A., Pereira, J.P., Pereira, I.: A user-centered interface for scheduling problem definition. In: Rocha, Á., Correia, A., Wilson, T., Stroetmann, K. (eds.) Advances in Information Systems and Technologies, vol. 206. Springer, Heidelberg (2013)
14. Cunha, B., Madureira, A., Pereira, J.P.: User modelling in scheduling system with artificial neural networks. In: 10th IEEE Iberian Conference on Information Systems and Technologies (2015)
15. Pereira, I.: Intelligent System for Scheduling Assisted by Learning, Ph.D. Thesis in Electrical and Computer Engineering, Department of Electrical and Computer Engineering, University of Trás-os-Montes and Alto Douro (2014). (in Portuguese)
16. Madureira, A., Pereira, I.: Intelligent bio-inspired system for manufacturing scheduling under uncertainties. In: 10th International Conference on Hybrid Intelligent Systems (2010)
17. Madureira, A., Santos, J.M., Gomes, S., Cunha, B., Pereira, J. P., Pereira, I.: Manufacturing rush orders rescheduling: a supervised learning approach. In: Sixth World Congress on Nature and Biologically Inspired Computing, pp. 299–304 (2014)

Towards Better SWRL Rules Dependency Extraction

Abir Boujelben[1(✉)], Tarak Chaari[2], and Ikram Amous[1]

[1] MIRACL Laboratory, Technological pole of Sfax, BP. 242, 3021 Sfax, Tunisia
abeer.boujelben@gmail.com, ikram.amous@isecs.rnu.tn
[2] ReDCAD Laboratory, National School of Engineers of Sfax,
BP 1173, 3038 Sfax, Tunisia
tarak.chaari@redcad.org

Abstract. Information systems knowledge bases often include inference rules. The continuous growth of the facts in the recent information systems environments has caused the exponential increase of rule bases sizes. Therefore, rule bases management becomes more and more difficult. Such a task should be automated and based on the extraction of dependencies between rules in order to have a better insight on their correct execution order and to detect conflicts between them. In this paper, we describe a rules dependency extraction approach for Semantic Web Rule Language (SWRL) rules. Our approach insures the automatic extraction of a rule dependency graph based on the semantics of their components. We evaluated our work by applying it to two different ontologies from medical and network security domains. We have implemented a prototype of our approach and we integrated it in a plug-in for Potégé-Owl editor.

1 Introduction

Currently we are overwhelmed with information. Business, scientists or even individuals are now forced to use tools that are based on knowledge management systems. The Semantic Web represents the next major evolution in connecting information for such systems. It enables data to be linked from a source to any other source and to be understood by computers so that they can perform sophisticated tasks on our behalf. A crucial component to the Semantic Web is the definition and use of ontologies. The Web Ontology Language (OWL) is an international standard for encoding and exchanging them [19]. The expressivity of this language can be extended by adding SWRL (Semantic Web Rule Language) rules to ontology [14]. Nowadays, rule sets associated to ontologies are getting more and more huger. As a result, they became difficult to comprehend and manage [12]. So many methods and tools [23,30,31] have been proposed to assist users to deal with these difficulties (e.g. Axiomé [9,10], SWRL Tab [21], SAP NetWeaver BRM [24]). Many of these methods and tools are based on the automatic detection of the dependency relations between the elements of a rule base. In this paper we aim to introduce a new method to accomplish such

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_77

a task. We address the problem for a set of SWRL rules. The main difference between proposed method and existent work consists in the fact that proposed idea allows a better use of semantics expressed by the ontology definition. It employs more its entities definitions and follows a new technique to extract more correct dependencies and less false ones. This is not widely recognized to be problematic in existent works. As far as we know, there is no proposed method evaluated to determine the number of correct dependencies among those extracted. The lack of available ontologies with associated rule sets with already known possible execution orders is a proof for our claim. The rest of this paper is organized as follows: Sect. 2 introduces OWL and SWRL languages in addition to some fundamentals that we used in our work. Section 3 presents related work. Section 4 describes our method and details its steps. Finally, Sect. 5 presents our experimental analysis and Sect. 6 provides a summary and an outlook for further research.

2 Background

OWL ontology is a set of axioms, which provide explicit logical assertions about three types of things: classes (Person, Food), individuals (classes' instances) and properties. There are two types of properties in OWL ontologies. Datatype-properties are binary relations that link a class to a piece of typed data (string, float, etc.). Object properties are binary relations that link a class to another. Knowledge about these entities can be enriched by adding to them some SWRL rules. Each of which is composed of two parts: an antecedent part (body) that implicates a consequent part (head). Each one of them is composed of positive conjunctions of atoms. $R0$ is an example of SWRL rules which indicates that a person whose age is greater than 17 is considered as adult.

$$R0 : Person(?p) \wedge hasAge(?p, ?age) \wedge swrlb : geaterThan(?age, 17) \rightarrow Adult(?p)$$

SWRL provides seven types of atoms. Each atom is composed of a predicate and one or more arguments. Rule bases are executed throw some piece of software called semantic reasoners. Some of these rules have to be executed after other ones as they need inferred facts. This means that latter executed rules depend on first ones. A rule dependency graph [8,15,16] can be constructed in order to determine the interdependency among the rules. It is a directed graph defined by $RDG = (V, E)$ [30] where (i) V is a set of vertices. There is one vertice for each rule in the rule base. (ii) E is a set of ordered pairs of vertices (called also arcs). An arc $e = (R1, R2)$ means that the rule represented by the vertice $R2$ depends on the rule represented by the vertice $R1$. Semantic links between classes (or words) are another notion used when dealing with semantic issues especially in Natural Language Processing (NLP). Different relationships are defined between words in natural language. Therefore, these relations are defined between classes in dictionaries (e.g. Wordnet [20], Wiktionary [7]) (see Table 1) [13,22,28,29] and they represent semantic relations which express upward, downward and horizontal directions [13]. Upward links express generalization (e.g. a wheel **is a**

Table 1. Semantic relations and expressed directions

Semantic relation	Example	Expressed direction
Synonymy (similar)	Sad/unhappy	Horizontal link
Antonymy (opposite)	Rapidly/slowly	Horizontal link
Hypernymy (superior)	Apple/fruit	Upward link
Hyponyymy (subordinate)	Animal/monkey	Downward link
Holonymy (hole)	Head/nose	Downward link
Meronymy (part)	Wheel/car	Upward link
Troponomy (manner)	Whisper/speak	Upward link
Entailment:	Drive/ride	
Cause		Downward link
Effect		Upward link

part of a vehicle) and downward links express specialization (a person **owns** a car). Table 1 shows links direction expressed by each type of relation.

3 Related Work

Recently, some work focused on the development of rule dependency analysis techniques [11] which are needed in front of the increasing use of rules in ontology-based systems. Dependency graphs are one of these techniques. They are directed graphs that give information about their state and reflect some of the knowledge they are modeling. There are three types of dependency graphs used while dealing with rule bases problems. The predicate dependency graph is defined by a set of vertices representing the predicates invoked in the rules, and a set of edges connecting them. Each edge indicates which predicate depends on another one. We cite also the position dependency graph. Its vertices represent pairs *(p, i)* with i is the position of a variable X of an atom p from rule R1's body. There is an edge from *(p, i)* to each position of X in R1's head [18,25]. Edges show how variables are propagated. The third type is the rule dependency graph mentioned in the previous section. As we aim in our future research to ensure an automatic maintenance of rule bases, in our work we employed this latter type of dependency graphs. Works based on this type of graphs have many purposes such as incoherence detection [11], ensuring a finite execution of queries [1,6], and generation of explanations [6]. There are two types of rules dependencies: (i) syntactic dependencies (or naïve dependencies), and (ii) semantic ones. The first type relies on the fact that an atom's predicate is referenced in the first rule's head and the second's body. The semantic dependency extends that by using the equivalence and the hierarchical relationships defined between the ontology entities. In literature, there are many methods and tools based on rules dependency graphs. [2] proposes a graphical tool to support graph-based rule creation and edition. [31] presents a tool, called S2REd, to assist rule base development. Both of them primarily aimed to facilitate rules creation, edition

and comprehension by representing each rule in a graph. In order to maintain rule bases, authors in [3] tried to represent them with a dependency graph according to dominance criteria. Authors in [17] founded also their approach to enhance the performance of the rule-based OWL reasoning on a rule dependency graph. These works built these graphs using only syntactic dependency. This drawback is addressed in Slider-p [5] which is a reasoning tool supported by generating a rules dependency graph using semantic dependency. However this tool considered only class atoms and object-property atoms to accomplish their missions. Employing the same technique, [9] describes a tool, called Axiomé, that helps in creating techniques to visualize, paraphrase, and categorize SWRL rules. It creates a rule dependency graph and offers visualization user interface for it. This work was improved [12] by using semantics expressed in the ontology between classes (classes hierarchy and object properties) involving atoms' arguments comparison. In addition to class atoms and object-property atoms, the authors used same individual and different individuals atoms. Existing works dealing with dependency extraction in SWRL rule bases analyzed only first rule's body and second rule's head atoms. Besides, they did not consider datatype property atoms or built-in atoms. Furthermore, they analyzed the dependencies between atoms having only the same type. We believe that all of these existing works do not take profit from the semantic richness of the ontology definition to deduce rule dependencies.

4 General Strategy

In this section we present DEx, our method for rules dependency extraction. It consists in producing a dependency graph for a given rule base elements. It is based on two major techniques while treating the dependency between two rules: (i) analyzing the dependency between first rule's head and second rule's body, and (ii) analyzing the dependency between two rules' heads. In each technique a weight is attributed according to the cause of the dependency. Then the strongest one is kept. A general overview of our method is presented in Algorithm 1 and details are provided by the further sub-sections.

Algorithm 1. DEx: Algorithm For **D**ependency **Ex**traction

Input: Σ {rule base}
Output: dep {dependency matrix}
1: **for** Ri$\in \Sigma$ **do**
2: **for** Rj$\in \Sigma$ **do**
3: dep[Ri][Rj]$\leftarrow$dependency from **Head**(Ri) to **Body**(Rj)
4: dep[Ri][Rj]$\leftarrow$ max(dep[Ri][Rj], dependency from **Head**(Ri) to **Head**(Rj))
5: **end for**
6: **end for**
7: **return** dep

4.1 Dependency Extraction by Head's and Body's Atoms Analysis

The first step in our method consists in generating the right dependency's direction between two rules using the first one's head and the second's body. If a first rule generates a fact concerning an OWL class and a second rule's body uses that fact, then the first rule must be executed before the second one. It means that the second rule depends on the first rule. Thus this step of our method is based on treating each couple of atoms (aH, aB) where aH is an atom from the first rule's head and aB is an atom from the second rule's body. So we are verifying if aB depends on aH. The analysis can extract a possible dependency between these two rules if aH and aB satisfy one of the following conditions: (i) **Condition 1:** if they have the same type, then we analyse their predicates and then their arguments: any entity referenced in aH must be equal or equivalent to or a sub-entity of the one referenced in aB (example 1); (ii) **Condition 2:** if they have different types, then we must find a class, referenced in aH, equal or equivalent to or a sub-class of a class referenced in aB (example 2).

Example 1: let be aH=hasLunch(?p1,?l1) and aB=hasFood(?p2,?f2). Here we have two object property atoms, so first we analyze the relation between the predicates. In this case aH refers to the property hasLunch, and aB refers to the property hasFood. hasLunch is a sub property of hasFood, so we analyze the classes referenced by the arguments' types. ?p1 refers to a Person and ?p2 refers to a Person (equality relation). In the other hand, ?l1 refers to the class Lunch and ?f2 refers to the class Food. In the ontology definition Lunch is a sub−class of Food. We conclude that atom aB depends on atom aH, so R2 depends on R1.

Example 2: let be aH=Person(?p1) and aB=hasAge(?p,18). Here we have atoms with different types. So we extract the classes referenced in aH (which are the only class Person) and the classes referenced in aB (which are also the only class Person). aH and aB refer to the same class, so aB depends on aH and R1 must be executed before R2.

In condition 1, we treat all atom types in the same way except for datatype property and built-in atoms. For datatype properties, besides the property referenced by the predicate and the class referenced by the first argument's type, we take into consideration the second argument whether it is a constant (e.g. hasAge(?p,?18)) or a variable (e.g. hasAge(?p,?a)). In this way we have more precise knowledge about facts needed for a rule's body and facts inferred by a rule's head. Besides, for a built-in atom, we take into consideration its input arguments' types when it is from a rule's body (e.g. if aB=swrlb:multiply(?rectangleArea,?w,?h) we consider the classes referenced by the arguments ?w and ?h which are the classes Width and Height). On the other hand, when it is a rule head atom we consider the modified entity's type (e.g. if aH=swrlb:multiply(?rectangleArea,?w,?h) we consider the class referenced by the argument ?rectangleArea which is the class Area). If it is a custom built-in, the modified entity is provided by the ontology designer. In this step of our method, we assign a weight for each cause of dependency. In case of Condition 1, when atoms predicates reference equal or equivalent entities and their

arguments reference equal or equivalent classes we assign to this dependency an "Extra Strong" weight. This latter would be "Strong" if atoms' arguments referenced classes having a hierarchical relationship. We note that possible assigned weights are "Extra Strong", "Very Strong" and "Strong". Cases where the dependency cause is judged to be weak are not considered as they extract false dependencies. In the same way, in Condition 2 we assign "Extra Strong", "Very Strong" and "Strong" weights. A weight raises when classes referenced in aH and aB are equal or equivalent. It would be weaker when they have hierarchical relationships.

Every time we analyse the dependency between two rules in this step of our method, we keep the strongest weight found.

4.2 Dependency Extraction by Heads' Atoms Analysis

The second step of our method extracts a possible dependency between two rules by analysing their heads' atoms. We propose this new technique to (i) extract more correct dependencies, (ii) to choose the right dependency's direction and (iii) to eliminate wrong dependencies. The main idea to accomplish this mission is based on the fact that first inferred facts should be about the more general or the more elementary classes. Therefore, this step of our method combines the type of the process used for rule base creation (top-down or bottom-up process) with directions of the semantic relations expressed by object-properties names (upward link or downward link).

Dependencies extracted in this step are weighted "Extra Strong" as they are based on following directions of knowledge flows in the ontology definition and the rule base. In the first hand, top-down process (resp. bottom-up process) means that, during rules execution, first generated facts concern individuals of the more general OWL classes (resp. the more elementary OWL classes) and the latter generated facts concern individuals of the more elementary OWL classes (resp. the more general OWL classes). In the other hand, classes are connected by object-properties in the ontology definition. And an object-property name expressing a downward-link (resp. upward-link) shows that this property's domain references an OWL class more general (resp. more elementary) than the OWL class referenced by its range. For example, belongsTo(Player, Team) is a defined OWL object property. Its name expresses a meronymy relation from class Player to class Team. So it expresses an upward link from class Player to class Team. Keeping these two notions in mind, we do the correspondence between top-down process and downward-links expressed by object-properties names; and bottom-up process and upward-links expressed by object-properties names. As example, we consider that the process used to create a rule base σ is a top-down process, so we have to follow downward links. Let be R1 and R2 two rules from σ.

$$R1 : Player(?p) \wedge hasNewPoints(?p, ?pts) \wedge hasTeam(?p, ?t) \wedge hasScore(?t, ?s) \wedge swrlb : add(?ns, ?s, ?pts) \rightarrow hasScore(?t, ?ns)$$

$R2 : Tournament(?r) \wedge isFinished(?r, true) \rightarrow FinishedTournament(?r) \wedge myBI : windUp(?r)$

Team is a class referenced in R1's head and class FinishedTournament, a subclass of Tournament, is referenced in R2's head. includes (Tournament, Team) is an OWL object property defined in the associated ontology. It expresses a holonymy relation. It means that it represents a downward-link from class Tournament to class Team. We conclude that R1 must be executed before R2.

To detect the direction expressed by an object property, we built a predicate corpus for this type of properties. We defined two sets of properties predicates. The first set includes predicates expressing upward links and the second one includes predicates that express a downward link. We based it on ConceptNet5 corpus [27] which is a multilingual knowledge base. It represents words and phrases that people use and the common-sense relationships between them. The knowledge in ConceptNet is collected from a variety of resources such as Wiktionary, Open Mind Common Sense [26] and WordNet.

5 Experimentation and Discussions

To check our approach feasibility, we integrated our work in a tool called Axiomé [9]. It is provided as a plugin for Protégé-OWL[1] the most widely used ontology modeling software for the Semantic Web. As mentioned before, Axiomé extracts rules dependencies using semantic dependency analysis. In this section, we present the evaluation of our method on two different ontologies. Each one of them is associated to a set of rules that we already know their possible execution orders.

Security Policy Matching ontology (SPM) [4] is developed in the framework of a thesis in the ReDCAD laboratory[2]. It defines a set of semantic relations that can exist between the provider and requestor security concepts in web services. Its rule set contains 79 elements developed in a bottom-up process.

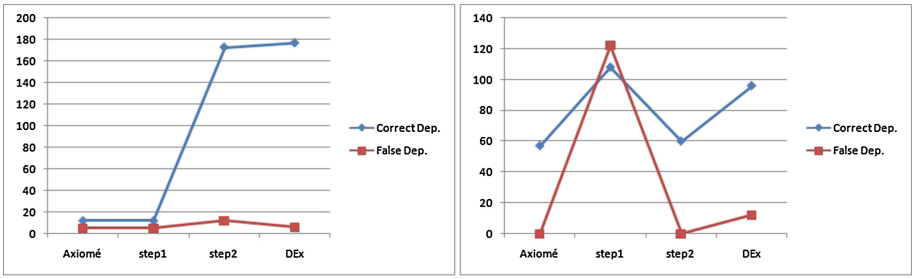
Diabete Food Ontology (DFO) is designed and developed in collaboration with domain experts. It is an ontology of nutrition for diabetic patients. It has 24 rules developed in a top-down process.

The results are given in Table 2. In both tests, our method outperforms the number of correct dependencies (78% of dependencies expected to be extracted in SPM case, and 89% in DFO case) with a high precision (about 90% in both cases). The first step of our method presents an extension for the existent techniques proposed for rules dependency extraction by (i) analyzing datatype property and built-in atoms, (ii) and analyzing the dependency between different types of atoms. This contribution has shown its performance in the DFO case as it highly raised the recall value compared to Axiomé. Moreover, the analysis of two rules' heads which represents the second step of our algorithm, improves

[1] http://protege.stanford.edu/.
[2] http://www.redcad.org/.

Table 2. Application of Axiomé and DEx

	SPM case				DFO case			
	Axiomé	DEx	DEx's 1st step	DEx's 2nd step	Axiomé	DEx	DEx's 1st step	DEx's 2nd step
Recall	0.533	0.78	0.53	0.77	0.54	0.89	1	0.55
Precision	0.7	0.97	0.7	0.93	1	0.88	0.45	1



(a) SPM use case : Axiomé and DEx application results

(b) DFO use case : Axiomé and DEx application results

Fig. 1. Evaluation results

its efficiency. According to the recall values, it is obvious that it has extracted a good number of correct dependencies and emended the direction of some ones. Also, we note that both steps of our algorithm are complementary. In SPM case, the second step has exponentially raised the number of correct dependencies while the first step has lowered the number of false ones (see Fig. 1a). Besides, in DFO's case, their roles have been switched as the first step has pulled up the number of correct dependencies and the second step has pulled down the number of false ones (see Fig. 1b).

6 Conclusions

In this paper we presented a novel method for extracting dependencies between elements of a SWRL rule base. Our work relies on the semantics expressed by the ontology definition in opposition to the existing work which was mainly focused on the syntactic definition of the rules. Experimentation results with two example ontologies from two different domains are very promising. For our future work we will improve our method by eliminating the possible incoherences (cyclic redundancy, conflicts, etc.) from the obtained graph. We plan to extend this method to be as closer as possible to an automatic management of a SWRL rule base.

References

1. Baget, J.-F., Mugnier, M.-L., Thomazo, M.: Towards farsighted dependencies for existential rules. In: Rudolph, S., Gutierrez, C. (eds.) RR 2011. LNCS, vol. 6902, pp. 30–45. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23580-1_4
2. Bak, J., Nowak, M., Jedrzejek, C.: Graph-based editor for SWRL rule bases. In: RuleML (2). Citeseer (2013)
3. Bouker, S., Saidi, R., Yahia, S.B., Nguifo, E.M.: Ranking and selecting association rules based on dominance relationship. In: 2012 IEEE 24th International Conference on Tools With Artificial Intelligence, vol. 1, pp. 658–665. IEEE (2012)
4. Brahim, M.B., Chaari, T., Jemaa, M.B., Jmaiel, M.: Semantic matching of web services security policies. In: 2012 7th International Conference on Risks and Security of Internet and Systems (CRiSIS), pp. 1–8. IEEE (2012)
5. Chevalier, J., Subercaze, J., Gravier, C., Laforest, F.: Incremental and directed rule-based inference on RDFS. In: Hartmann, S., Ma, H. (eds.) DEXA 2016. LNCS, vol. 9828, pp. 287–294. Springer, Heidelberg (2016). doi:10.1007/978-3-319-44406-2_22
6. Erdem, E., Erdem, Y., Erdogan, H., Öztok, U.: Finding answers and generating explanations for complex biomedical queries. In: AAAI (2011)
7. Fuertes-Olivera, P.A.: The function theory of lexicography and electronic dictionaries: wiktionary as a prototype of collective free multiple-language internet dictionary. In: Bergenholtz, H., Nielsen, S., Tarp, S. (eds.) Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow, pp. 99–134. Peter Lang, Bern (2009). ISBN: 978-3-03911-799-4
8. Gantayat, N., Das, R., Cherukuri, S.C.: Automated methodology comprised of supervised techniques to assist product selection. In: Recent Advances and Innovations in Engineering (ICRAIE), pp. 1–6. IEEE (2014)
9. Hassanpour, S., O'Connor, M., Das, A.: Axiomé: a tool for the elicitation and management of SWRL rules. In: Proceedings of the 6th International Conference on OWL: Experiences and Directions, vol. 529, pp. 204–207. CEUR-WS.org (2009)
10. Hassanpour, S., O'Connor, M.J., Das, A.K.: Exploration of SWRL rule bases through visualization, paraphrasing, and categorization of rules. In: Governatori, G., Hall, J., Paschke, A. (eds.) RuleML 2009. LNCS, vol. 5858, pp. 246–261. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04985-9_23
11. Hassanpour, S., O'Connor, M.J., Das, A.K.: Visualizing logical dependencies in SWRL rule bases. In: Dean, M., Hall, J., Rotolo, A., Tabet, S. (eds.) RuleML 2010. LNCS, vol. 6403, pp. 259–272. Springer, Heidelberg (2010). doi:10.1007/978-3-642-16289-3_22
12. Hassanpour, S., O'Connor, M.J., Das, A.K.: Clustering rule bases using ontology-based similarity measures. Web Semant. Sci. Serv. Agents World Wide Web **25**, 1–8 (2014)
13. Hirst, G., St-Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, vol. 305, pp. 305–332. MIT Press, Cambridge (1998)
14. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M., et al.: SWRL: a semantic web rule language combining OWL and ruleml. In: W3C Member submission, vol. 21, p. 79 (2004)
15. Katta, N., Alipourfard, O., Rexford, J., Walker, D.: Cacheflow: dependency-aware rule-caching for software-defined networks. In: Proceedings of the ACM Symposium on SDN Research (SOSR) (2016)

16. Krötzsch, M., Rudolph, S.: On the relationship of joint acyclicity and super-weak acyclicity. Technical report, Tech. rep. 3037, Institute AIFB, Karlsruhe Institute of Technology (2013). http://www.aifb.kit.edu/web/Techreport3013
17. Liu, Z., Feng, Z., Zhang, X., Wang, X., Rao, G.: RORS: enhanced rule-based owl reasoning on spark. arXiv preprint arXiv:1605.02824 (2016)
18. Lukasiewicz, T., Martinez, M.V., Simari, G.I.: Complexity of inconsistency-tolerant query answering in datalog+/–. In: Meersman, R., Panetto, H., Dillon, T., Eder, J., Bellahsene, Z., Ritter, N., Leenheer, P., Dou, D. (eds.) OTM 2013. LNCS, vol. 8185, pp. 488–500. Springer, Heidelberg (2013). doi:10.1007/978-3-642-41030-7_35
19. McGuinness, D.L., Van Harmelen, F., et al.: OWL web ontology language overview. In: W3C recommendation, vol. 10(10), p. 2004 (2004)
20. Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
21. O'Connor, M.: Swrltab (2008)
22. Panchenko, A., et al.: Similarity measures for semantic relation extraction. Ph.D. thesis, UCL (2013)
23. Rivolli, A., Orlando, J.P., Moreira, D.A.: An analysis of rules-based systems to improve swrl tools. In: Proceedings of the ICEIS (4), pp. 191–194 (2011)
24. Sap, A.: See your business clearly. SAP BusinessObjects Dashboards. http://www.sap.com/uk/solutions/sapbusinessobjects/large/business-intelligence/dashboards/sapbusinessobjects-dashboards/index.epx
25. Seipel, D.: Knowledge engineering for hybrid deductive databases. In: 29nd Workshop on (Constraint) Logic Programming (WLP 2015), p. 66 (2015)
26. Singh, P., Lin, T., Mueller, E.T., Lim, G., Perkins, T., Zhu, W.L.: Open mind common sense: knowledge acquisition from the general public. In: Meersman, R., Tari, Z. (eds.) OTM 2002. LNCS, vol. 2519, pp. 1223–1237. Springer, Heidelberg (2002). doi:10.1007/3-540-36124-3_77
27. Speer, R., Havasi, C.: Conceptnet 5: a large semantic network for relational knowledge. In: Gurevych, I., Kim, J. (eds.) The People's Web Meets NLP, pp. 161–176. Springer, Heidelberg (2013)
28. Vossen, P.: From wordnet to eurowordnet to the global wordnet grid: anchoring languages to universal meaning. Guest lecture, Language Engineering Applications, 26 February 2009
29. Yang, F., Xing, Y., Sun, H., Sun, T., Chen, S.: An ontology-based semantic similarity measure considering multi-inheritance in biomedicine. Math. Probl. Eng. **2015**, 1–9 (2015)
30. Zacharias, V., Borgi, I.: Exploiting usage data for the visualization of rule bases. In: Proceedings of the 3rd International Semantic Web User Interaction Workshop SWUI. Citeseer (2006)
31. Zetta, T., Kontopoulos, E., Bassiliades, N.: S 2 red: a semantic web rule editor (2011)

Clustering of Maintenance Tasks for the Danish Railway System

Shahrzad M. Pour[1(✉)] and Una Benlic[2]

[1] DTU Management Engineering, Technical University of Denmark, Produktionstorvet, 2800 Kongens Lyngby, Denmark
shmp@dtu.dk

[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK
u.benlic@qmul.ac.uk

Abstract. Standardisation of the European rail traffic signalling system is an ongoing project for faster travel within the EU, which entails very strict time limits and constraints on recovery operations. Denmark will be the first country to upgrade its entire signalling system to implement the new standards. In this paper, we present a mathematical model for allocation of maintenance tasks to maintenance team members, which is a variant of the Generalized Assignment Problem. The aim is to optimise the following three criteria: (i) the total distance travelled from depots to tasks, (ii) the maximal distance between any maintenance task and its allocated crew member, and (iii) the imbalance in workload among crew members. As test cases, we use a set of instances that simulate the distribution of tasks in the Jutland peninsula, the largest region of Denmark.

Keywords: European Rail Traffic Management System · Maintenance scheduling · Clustering · Mathematical model

1 Introduction

The European Rail Traffic Management System (ERTMS) [1] is a leading European industrial project that aims to standardize rail traffic signaling that will contribute to faster and safer travel within the European Union. Currently, several countries have gradually started substituting their signaling system to ERTMS, not only in the Europe but across the world. This is the result of characteristic of ERTMS which is independent from any signaling system and at the same time consistent with them. Thereby, it provides an opportunity for gradual movement of existing incompatible signaling systems towards ERTMS, depending on the potential national restriction and the economic standards of various countries.

Denmark will be the first country to upgrade its entire signaling system to the ERTMS standards, and has invested approximately 3 billion euros for this

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_78

purpose [2]. The motivation behind such a total renewal is that railway signalling assets are getting over aged. The current system is based on a national Automatic Train Protection(ATP) and color-light signalling, which has been implemented between 1986 and 1988 [2]. According to Banedanmark, the infrastructure owner of most railways in Denmark, the failure of signalling components are the main source of more than 50% of the delays in railway traffic.

ERTMS introduces a completely different hardware in the new system mainly in the form of a new on-board signaling equipment [3]. Due to new hardware, maintenance tasks of the new system differ significantly from the previous ones, involving very strict time limits and constraints on recovery operations. Given the complexity and high interdependence of the railway track and signaling systems, failure of a single component can have an impact on a large part of the railway network [4]. Therefore, system maintenance is crucial for effective functioning of the system. Nevertheless, little research has been conducted on maintenance issues related to the implementation of ERTMS as the project itself is still in its early stages [5].

As specified by the industrial partner of the ERTMS project in Denmark, the railway network should be divided into sub-regions where each sub-region is maintained by a different member of the maintenance team. Apart from the requirement that the workload needs to be fairly distributed across sub-regions (i.e., crew members), the geographic position of the maintenance points (tasks) should ensure that crew members can travel from one task to another quickly when needed so as to handle unexpected disruptions and failures. This stage in the maintenance planning process is termed the *assignment phase* or the *clustering phase*. Once the sub-regions are defined in the assignment phase, the planner can determine the maximum distance availability of the crew members in case of future failures prior to the routing phase. Finally, the *routing phase* is to find the best route for each crew member so as to minimise the total distance traversed for the entire maintenance plan. In the current maintenance planning system in Denmark, crew may start their duties from different locations rather than from a single depot. This requires an efficient allocation of tasks to avoid long total distance, or in some cases to ensure that a feasible plan is found. Each crew member is then responsible for maintenance within its own sub-region.

The combined assignment and routing phases of the above described maintenance process can be seen as a Multi Depot Vehicle Routing Problem (MDVRP) [6], where each vehicle (crew member) operates on its own routes starting and ending from a specific depot. The decision makers have to determine which tasks are served by which depots, prior to the routing and scheduling problem. As MDVRP is NP-hard, a wide range of heuristics have been used for this problem, including Tabu Search [7] and Adaptive Large Neighborhood Search [8]. A recent survey on approaches to MDVRP can be found in [9]. A popular strategy to tackle MDVRP is the "cluster-first route-second" approach, where the clustering phase is usually solved as a Generalized Assignment Problem (GAP) [10] which itself is NP-hard. In GAP, there are a set of tasks and a set of agents. The tasks need to be assigned to the agents with the goal of minimum total

assignment cost. Furthermore, each agent has a workload limit and the sum of the weighted tasks assigned to it cannot exceed this limit. The problem then aims to find an allocation of agents to tasks such that the total profit of the assignment is maximised without exceeding the workload limit of each agent. In [11], the authors propose several algorithms dedicated to the assignment (clustering) phase for MDVRP.

This work tackles the clustering phase of the railway maintenance plan at Jutland peninsula, the largest region of Denmark. We propose a mathematical model and test it on a set of instances that simulate the distribution of tasks in the Jutland peninsula.

2 Mathematical Model

As previously mentioned, the problem of allocating maintenance tasks to the maintenance team of a railway system can be viewed as an instance of the Generalized Assignment Problem (GAP). Let C be a set of crew members, let M be a set of tasks, and let $Q_{k,l}$ be a matrix indicating a distance between each depot/crew member $k \in C$ and each maintenance task $l \in M$. Furthermore, let d_l be the weight (e.g., duration) associated to $l \in M$, and let b be the maximum workload limit per depot. The model for the general assignment problem is given in Eq. (1)–(3). The constraint enforced by Eq. (2) is to assign exactly one crew member $k \in C$ to each task $l \in M$, and exactly one task to each crew member, while Eq. (3) imposes a maximum workload limit b to $k \in C$ for a balanced distribution of tasks across sub-regions. We assume that $d_l = 1$ for each $l \in M$ and that $b = \lceil |M|/|C| \rceil$. The objective is to minimise the total distance travelled from depots to tasks.

$$min \sum_{k \in C} \sum_{l \in M} x_{k,l} * Q_{k,l} \tag{1}$$

$$\sum_{k \in C} x_{k,l} = 1 \qquad \forall l \in M \tag{2}$$

$$\sum_{l \in M} d_l * x_{k,l} \leq b \qquad \forall k \in C \tag{3}$$

We extend and adapt the above defined GAP model to take into account the specific requirements considered during depot to task allocation for the Danish railway system. The objective of the proposed model is the minimisation of the following three criteria: (i) the total distance traveled from depots to tasks (as defined in Eq. 1), (ii) the maximal distance between any maintenance task and its allocated crew member, and (iii) the imbalance in workload among crew members. Coefficients w_1, w_2 and w_3 correspond to the weights associated to the three respective problem criteria. Note that unlike in the GAP model, we consider the workload balance as an objective rather than a constraint. The complete model is defined in Eq. (4)–(7). Let ψ denote the upper bound for the maximal distance between any depot-task pair. The constraint in Eq. 5 ensures that each depot-task is separated by at most ψ, where ψ is one of the terms

to be minimised. In this way, all the tasks maintained by a crew member are distributed in circles, where ψ is the circle diameter. Let δ denote the upper bound for the imbalance in workload distribution across different sub-regions on the railway network, and let d_l denote the duration of task $l \in M$. The constraint in Eq. 6 ensures that the workload distribution between any two depots $k \in C$ and $v \in C$ differs by at most δ, where δ is to be minimised. Finally, constraint in Eq. 7 serves the same purpose as in the GAP model, i.e., to enforce the assignment of each task to exactly one crew member.

$$min\ w_1 * \sum_{k \in C} \sum_{l \in M} x_{k,l} * Q_{k,l} + w_2 * \psi + w_3 * \delta \tag{4}$$

$$x_{k,l} * Q_{k,l} \leq \psi \qquad \forall k \in C, l \in M \tag{5}$$

$$\sum_{l \in M} x_{k,l} * d_l - \sum_{l \in M} k_{v,l} * d_l \leq \delta, \qquad \forall k, v \in C \tag{6}$$

$$\sum_{k \in C} x_{k,l} = 1 \qquad \forall l \in M \tag{7}$$

3 Experimental Results

3.1 Experimental Protocol

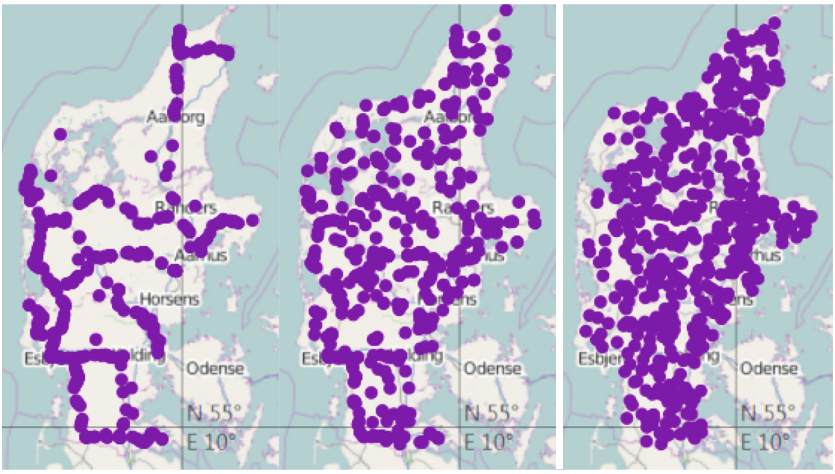
The proposed model, as well as the GAP model used for comparisons in Sect. 3.4, is coded in GAMS and executed on a machine equipped with Intel (R) Core (TM) i7-4600U CPU at 2.10 GHz and 8 GB of RAM. We use the default setting for GAMS, i.e., $optca = 0.0$ and $optcr = 0.1$.

3.2 Benchmark Instances

The coordinates, representing geographical locations of maintenance tasks, are located in the Danish peninsular of Jutland and are extracted by using the Google Map API. To evaluate the performance of the proposed model on instances with different structures (distributions of the maintenance tasks), we generate and group instances into three classes based on geographical locations of the tasks. More precisely, tasks can be located at:

- Rail tracks of the maintenance area. Instances from this class are prefixed with 'E';
- Mixture of on-tack or off-track positions within the maintenance area. Instances of this class are prefixed with 'M';
- Scattered at random positions around the entire Jutland region. Instances from this class are prefixed with 'R'.

For each instance class, we generate three instances with 100, 500 and 1000 tasks respectively (a total of 9 instances), according to the number of maintenance tasks performed on a daily, weekly and monthly basis. We suppose that these are served by a team of 8 crew members. The used datasets, as well as the associated documentation describing the instance generator, are available at http://github.com/ShahrzadMP/RegionSplitterDataset. To standardise the test cases, we follow the file format of the popular data set that is generally used to test algorithms for the Vehicle Routing Problem with Time Windows (VRPTW)[1]. Instances are referred to by their class and number of tasks - $Class\#Tasks$ - e.g., $E100$, $R1000$, etc. Figure 1 visualises geographical locations corresponding to the three types of instances with 500 tasks.



(a) E500 (b) M500 (c) R500

Fig. 1. Geographical visualization of the maintenance task distributions for three types of datasets.

3.3 Trade-Off Between the Three Problem Criteria

The purpose of this section is to investigate the trade-off between the three problem objectives: (i) the total distance travelled from depots to tasks, (ii) the maximal distance between any maintenance task and its allocated crew member, and (iii) the imbalance in workload among crew members. According to Denmark's rail traffic management, the most important solution quality criterion is objective (ii), followed by objectives (iii) and (i). To analyze the trade-off in terms of the three criteria, we try four different combinations of the corresponding weight coefficients such that $w_2 \geq w_3 \geq w_1$, where $w_1, w_2, w_3 \in [0, 1]$.

[1] http://w.cba.neu.edu/~msolomon/problems.htm.

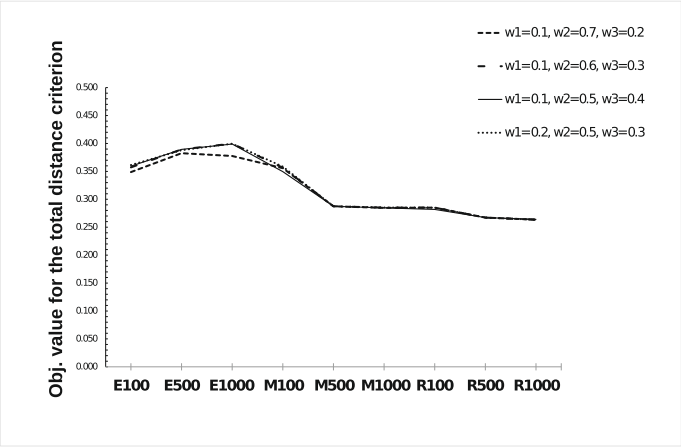


Fig. 2. Comparison of solutions in terms of objective (i), i.e., in terms of the total distance traveled from depots to tasks.

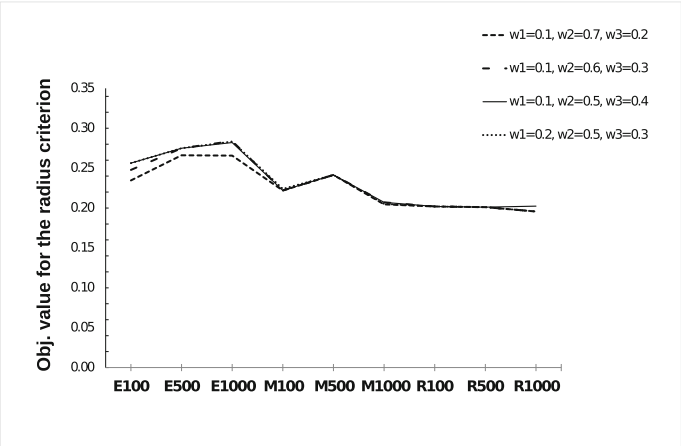


Fig. 3. Comparison of solutions in terms of objective (ii), i.e., in terms of the maximal distance between any maintenance task and its allocated crew member. The maximal distance is also termed the "radius".

Furthermore, in all our experiments, we normalise the objective values associated with each individual criterion to take on a value in the range $[0, 1]$.

For our set of 9 benchmark instances, Figs. 2, 3 and 4 illustrate the normalised solution quality in terms of the three objective criteria respectively, obtained with the four selected settings of the weight coefficients. When varying weights of coefficients $w_1 - w_3$, we observe a significant difference in performance regarding the workload balance criterion (objective (iii)), while this difference is less evident in terms of objectives (i) and (ii).

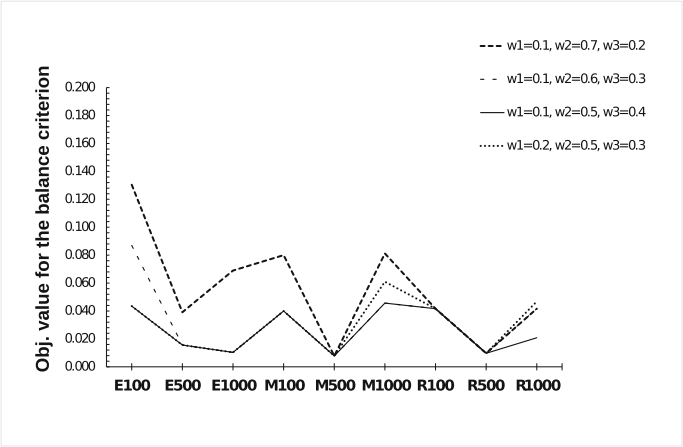


Fig. 4. Comparison of solutions in terms of objective (i), i.e., in terms of imbalance in the workload distribution.

3.4 Results and Comparisons

The aim of this section is to report and compare the output of the proposed model (denoted as PM) with that of the model for the Generalized Assignment Problem (GAP). Considering the prioritization of the objectives specified by the rail traffic management and the performances observed in Figs. 2, 3 and 4, we set the objective weight coefficients to $w_1 = 0.1$, $w_2 = 0.6$, $w_3 = 0.3$.

For each problem instance, Table 1 shows the upper bound, the lower bound and the relative gap obtained with the proposed model, as well as the CPU time in seconds required to reach the presented result. As mentioned in the previous section, the bounds are normalized in the range $[0, 1]$. Based on the default setting of GAMS, the relative optimality tolerance is 0.1, and the search stops as soon as this threshold is reached. As the reported relative gap for PM is never greater than 0.1, this implies that the PM solutions are optimal or very close to the optimal ones. The GAP model has shown to be easier to solve with GAMS as the optimality gap is 0% in all the cases. The computing times are short for both models, ranging from 2 to 65 s for PM and from 1 to 15 s for GAP.

Table 2 compares solution qualities produced with PM and GAP in terms of the three problem criteria. The values for each criterion are non-normalized. For each objective, column '%Diff.' denotes the percentage difference between a PM and a GAP solution regarding the given objective. A negative '%Diff' value implies that the GAP solution is of a better quality, while a positive '%Diff' means that the PM solution is more favourable in terms of the given criterion. When compared solely against the total distance criterion, we notice a slight advantage of the GAP model in 7 out of the 9 cases with an 8.86% difference for instance M100. On the other hand, solutions obtained with PM tend to have a more compact distribution of the maintenance tasks (with less outliers).

Table 1. Upper bounds (UB), lower bounds (LB) and the relative gap of the solutions obtained with PM, as well as the CPU time in seconds required to solve the PM and the GAP model.

Instance	UB	LB	Relative gap	CPU_{PM}(s)	CPU_{GAP}(s)
E100	0.210	0.189	0.100	3.34	1.92
E500	0.209	0.188	0.096	16.24	6.58
E1000	0.213	0.192	0.100	65.28	9.42
M100	0.181	0.165	0.089	2.4	2.38
M500	0.176	0.174	0.014	5.45	4.35
M1000	0.166	0.151	0.093	12.16	10.91
R100	0.162	0.150	0.077	2.47	1.34
R500	0.150	0.148	0.018	9.78	5.2
R1000	0.156	0.144	0.080	11.01	14.66

In other words, the maximal distance between any maintenance task and its corresponding crew member is significantly shorter with the PM model in 6 out of the nine instances. Such results are as expected as the GAP model does not take into account the radius objective, which is the primary objective for the Danish rail traffic management. As for the balanced distribution of workload, we observe the advantage of GAP only for the three largest problem instances.

Table 2. Comparison of solutions obtained with the proposed model (PM) and the GAP model in terms of the three problem objectives.

Instance	Total distance - Obj. (i)			Radius - Obj. (ii)			Imbalance - Obj. (iii)		
	GAP	PM	%Diff.	GAP	PM	%Diff.	GAP	PM	%Diff.
E100	3566.44	3605.84	−1.10	69.99	65.64	6.22	3	2	33.33
E500	19582.29	19870.08	−1.47	121.71	72.61	40.34	4	2	50.00
E1000	40002.31	40416.25	−1.03	122.22	75.25	38.43	0	3	–
M100	3071.67	3343.75	−8.86	68.47	57.98	15.32	4	1	75.00
M500	14317.46	14474.94	−1.10	63.95	63.95	0.00	4	1	75.00
M1000	29207.42	28911.66	1.01	69.93	57.38	17.95	0	9	–
R100	2467.56	2570.14	−4.16	52.24	52.24	0.00	3	1	66.67
R500	12492.07	12528.54	−0.29	53.12	53.12	0.00	4	1	75.00
R1000	24693.42	24596.10	0.39	56.93	50.61	11.10	0	8	–

4 Conclusion

In this paper, we propose a mathematical model for partitioning a spatial area into sub-regions considering the geographical locations of the Danish railway network. Due to the specification of the danish railway network, maintenance scheduling can be seen as a Multi Depot Vehicle Routing Problem (MDVRP), where each vehicle (crew member) operates on its own routes starting and ending from a specific depot. This paper deals with the clustering phase of MDVRP, which can be viewed as a variation of the Generalized Assignment Problem (GAP). More precisely, the problem consists in determining which tasks are served by which depots, prior to the routing and scheduling phase, by taking into account the specific requirements to deal with failures and breakdown of the railway network. For this purpose, we presented a mathematical model that optimises the three problem criteria, and applied it on a set of instances that simulate the distribution of tasks in the Jutland peninsula, the largest region of Denmark. This clustering phase, results in rescaling the size of the problem in the scheduling phase, having the possibility to do parallel scheduling, and more sophisticated planning framework. As future work, a comparison between different scheduling approaches with and without clustering phase is proposed.

References

1. Barger, P., Schön, W., Bouali, M.: A study of railway ERTMS safety with Colored Petri Nets. In: Proceedings of the European Safety and Reliability Conference (ESREL 2009), vol. 2, pp. 1303–1309 (2009)
2. Banedanmark., Trafikministeriet.: The signalling programme - a total renewal of the Danish signalling infrastructure, Banedanmark (2009)
3. Zimmermann, A., Hommel, G.: Towards modeling and evaluation of ETCS real-time communication and operation **77**(1), 47–54 (2005)
4. Liden, T.: Survey of railway maintenance activities from a planning perspective and literature review concerning the use of mathematical algorithms for solving such planning and scheduling problems. Technical report, Linköping University, Department of Science and Technology (2014). http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-111228
5. Patra, A.P., Dersin, P., Kumar, U.: Cost effective maintenance policy - a case study. Int. J. Perform. Eng. **6**(6), 595–603 (2010)
6. Lenstra, J., Kan, A.: Complexity of vehicle routing and scheduling problems. Networks **11**(2), 221–227 (1981)
7. Cordeau, J., Laporte, G., Mercier, A.: A unified tabu search heuristic for vehicle routing problems with time windows. J. Oper. Res. Soc. **52**(8), 928–936 (2001)
8. Pisinger, D., Ropke, S.: A general heuristic for vehicle routing problems. Comput. Oper. Res. **34**(8), 2403–2435 (2007)
9. Montoya-Torres, J.R., Franco, J.L., Isaza, S.N., Jimenez, H.F., Herazo-Padilla, N.: A literature review on the vehicle routing problem with multiple depots. Comput. Ind. Eng. **79**, 115–129 (2015)
10. Shmoys, D.B., Tardos, E.: An approximation algorithm for the generalized assignment problem. Math. Program **62**(3), 461–474 (1993)
11. Giosa, I., Tansini, I., Viera, I.: New assignment algorithms for the multidepot vehicle routing problem. J. Oper. Res. Soc. **52**(5), 977–984 (2001)

Evaluating the Quality of Business Process Models Based on Measures and Criteria in Higher Education: Developing a Framework for Continuous Quality Improvement

Fouzia Kahloun[1(✉)] and Sonia Ayachi Ghannouchi[2]

[1] Laboratory RIADI-GDL, ENSI, Mannouba, Tunisia
foufoukahloun@yahoo.fr
[2] High Institute on Management of Sousse, Sousse, Tunisia
s.ayachi@coseleam.org

Abstract. In recent years, modeling and improving the quality of business processes have been increasing interest in several domains, especially in higher education. In fact, quality of business process models could be the object of various interpretations; it must be treated in the most objective way, by using measures. Consequently, this paper presents a framework where we apply these steps for the purpose of quality management in the higher education domain. After that, we illustrate these ideas with a real business process model related to the tracking of curriculum offers process more commonly named "habilitation process".

Keywords: Business process model · Framework · Quality · Criterion · Higher education · Measure

1 Introduction

BPM provides a significant way to ensure changes in the structure of a company and particularly in higher education, since it corresponds to one of the most important domains for development in the economic world. For that reason, we must boost the culture of quality within higher education institutions on the one hand and improve the quality of business processes on the other hand.

In this context, we developed a framework for continuous improvement of the higher education process founded on Business Process Management Modeling.

A thorough study is conducted in literature allowing us to classify these approaches into three categories. The first category consists on approaches based on methods, defining methodological guides and good practices to determine the quality of BP models. Second, approaches of quality assessment for business processes, which are mainly applied during their execution and their control. Finally, approaches of quality assessment applied to BP models. Through our research, we focus on the assessment and improvement of business process models (last category). This approach is founded on quality criteria and metrics. Authors who used such approaches have focused on

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_79

software measures because there is a strong analogy between software programs and business processes [1].

In this paper, we present a framework and we apply it on a real case of a process in the domain of higher education. First, we present an overview of selected existing metrics for defining different aspects of BP models. Then, we give an overview on the way of applying them for higher education processes. In Sect. 2, we define a framework and we specify its components. Section 3 refers to the diverse criteria and measures in the software engineering domain. As for Sect. 4, it is an illustration devoted to the application of such measures on a process model in higher education domain. Section 5, is based on assessment and improvement of the result of the previous section. As for Sect. 6, it spots the light on some recommendations and metrics for ensuring quality in BP models. We end up with the seventh section which sums up our conclusion and future works.

2 Developing a Framework for Continuous Quality Improvement

To obtain the quality management model, we took into account the BPM life cycle which is composed of four steps: Design & model, development & testing, execution, analyzes & optimization. We developed a framework for defining a business process model, describing the methods and means used by this latter to achieve a higher-quality model that is comprehensible, interpretable and which can ensure good execution.

Indeed, our framework considers a business process model as input and an improved business process model as output. The major difference between then is that the input model is without control and assessment of quality but the output model meets the criteria and threshold values that define the targeted level of quality. This framework is based on three main elements which are criteria, measures or performance indicators

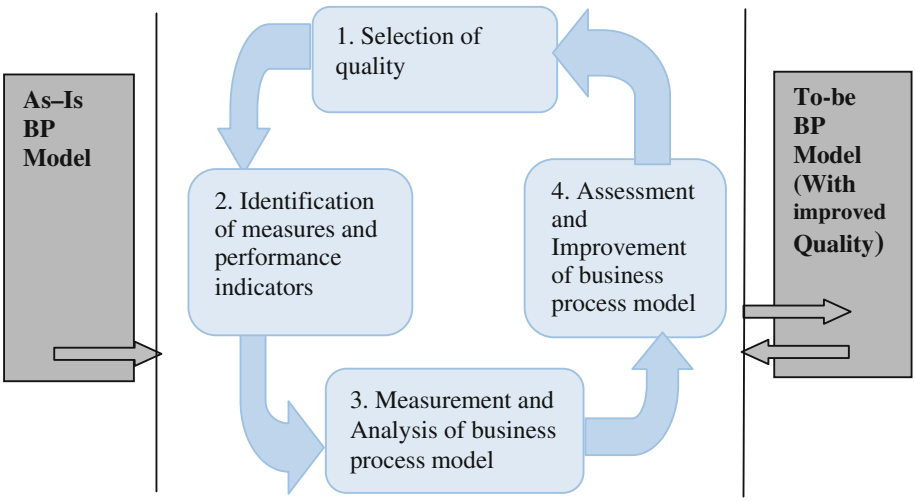


Fig. 1. A framework for continuous improvement of higher education

and finally the models of business process management. These elements directly influence the output models. First, we select the criteria then for each criteria we determine all measures or performance indicators, afterwards we can apply on a real business process model in order to assess and improve its level of quality. Thus the above mentioned steps conduct to a measurement level analysis and improvement that will contribute to the evaluation of a process model [2] (Fig. 1).

3 Selection of Criteria and Quality Metrics Stirred from Software Engineering Domain

This part is dedicated to present the quality criteria and measures of business processes that have been adapted from the software engineering domain. So these two tasks allow achieving the first and the second step of our framework.

3.1 Coupling

The coupling (CP) measure is used to describe how the activities of a BP model are connected. An activity is associated to another in case they share information with each other (e.g. data).

Differently expressed, it counts all pairs of activities in the BP model that are connected to each other [3].

Vanderfeesten et al. (2007) presented a mathematical formula of measure for this criterion. More details on the formula itself and the explanation are available in [4, 5]. Moreover the average value of the measurements is determined by a ratio between the total numbers of connections for all activities and the total number of activities.

It is worthy to mention that Reijers and Vanderfeesten (2004) also established a similar coupling metric counting the overlap of data elements for each pair of activities using a static description of the product represented by a Product Data Model (PDM) [6]. The outcome is always between 0 and 1.

3.2 Cohesion

First, we need to state that Vanderfeesten et al. [7] introduced cohesion (ch) metrics that are adapted from software engineering domain. This measurement determines the different operating numbers that are connected within an activity. Moreover, it determines, for each activity operation, how many other operations it overlaps with, by sharing an input or output [3]. We can conclude that we will have more errors when we have a low cohesion compared with a model with high cohesion. The cohesion of a BP model is calculated via a ratio between the sums of all the values of cohesion of activities and the number of activities.

In addition, concerning this criterion, Vanderfeesten et al. (2008) define the cross connectivity metric which quantifies the ease of understanding and the interplay of any pair of model elements [8] (for the mathematical formula and the explanation of this metric, the readers is referred to [9]).

Moreover, a high value of cohesion in the activities of a BP model defines a better modular decomposition of its activities. Both Principles of cohesion and coupling are important for BP model to build processes that are highly coherent and weakly coupled to one another. It is noteworthy that the cohesion value is always between 0 and 1.

3.3 Complexity

Cardoso (2006) proposed the control flow complexity (CFC) metric for measuring control flow complexity of a BP model [10]. This metric is an adaptation of Cyclomatic Metric (MCC). This measurement determines the simplicity as well as the understandability of a model. It can be handled from different perspectives since they are high-level concepts, which are made up of countless different elements (splits, joins, resources, data, etc.). The main idea behind this metric is to assess the complexity introduced in a process because of the existence of gateways such as XOR–Split, OR–Split, AND–Split. For XOR-split the CFC is simply fan-out (a); for OR-split is 2 fan-out (a)-1. Where fan-out is the output points number of splits [11], and for AND-split is simply equal to 1. Accordingly, it is easy to calculate the complexity of a BP model, by the sum of different CFC from all splits in the business process model.

We can also measure the complexity by another method based on HPC measure. Cardoso et al. [12] introduced the notion of Halstead-based Process (HPC) measures. This requires calculating: length N, volume V and difficulty D of a BP model, calculated as follows:

$$N = n_1 * \log_2(n_1) + n_2 * \log_2(n_2). \tag{1}$$

$$V = (N_1 + N_2) * \log_2(n_1 + n_2). \tag{2}$$

$$D = (n_1/2) * (N_2/n_2). \tag{3}$$

Where:

n_1: describes the number of activities, control flow, and different gateways for a BP model.

n_2: describes the unique number of treated data by the activities of the process.

N_1 and N_2 can be easily derived directly from n_1 and n_2 as the total number of control flows and elements [13].

In addition, Ivanchikj (2014) presents another metric to calculate the complexity of model named coefficient of connectivity (CC). This latest defines how well connected are different nodes in the model (activities, events, gateways) between each other [14].

Furthermore, Sanchez et al. [15], mentioned that a model is considered as very easy to understand if these complexity measures do not exceed 6 for $CFC_{XOR\text{-}split}$, 1 for $CFC_{OR\text{-}split}$, and 1 for $CFC_{AND\text{-}split}$. However, it is hard to understand it if it contains 46 for $CFC_{XOR\text{-}split}$, 14 for $CFC_{OR\text{-}split}$, and 7 for $CFC_{AND\text{-}split}$.

3.4 Modularity

Modularity is a measure that describes the modularity (M) of a BP model. It corresponds to the degree of decomposition of the model into sub-models. Besides, modularization may encourage the understanding of a process model by its "information hiding" quality [16]. The degree of modularization affects the quality of a design. As for the over-modularization, it is as undesirable as under-modularization [6].

Moreover, Mendling et al. [17] proposed that a model with more than 50 elements should be modularized. It is determined as follows:

$$M = (\text{fan} - \text{in} * \text{fan} - \text{out})^2. \tag{4}$$

Where Fan-in: describes all the modules calling another module m of the model and Fan-out: describes all the modules called by a module m.

3.5 Size

Speaking about the size, we can measure it by adopting the LOC (Line of Code). This metric comes from software engineering and measures the size of a BP model. More precisely, we adopted the three following metrics:

NOA: sums up activities (tasks and sub-processes) in a BP model [18]
NOAC: Number of activities and control-flow elements [19]
NOAJS: Number of activities, splits, and joins, is NOAC used for processes that are not well-structured (there is at least one splitting of process flow that is not closed) [20].Furthermore, Ivanchikj [15] defines the size of the model which can be measured as the number of any of the elements that the model comprises. Hence, he defines many measures for this criterion: for instance total number of events (TNE), total number of Gateways (TNG), total number of Nodes is the sum of routing elements in a business model and activities, total number of Lanes, total number of sequence flow (TNSF), etc.

We justify our choice of all these selected metrics by the fact that they are the most cited in literature. Moreover, other metrics can be derived from them.

In the next section, we present a real case of a business process model for the higher education, which is the tracking of curriculum offer process. At that point, we are going to indicate obtained values for the selected quality metrics.

4 Measure and Analysis of a Real Case of the Tracking of Curriculum Offer Process in Higher Education Domain

In this part we will deal with the third step of framework. In fact, we developed a process model of higher education then we realized measures on the latter in order to analyze. Figure 2 presents an overview of the tracking of curriculum offer process in higher education domain. We realized the BPMN model from a document presented in the

official website of the higher education and scientific research Ministry in Tunisia. Moreover, it has been validated by involved actors.

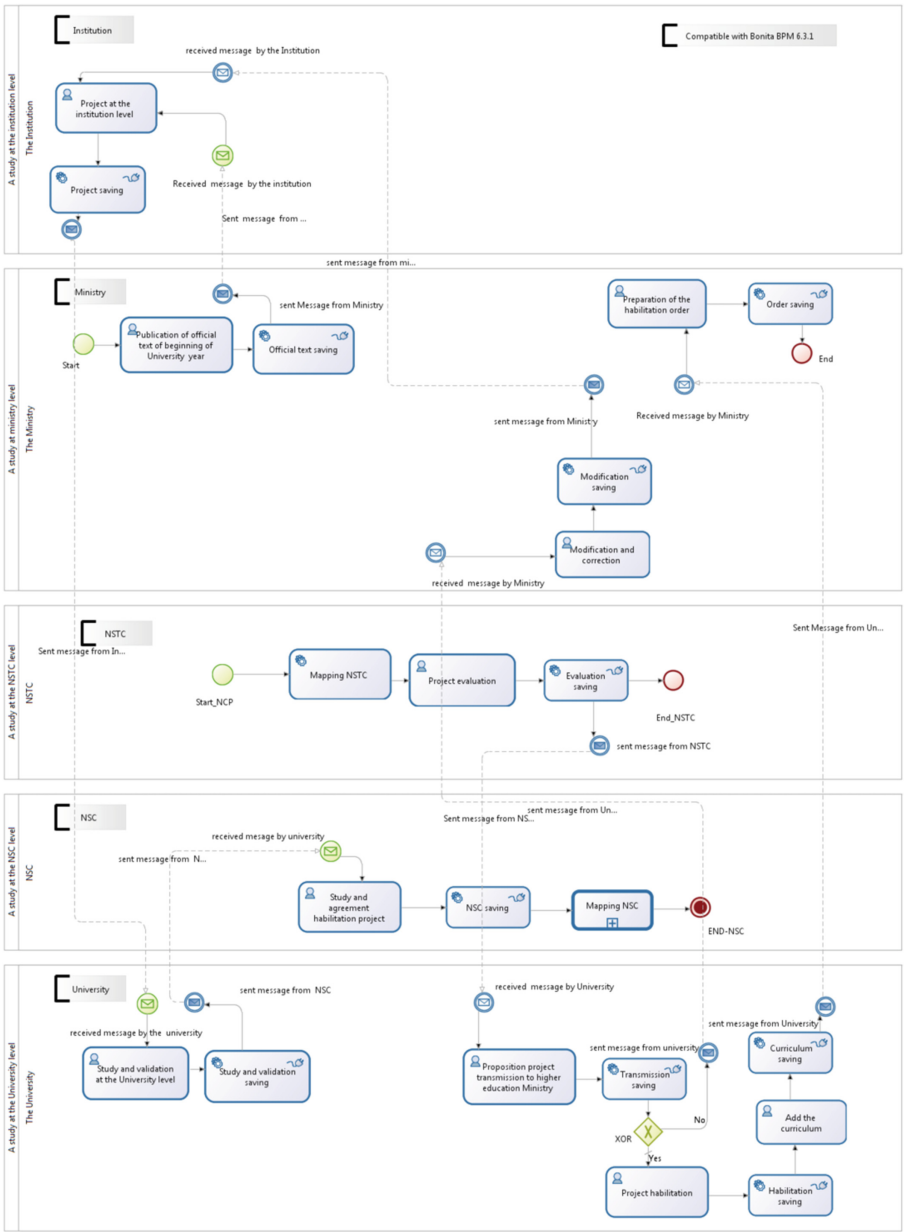


Fig. 2. BPMN model of the tracking of curriculum offers process in higher education domain.

4.1 Description and Modeling

In order to improve the quality of processes and increase their benefits on several levels, through the establishment of the Business Process Management (BPM) approach, we illustrate our ideas on this case study.

Through this paper, we tend to present the case of the tracking of curriculum offer process. We can justify our choice concerning this real case by the importance it has in the domain of higher education. In fact, a curriculum offer is defined as a starting point which is grounded on the offered curricula in higher education institutions. Its adaptations can define new employment opportunities for the upcoming graduates. However, it should be noted that it is quite difficult to deal with curricula offers because they require a long time to be established and a large number of involved actors. This tracking of curriculum offer process is known as the habilitation process. Five actors intervene in this process which is the ministry, the university, the higher education institution, the national sectorial committee (NSC), and the national Steering committee (NSTC).

This process of habilitation starts with the publication of the official text of beginning the university year by the ministry. At that juncture, a proposition of project is triggered at the institution level. A study of this proposition will be fulfilled at the university level. Afterwards, an assessment of the NSTC of the project proposition following a validation and an agreement of the NSC are dealt with. After that, the university transmits the project proposition to the Higher Education ministry. In this phase, a decision will be taken in order to decide if the project has to be modified. We must return to the step of project proposition at the institution level and repeat these phases after a modification is required by the ministry. However, if the decision is positive (yes), then the habilitation of the project is accomplished, the curriculum is added by the university for interested students and the ministry prepares the habilitation order. In order to make our process executable, after each activity, saving data in MYSQL database must be accomplished.

4.2 Measures and Their Interpretation

We will calculate a number of measures for this business process model, by considering the criteria mentioned above. It is not possible to apply all of them due to the absence of some specific elements in this model. These measures are automatically calculated by using two tools. Vanderfeesten et al. (2008) developed a tool to support the design of a workflow process model based solely on coupling and cohesion metrics. This tool, named CoCoFlow tool (Cohesion- Coupling metrics for workFlow models) [7]. Only loads XML files with predefined format. Its main function is the calculation of metrics for business process design. On the other hand, for the other measures, Ivanchikj et al. (2015) developed a web application named BPMeter which provides a simple Web application to visualize the results of applying over 100 different size and structure metrics to BPMN 2.0 process models [21]. We present in Table 1 some measurement results that are obtained.

Table 1. Selected calculated metrics for our process model.

Metric	Result
CP	0.244
Cohesion (Ch)	0.32
$CFC_{AND\text{-}split}$	0
$CFC_{OR\text{-}split}$	0
$CFC_{XOR\text{-}split}$	2
HPC-N	134.36
HPC-V	254.53
HPC-D	21.84
Coefficient of connectivity	0.767
FanIn	1
FanOut	1
M	1
NOA = NOAC	22
NOAJS	23
TNE	20
TNG	1
Number of nodes	43
Number of lanes	5
TNSF	34

5 Model Assessment and Improvement

After getting the results of the selected measures, we must evaluate and try to improve our model. Therefore we compared them to threshold values defined in literature. We conclude that according to Sanchez et al. (2011), our model is very easy to understand. We can thus judge that our model has low complexity.

We also evaluate the other measures related to size via the threshold presented in [22]. Thus, we conclude:

TNE is 20, this means that the model is very inefficient. TNG is 1, thus the model is considered as very efficient. Number of nodes is 43, thus the model is fairly inefficient. TNSF is 34, thus the model is fairly efficient.

Following the evaluation, we remarked that TNE, and Number of nodes had a high value. Consequently, the number of events and nodes must be reduced, but TNSF got acceptable value. Hence this result did not require to be analyzed for other improvement.

Cp = 0.244, Ch = 0.32, M = 1 obtained values are considered as acceptable results for a good quality model.

Otherwise, measuring HPC gave very high values specially HPC-N and HPC-D. These results indicate that the number of data variables and elements in the model must be reduced. Therefore, we will have to focus on an analysis of some parts of the model in order to improve its quality.

6 General Recommendations for Ensuring Quality in BP Models

We devote this paragraph to recommendations and tips for business process modelers. In order to achieve a BP model with good quality, Mendling et al. [17] proposed seven recommendations for the way of presenting a process model from scratch, or for improving the quality of an existing process model. They are named seven process modeling guidelines (7PMG). These guidelines are the following:

1. Reducing the number of the elements because the size affects the understandability and the error probability of a model;
2. Minimizing the routing paths per element;
3. Attempting the use of one start and one end event;
4. Having a well-structured model for which every split connector matches a respective join connector of the same type.
5. Avoiding the use of OR routing elements;
6. Applying a verb-object form of the activity label;
7. Decomposing a model if it contains more than 50 Elements.

After giving recommendations on designing a good model of business processes, we give recommendations on BP model measures. In fact, a low coupling measurement value is sought after. The higher is the coupling value of the process, the more problematic it is to change the process and the higher is the probability that there will be errors in the process [23]. We can solve this problem if we have a high value for this measurement by applying the first and the fifth guidelines from 7PMG.

High cohesion measure values are preferable for high quality Models. Vanderfeesten et al. (2008) have stated that a model with a low measure of cross connectivity is more likely to include errors. Low NOA, NOAC, NOAJS, TNG, Number of Nodes, Number of Lanes, and TNSF measures values are desired. These metrics presented some information to understand the model design. To remedy to the raised problem by these values, the idea is to apply the first guideline from 7PMG. On the other hand for NOAC metric, if we moreover apply the fourth guideline from 7PMG, we get a high value for this metric. Low TNE is desired. Hence, we improve this value by applying the third guideline from 7PMG. Low control flow complexity (CFC) measures are wanted. The greater the overall structural complexity of a process is, the higher value of the CFC will be obtained [13]. If CFC value of the model is high, we can reduce this value by applying the first and the fifth guidelines from 7PMG. As a final point, low HPC measures are desired. Cardoso et al. (2006) noted that the measures calibration needs should be examined through the empirical experiments. If HPC value of the model is high, we can solve this problem by applying the first guideline from 7PMG. If HPC value of the model is high, we can solve this problem by applying the first guideline from 7PMG.

7 Conclusion and Future Works

Currently, Quality measure and improvement for business process models is important in the domain of model-based software development. Through this paper, we present a

framework for measuring and analyzing the business process model characteristics in terms of quality. First we evaluate a number of criteria and measures that are considered as important in order to describe the various aspects of business processes and we apply these measures to a real process related to the tracking of a curriculum offer in the domain of higher education and finally we check the quality of the latter according to existing thresholds in literature to improve its quality. In future research work, we aim to develop a prototype to automate the measurement of quality metrics values for process models in higher education and obtain recommendations on the way to transform them in order to improve their quality.

References

1. Gokçen, To.: Evaluating the complexity of business process models. MS thesis, Atilim University (2011)
2. Dragan, M., Ivana, D., Arba, R.: Business process modeling in higher education institutions developing a framework for total quality management at institutional level. Procedia Econ. Finance **16**, 95–103 (2014)
3. Khlif, W., Makni, L., Zaaboub, N., Ben-Abdallah, H.: Quality metrics for business process modeling. In: Proceedings of the 9th WSEAS International Conference on Applied Computer Science, pp. 195–200. World Scientific and Engineering Academy and Society (WSEAS) (2009)
4. Cardoso, J., Vanderfeesten, I., Reijers, H.A.: Computing coupling for business process models (2010). http://eden.dei.uc.pt/~jcardoso/Research/Papers/Old%20paper%20format/Caise-19th-Coupling-Cardoso-Vanderfeesten.pdf
5. Vanderfeesten, I., Cardoso, J., Reijers, H.A.: A weighted coupling metric for business process models. In: CAiSE Forum (2007)
6. Vanderfeesten, I., Cardoso, J., Reijers, H.A., Van der Aalst, W.: Quality metrics for business process models. In: BPM and Workflow Handbook, pp. 179–190 (2007)
7. Vanderfeesten, I., Reijers, H.A., Vander Aalst, W.M.: Evaluating workflow process designs using cohesion and coupling metrics. Comput. Ind. **59**(5), 420–437 (2008)
8. Makni, L., Khlif, W., Haddar, N.Z, Ben-Abdallah, H.A.: Tool for evaluating the quality of business process models. In: ISSS/BPSC, pp. 230–242. Citeseer (2008)
9. Vanderfeesten, I., Reijers, H.A., Mendling, J., Aalst, W.M.P., Cardoso, J.: On a quest for good process models: the cross-connectivity metric. In: Bellahsène, Z., Léonard, M. (eds.) CAiSE 2008. LNCS, vol. 5074, pp. 480–494. Springer, Heidelberg (2008). doi: 10.1007/978-3-540-69534-9_36
10. Muketha, G., Ghani, A., Selamat, M., Atan, R.A.: Survey of business process complexity metrics. Inf. Technol. J. **9**(7), 1336–1344 (2010)
11. Kherbouche, M.O.: Contribution à la gestion de l'évolution des processus métiers. In: Laboratoire d'informatique signal et image de la Cote d Opale, p. 195. Université du Littoral Cote d Opale (2013)
12. Cardoso, J., Mendling, J., Neumann, G., Reijers, H.A.: A discourse on complexity of process models. In: Eder, J., Dustdar, S. (eds.) BPM 2006. LNCS, vol. 4103, pp. 117–128. Springer, Heidelberg (2006). doi:10.1007/11837862_13
13. Kluza, K., Nalepa, G.J.: Proposal of square metrics for measuring business process model complexity. In: 2012 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 919–922. IEEE (2012)

14. Ivanchikj, A.: Characterising representative models for BPMN 2.0 workflow engine performance evaluation. Ph.D. thesis (1994)
15. Sanchez, G.L, Ruiz, F., Garcia, F., Cardoso, J.: Towards thresholds of control flow complexity measures for BPMN models. In: Proceedings of the 2011 ACM Symposium on Applied Computing, pp. 1445–1450. ACM (2011)
16. Sanchez, G.L., Garcia, F., Piattini, M.: Toward a quality framework for business process models. Int. J. Coop. Inf. Syst. **22**(01), 1350003 (2013)
17. Mendling, J., Reijers, H.A., van der Aalst, W.M.: Seven process modeling guidelines (7PMG). Inf. Softw. Technol. **52**(2), 127–136 (2010)
18. Sadowska, M.: An approach to assessing the quality of business process models expressed in BPMN. e- Informatica. Softw. Eng. J. **9**, 57–77 (2015)
19. Antonini, A., Ferreira, A.M., Morasca, S., Pozzi, G.: Software measures for business processes. In: Advances in Databases and Information Systems (2011)
20. Frece, A., Juric, M.B.: Modeling functional requirements for configurable content- and context-aware dynamic service selection in business process models. J. Vis. Lang. Comput. **23**(4), 223–247 (2012)
21. Ivanchikj, A., Vincenzo, F., Pautasso, C.: BPMeter: web service and application for static analysis of BPMN 2.0 collections (2015)
22. Sanchez, G.L., Ruiz, F., Garcia, F., Cardoso, J., Piattini. M.: Business process model improvement based on measurement activities. In: ENASE, pp. 104–113 (2011)
23. Sadowska, M.: Quality of business models expressed in BPMN. MS thesis (2013)

Age and Gender Classification from Finger Vein Patterns

Wafa Damak[1(✉)], Randa Boukhris Trabelsi[1], Alima Damak Masmoudi[1], Dorra Sellami[1], and Amine Nait-Ali[2]

[1] Computers Imaging Electronics and Systems Group (CIELS) from Advanced Control and Energy Management Laboratory (CEM-Lab), Sfax Engineering School, University of Sfax, Sfax, Tunisia
damakwafa@gmail.com, trabelsiboukhrisranda@live.fr, damak_alima@yahoo.fr, sellamimasmoudidorra@yahoo.com

[2] Laboratoire d'Images, Signaux et Systèmes Intelligents (LISSI), Université Paris Est-Creteil, Paris, France
naitali@u-pec.fr

Abstract. The main goal of this paper is to build a system able to recognize the age range and the gender of individuals from their venous network characteristic. Accordingly, we develop an algorithm able to detect changes related to aging. Proposed age and gender recognition system is composed by 4 key steps: image acquisition, image preprocessing, feature extraction and age/gender classification. Image preprocessing is established by ROI extraction and image enhancement. ROI extraction separates the informative region from finger vein image. For image enhancement, we use Guided Filter based Singe Scale Retinex (GFSSR) method. In feature extraction step, we implement the LBP descriptor in order to characterize venous texture from finger veins. Our study is based on MMCBNU_6000 finger vein database. Experimental results prove that extracted attributes from finger vein can define the gender and the age class. Proposed age and gender classification process gives a recognition rate of 98% for gender classification and a recognition rate of 99.67%, 99.78% and 97.33% for respectively 2, 3 and 4 classes, for age classification.

Keywords: Finger vein · Age recognition · Gender recognition · ROI extraction · Image enhancement

1 Introduction

Although there are many studies on hand veins in the last decades, researchers focused only on identification [1–4]. A preliminary analysis indicates that the indices of human health, such as hand aging, can be detected from hand veins images [5]. In [6], Nadort et al. tested various factors affecting the vascular system during human life. They have found that the skin becomes thinner and less

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_80

elastic with increasing age. As a consequence, veins may appear more important. Despite their cambered appearance, veins do not change. Veins often swell during physical activity, but decrease thereafter once the person is at rest. Hand veins appearance can also become more pronounced if a person lifts weights. Other factors influence the vascular system during the normal life, such as the environment temperature and alcohol consumption. In [7], Robert et al. tried to determine the factors that could influence on the patient's age. They treated factors that characterize hand age and they concluded that hands with less visible veins generally appear younger. According to Robert et al., wrinkles, veins, prominent joints, thin skin, deformity, and spots characterized old hands, by against, fullness, lack of veins, and lack of wrinkles characterized young hands. Yang et al. analysed statistical characteristics (ages, genders, BMI (Body Mass Index) and FMI (Fat Mass Index)) of hand vein image [8]. Experiments are conducted in a population of 3 to 60 years old and the total number of samples is 400. Through their experience and observation, there is little effect of age on the acquired image of the hand veins and the main factor affecting the quality of hand vein image is the FMI. In [5], Wang et al. analyzed dorsal hand veins for aging detection. They collected a database containing 1000 images from 50 people: 40 aged people and 10 young people. They predefined people aged over 60 years as aged people and people aged under 30 years as young people. There is a difference of brightness distribution between the aged person image and the young person image. The remarkable difference between dorsal veins of young person and aged person is characterized by the blood flow and the skin state. Therefore, gray levels of veins and skin account for aging detection. So, in order to distinguish a young person from aged person, they used the grayscale histogram and the mean pixel gray value of the vein and the skin as features. From these features, they employed k-means clustering for classification and the Euclidean distance for recognition. The aging recognition rate from the old is 82% using histogram and 76% using mean pixel. On the other hand, the aging recognition rate from the young is 5% using histogram and 2% using mean pixel.

The interest about age recognition is in growing as there is recently many practical applications relying on human demographic data. It is recognized that human-computer interaction differs for various age groups. So, an application that adapts its interface to the human age is considered as an important task. Moreover, in the case of an access control based age, there exist restrictions related to age, for example, limiting the access to a site web, preventing smoking or drinking alcohol. Age recognition can also be utilized to ameliorate others image analysis applications.

The research in age estimation is investigated by facial images [9,10]. Facial wrinkles, skin smoothness and other signs are analysed for the estimation of age and gender. But, presence of makeup, anti-aging treatment and wrinkles surgery disrupt age detection from face features. Our goal in this paper is to implement an age and a gender recognition system that ensures comfort and acceptability by the user. We decided to choose venous hand due to various benefits. The key advantage of hand veins is that they are hidden biometric, since they are

under the skin. Also, they are independent of the appearance of the body (hands in gloves, dusty hands, etc.). The research in aging detection from hand vein patterns is very restricted. Little investigations have been concentrated to this topic. This have induced more research for an effective strategy able to classify the age range and the gender of individuals.

The remainder of this paper is organized as follows: Sect. 2 presents our proposed method for age and gender classification from finger vein pattern. Experimental results are exposed in Sect. 3. Finally, we conclude and give some perspectives.

2 Age and Gender Recognition from Finger Vein Patterns

We are interested in implementing a system of age range and gender recognition through finger veins. For this goal, we aim to propose an algorithm able to detect changes related to aging. This motivates to study the signs that can predict the age and the gender of a person. The proposed age and gender classification block diagram is shown in Fig. 1. It contains four main steps: image acquisition, image preprocessing, feature extraction and age/gender classification.

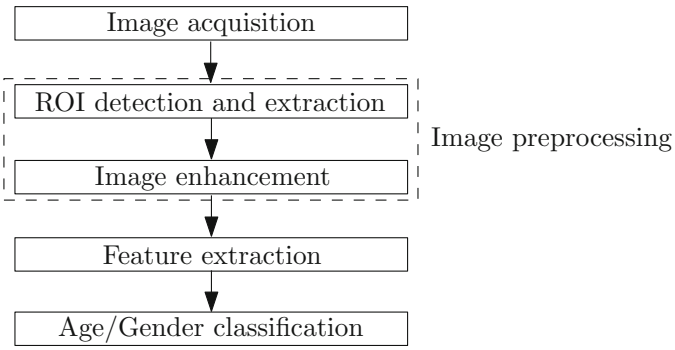


Fig. 1. Proposed age and gender classification block diagram.

2.1 Image Acquisition System

There are two database available for researchers that records personal information (age, nationalities, blood groups, gender, etc.) with finger vein images:

– The UTFVP (University of Twente Finger Vascular Pattern) Database: contains personal information such as age and gender [11]. The number of volunteers who contribute in the elaboration of the database is 60 (73% male and 27% female), in the age range of 19–57.

– The MMCBNU_6000 (Multimedia Lab Chonbuk National University 6000) Database: contains statistical information about nationality, age, gender, and blood group [12]. 100 volunteers (83 men and 17 women) have participate in the construction of the database. Their age varies from 16 to 72 years.

In Fig. 2, we present an example of some samples from MMCBNU_6000 Database and UTFVP Database with different ages and genders. Figures 3 and 4 show the age and the gender distribution of volunteers in MMCBNU_6000 Database and UTFVP Database. The MMCBNU_6000 Database is better than the UTFVP Database, such as it includes more number of volunteers and the quality of postures is better. For this, our experimental study is based on images from MMCBNU_6000 Database.

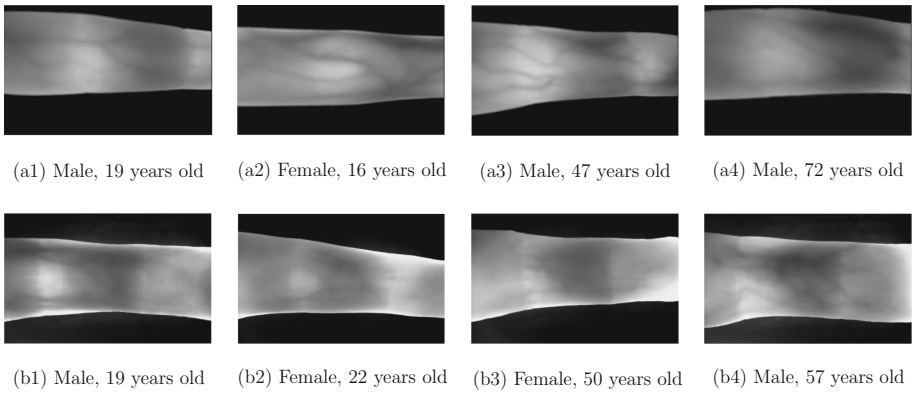


Fig. 2. Samples from MMCBNU_6000 (a1–a4) and UTFVP (b1–b4) with different ages and genders.

2.2 Image Preprocessing

– ***ROI detection and extraction:*** The Region Of Interest (ROI) detection and extraction is an important step in finger vein recognition process. It aims to extract the effective region from an acquired image. In [13], Lu et al. proposed a finger vein ROI localization based on flexible segmentation. This method is composed by the following steps: segmentation, orientation correction, and ROI detection. They investigated a robust segmentation method for MMCBNU_6000 database images against the complicated background and the uneven illumination. An appropriate segmented finger region can lead the finger orientation correction. This later tries to rotate images to the identical angle (horizontal and vertical). Thereafter, ROI detection purpose is to extract the informative region from finger vein. The extracted ROI is shown in Fig. 5.
– ***Image enhancement:*** The main goal of image enhancement is to improve the visual quality and the perception of image. Many techniques have been developed in this field:

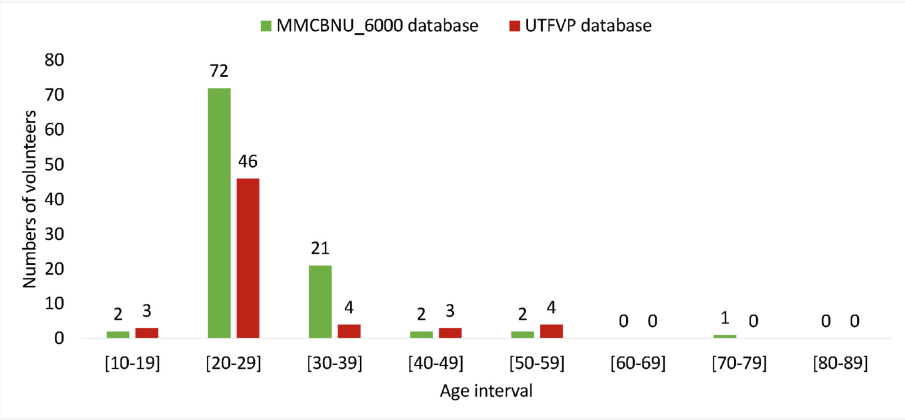


Fig. 3. The age distribution of volunteers in MMCBNU_600 Database and UTFVP Database.

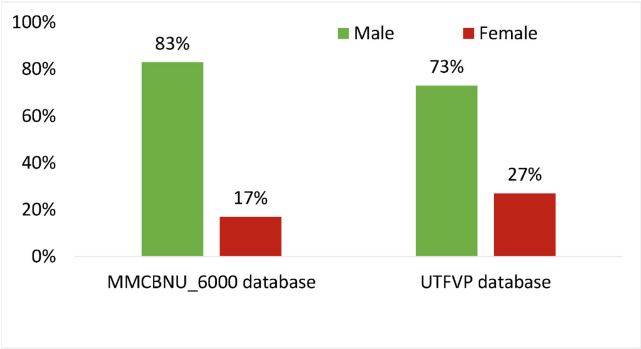


Fig. 4. The gender distribution of volunteers in MMCBNU_600 Database and UTFVP Database.

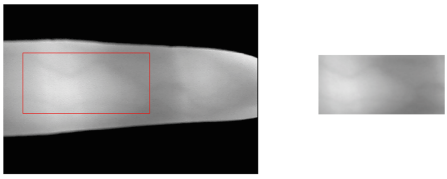


Fig. 5. ROI localization: (a) ROI detection (b) ROI extraction.

- ***Guided Filter based Singe Scale Retinex (GFSSR):*** In [14], Shan proposed an intensity variation normalization for finger vein image enhancement using guided filter based Singe Scale Retinex (GFSSR). Single Scale Retinex (SSR) is based on the hypothesis that an input image I

is considered as the multiplication of the intensity variation IV and the reflectance R, as follow:

$$I(x, y) = IV(x, y) \times R(x, y) \tag{1}$$

In order to remove the illumination from the acquired image, a subtraction operation is used in the logarithmic domain.

$$\log R(x, y) = \log I(x, y) - \log IV(x, y) \tag{2}$$

It is supposed that the illumination is modified clearly and the reflectance is modified slowly, the intensity variation IV can be estimated by a convolution operation of a smoothing filter on the acquired image.

$$IV(x, y) = I(x, y) \times F(x, y) \tag{3}$$

Figure 6 shows a example of an output of guided filter, an illumination image and an intensity variation normalization. The Guided filter is selected owing to its adjustable edge-preserving smoothing ability [15]. It uses a guidance image which can be the input image or another image.

- ***Contrast-Limited Adaptive Histogram Equalization (CLAHE):*** is an histogram equalization method which extends the intensity values on the full intensity range in order to visualize the maximum contrast.

To measure the contrast of the enhanced image, we use the Second Derivative Measure of Enhancement (SDME) [16], which is presented in the following equation:

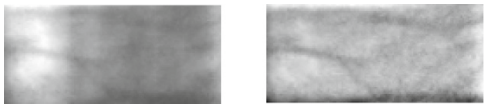
$$SDME = \frac{-1}{k1k2} \sum_{l=1}^{k1} \sum_{k=1}^{k2} 20 \ln \left| \frac{I_{max;k,l} - 2I_{ctr;k,l} + I_{min;k,l}}{I_{max;k,l} + 2I_{ctr;k,l} + I_{min;k;l}} \right| \tag{4}$$

(k1, k2) represent the size of the image blocks. $I_{max;k,l}$, $I_{min;k,l}$ are the maximum and the minimum pixel values in each block. $I_{ctr;k,l}$ is the center pixel value in the block. Figure 7 presents the measure of SDME of the image enhanced by CLAHE and the image enhanced by GFSSR. It is clear that the image enhancement using GFSSR can contribute to high contrast of finger vein image compared to the image enhancement using CLAHE.



(a) (b) (c) (d)

Fig. 6. Intensity variation normalization using GFSSR: (a) Acquired image (b) Output of guided filter (c) Illumination image (d) IV normalization.



(a) SDME = 90.76 (b) SDME = 97.18

Fig. 7. Measure of SDME of different enhancement techniques (a) Image enhanced by CLAHE (b) Image enhanced by GFSSR.

2.3 Feature Extraction

Feature extraction is an essential step in finger vein identification or verification systems. It aims to represent each image by a vector of features in order to distinguish it. Several methods have been proposed for finger vein feature extraction. They are based on geometric approaches [17,18] or/and textural approaches [19,20]. Among texture based approaches, Local Binary Pattern (LBP) is the most attractive feature texture analysis [21]. The mathematical LBP code of each pixel is defined as follows:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, s(x) = \begin{cases} 1 \text{ if } x \leq 0 \\ 0 \; else \end{cases} \tag{5}$$

Where P and R are respectively neighbors number and neighborhood radius. g_c represent central pixel gray level value and g_p is gray level value neighbors, $\{p = 0, 1, 2, ..., P - 1\}$.

3 Experimental Results

In gender and age range classification step, 3 classifiers are used: Support Vector Machine (SVM), k-Nearest Neighbors (KNN) and Artificial Neural Network (ANN). For SVM, a quadratic kernel is investigated, since it presents slightly superior performance than others kernel (polynomial, RBF, linear). For KNN, we use the Euclidian distance with a value of k = 1. For ANN, we consider a Feed Forward Backpropagation Neural Network (FFBPN). The neural network has 256 neurons used for the input layer representing finger vein features. There are 2 or 3 or 4 neurons used for the output layer, depending the number of classes (2 or 3 or 4). The used activation function is the sigmoid function.

For each volunteer, 60 finger vein images have been acquired, from middle (M) fingers, ring (R) fingers and fore fingers (F), from both right and left hands. The acquisition of 6 fingers is repeated 10 times for a total of 6000 images. In order to evaluate the performance of our work, data are divided in two set: training set (70% of data) and test set (30% of data). In our experimental study, we tested the effect of the type of hand (left or right) and the effect of finger type (M, R, F). Table 1 shows the different experimental results for various fingers of left and right hands. Each time, we vary the number of classes and we

Table 1. Age recognition rate.

		Left hand			Right hand		
		F	M	R	F	M	R
2 Classes	SVM	98%	97.67%	98.33%	97.67%	97.33%	98.67%
	ANN	98.33%	98%	97.67%	97.33%	97.67%	98%
	KNN	99.33%	99%	99%	99.33%	99.67%	99.67%
3 Classes	SVM	98.67%	98.44%	98.89%	98.44%	98%	99.11%
	ANN	99.33%	98.67%	98.67%	98.22%	98.89%	98.89%
	KNN	99.56%	99.33%	99.33%	99.56%	99.78%	99.78%
4 Classes	SVM	89%	89.33%	87.83%	89.50%	88.17%	89.5%
	ANN	87.33%	90.33%	86.50%	87.83%	89%	88.5%
	KNN	97%	96.17%	94.83%	96.67%	97.33%	96%

evaluate the response of the age recognition rate. KNN provide better recognition rate compared to SVM and ANN. The best recognition rate is 99.67%, 99.78% and 97.33% for respectively 2, 3 and 4 classes. The description of classes is exposed in Table 2. Each class is characterized by an age distribution and a volunteers number. Middle finger of right hand provides the best results (99.67%, 99.78%, 97.33%) when the classes number is equal to 2, 3 or 4 respectively. The recognition rate for 3 classes is a bit better than 2 classes, due to the small sample size for the third class. We can explain the improvement of recognition rate by the increasement of true positive and true negative for the minority class. We observe that the recognition rate decreases for 4 classes. But, the results remain in good performance. We try also to classify the gender of individuals. Table 3 shows the gender recognition rate by varying the type of the finger and the hand. The best gender recognition rate (98%) is achieved by middle finger of left hand, by KNN classifier. Experimental results demonstrate that age and gender can be detected

Table 2. Classes description.

Classes number	Age distribution
2 Classes	Classe 1: Age in [10..50[
	Classe 2: Age in [50..90[
3 Classes	Classe 1: Age in [10..50[
	Classe 2: Age in [50..70[
	Classe 3: Age in [70..90[
4 Classes	Classe 1: Age in [10..30[
	Classe 2: Age in [30..50[
	Classe 3: Age in [50..70[
	Classe 4: Age in [70..90[

Table 3. Gender recognition rate.

	Left hand			Right hand		
	F	M	R	F	M	R
SVM	84.33%	87.33%	87.67%	86.33%	89.33%	90%
ANN	87.67%	90.33%	88%	84.67%	90%	88.33%
KNN	95.33%	98%	94.33%	92.67%	95.33%	93.67%

from finger veins. However, we are unable to present comparisons between the results obtained by the proposed method and those from the literature, since research in age detection from hand vein is very restricted.

4 Conclusion

In this paper, our research is concentrated on gender and age detection from finger vein. Accordingly, we implement a strategy for age classification with fixed ranges. The experimental results prove that extracted features from finger vein can indicate the gender and the age class. KNN provide better recognition rate for age and gender compared to SVM and ANN. The recognition rate is 98% for gender classification and is 99.67%, 99.78% and 97.33% for respectively 2, 3 and 4 classes in the case of age classification. However, MMCBNU_6000 Database is not suitable and sufficient for age and gender recognition. The majority of age distribution is in [20–29]: 72% of volunteers. So, the future directions of this work is the construction of a hand vein database in order to test and evaluate an age and a gender recognition system. An equal distribution of age and gender volunteers is primordial to elaborate a classification strategy able to define the age and the gender of individuals. The database must take into account different information for person such as gender, age, weight, size, blood type, ambient temperature, physical activity, etc.

References

1. Li, K., Zhang, G., Wang, Y., Wang, P., Ni, C.: Hand-dorsa vein recognition based on improved partition local binary patterns. In: Yang, J., Yang, J., Sun, Z., Shan, S., Zheng, W., Feng, J. (eds.) Biometric Recognition. LNCS, vol. 9428, pp. 312–320. Springer, Heidelberg (2015). doi:10.1007/978-3-319-25417-3_37
2. Randa, B.T., Alima, D.M., Dorra, S.M.: Hand vein recognition system with circular difference and statistical directional patterns based on an artificial neural network. Multimedia Tools Appl. **75**(2), 687–707 (2016)
3. Randa, B.T., Alima, D.M., Dorra, S.M.: A novel biometric system based hand vein recognition. J. Test. Eval. **42**(4), 1–10 (2013)
4. Randa, B.T., Alima, D.M., Dorra, S.M.: A new multimodal biometric system based on finger vein and hand vein recognition. Int. J. Eng. Technol. **4**, 3175 (2013)

5. Wang, Y., Zheng, H.: A preliminary analysis of the aging dorsal hand vein images. In: 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), vol. 2, pp. 271–274. IEEE (2013)
6. Nadort, A.: The hand vein pattern used as a biometric feature. Master Literature Thesis of Medical Natural Sciences at the Free University, Amsterdam (2007)
7. Bains, R.D., Thorpe, H., Southern, S.: Hand aging: patients opinions. Plast. Reconstr. Surg. **117**(7), 2212–2218 (2006)
8. Yang, H., Yang, X., Xu, W., Yao, S., Wei, C., Qu, H., Qu, B.: Research of the types of applicable people and the statistical characteristics of hand vein image. In: SPIE/COS Photonics Asia, p. 92681M. International Society for Optics and Photonics (2014)
9. Thakur, S., Verma, L.: Age identification of facial images using neural network. Int. J. Comput. Sci. Inf. Technol. **3**(3), 4244–4247 (2012). Tech Science
10. Kumar, P.S., Reddy, P.K.K., Kumar, V.V.: Age classification based on features extracted from third order neighborhood local binary pattern. ICTACT J. Image Video Process. **5**(02), 926–931 (2014)
11. Ton, B.T., Veldhuis, R.N.: A high quality finger vascular pattern dataset collected using a custom designed capturing device. In: 2013 International Conference on Biometrics (ICB), pp. 1–5. IEEE (2013)
12. Lu, Y., Xie, S.J., Yoon, S., Wang, Z., Park, D.S.: An available database for the research of finger vein recognition. In: 2013 6th International Congress on Image and Signal Processing (CISP), vol. 1, pp. 410–415. IEEE (2013)
13. Lu, Y., Xie, S.J., Yoon, S., Yang, J., Park, D.S.: Robust finger vein ROI localization based on flexible segmentation. Sensors **13**(11), 14339–14366 (2013)
14. Xie, S.J., Lu, Y., Yoon, S., Yang, J., Park, D.S.: Intensity variation normalization for finger vein recognition using guided filter based singe scale retinex. Sensors **15**(7), 17089–17105 (2015)
15. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Trans. Pattern Anal. Mach. Intell. **35**(6), 1397–1409 (2013)
16. Zhou, Y., Panetta, K., Agaian, S.: Human visual system based mammogram enhancement and analysis. In: 2010 2nd International Conference on Image Processing Theory Tools and Applications (IPTA), pp. 229–234. IEEE (2010)
17. Alima, D.M., Randa, B.T., Krid, M., Dorra, S.M.: Implementation of a fingervein recognition system based on improved gaussian matched filter. J. MAGNT Res. Rep. **2**(4), 251–260 (2014)
18. Yuksel, A., Akarun, L., Sankur, B.: Hand vein biometry based on geometry and appearance methods. IET Comput. Vis. **5**(6), 398–406 (2011)
19. Alima, D.M., Randa, B.T., Dorra, S.M.: A novel finger vein recognition system based on monogenic local binary pattern features. Int. J. Eng. Technol. (IJET) **5**(6), 4528–4535 (2014)
20. Alima, D.M., Randa, B.T., Dorra, S.M.: A new biometric human identification based on fusion fingerprints and finger veins using monoLBP descriptor. World Acad. Sci. Eng. Technol. **78**, 1658–1663 (2013)
21. Randa, B.T., Imene, K.K., Dorra, S.M.: Person identification based on a new multimodal biometric system. Trans. Syst. Sig. Devices **7**(3), 273–289 (2012)

Coupling Event-B/ProB for the Analysis of the Software Architecture Evolution Described in PDDL

Farah Fourati[1(✉)], Mohamed Tahar Bhiri[2], and Riadh Robbana[3]

[1] National Engineering School of Sfax, Sfax, Tunisia
farah.fourati@ymail.com
[2] Faculty of Sciences of Sfax, Sfax, Tunisia
tahar_bhiri@yahoo.fr
[3] National Institute of Applied Science and Technology, Tunis, Tunisia
riadh.robbana@gmail.com

Abstract. The PDDL language is used to formally describe planning problems. It has tools called planners to solve planning problems described in PDDL. Equally, it has plan validation tools to validate the solution plans. Our automatic approach to transform a planning problem written in PDDL to Event-B method allows the use of the correct-by-construction paradigm associated to Event-B formal method. We propose a modeling PDDL for Software Architecture Evolution. Then, we translate our PDDL description in Event-B using our plugin PDDL2EventB. Finally, we reason on the Event-B model generated by the use of ProB including: animation, model checking and deadlock-free.

Keywords: PDDL · Event-B · ProB · Static/Dynamic analysis · Software architecture evolution

1 Introduction

The problem of architectures reconfiguration which essentially consists of adding or removing components in the system. Also it introduces changes in the architecture such as addition and deletion of interfaces. In this work, the problem of Software Architecture Evolution is considered as a planning problem [1] written in PDDL (Planning Domain Definition Language) standard language [2]. The initial state corresponds to the software architecture **before** reconfiguration. The goal state corresponds to the software architecture **after** reconfiguration. The path leading from the initial state to the goal state corresponds to the reconfiguration program. This path corresponds to a sequence of permitted reconfiguration operations. PDDL language has currently few tools able to support the PDDL model such as planners [3] and plan validation tools [4]. In fact, these tools are insufficient for the verification and validation of PDDL descriptions. These descriptions are often difficult to write and include errors even those

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_81

that are considered reference (benchmark problems) for evaluating planners [3]. In this work, we propose a modeling PDDL for Software Architecture Evolution. Then, we translate our PDDL description in Event-B [5] using our plugin PDDL2EventB. Finally, we reason on the Event-B model generated by the use of ProB [6] including: animation, model checking and deadlock-free.

This paper is organized as follows: Sect. 2 presents the PDDL language. Section 3 presents our PDDL description of Software Architecture Evolution. Section 4 describes the Event-B formal method and presents our transformation from PDDL to Event-B via our PDDL2EventB tool. Section 5 demonstrates the reasoning on the generated Event-B models corresponding respectively to PDDL domain and to PDDL problem using proof tools and validation tools associated with Event-B [7].

2 The PDDL Language

PDDL (Planning Domain Definition Language) [2] is a de facto standard for the planning domain in artificial intelligence. PDDL is developed by McDermott and coauthors, and evolved due to the researches of ICAPS (International Conference on Automated Planning and Scheduling) [3].

States of a planning problem that are represented as a conjunction of literals can be propositional literals or first order literal without variable or function. The closed-world assumption used is: any unverified requirement is deemed false. Also, PDDL supports other types of representations such as the numeric representations as well as the temporal representations. A PDDL specification contains two parts presented in two separate files. First, a description of the domain contains: types, predicates and possible actions (See Fig. 1). Second, a description of the problem contains: the objects, the initial state and the goal state (See Fig. 2). Thus, a domain description can be shared among several planning problem. Both the domain file and the problem file are expressed in a like Lisp syntax. The actions are represented by a name, they are composed of a list of preconditions and a list of effects: pre/post-conditions of PDDL actions. The preconditions list contains a conjunction of literals indicating what must be true **before** the execution of the action. The list of effects is a combination of positive and negative literals indicating what it changes **after** the execution of the action. The positive literals should be added to the state if they are not already there, while the negative literals are deleted.

A planner [3] is able to find solutions for planning problems. From a formal PDDL description of a planning problem (initial state, goal state and actions), a planner calculates a sequence of actions whose execution is changing the modeled system from the initial state to a goal state.

Planners are not equal vis-à-vis the constructions offered by PDDL. In fact, planners try to track different versions of PDDL. In addition, different categories of planners can be identified such as sequential satisficing track, sequential agile track, sequential optimal track, sequential multicore track and temporal satisficing track.

```
(define (domain ArchReconfigurable) ( :requirements :strips :typing :adl)
( :types system component_type component port connector_type
connector role )
( :predicates (exist_component ?c - component)
(exist_connector ?c - connector)
(contains_component ?s - system ?c - component)
(contains_connector ?s - system ?c - connector)
(has_component_type ?c - component ?t - component_type)
(has_connector_type ?c - connector ?t - connector_type)
(type_has_port ?t - component_type ?p - port)
(type_has_role ?t - connector_type ?r - role)
(bound ?c - component ?p - port ?co - connector ?r - role))
( :action create_component :parameters ( ?c - component ?s - system)
:precondition ( and(not (exist_component ?c)) (exists ( ?t - component_type )(
has_component_type ?c ?t)) ) :effect (and (exist_component ?c)
(contains_component ?s ?c)) )
( :action create_connector ...)
( :action remove_component :parameters ( ?c - component ?s - system)
:precondition (and(contains_component ?s ?c) (forall( ?co - connector ?p - port ?r
- role)(not(bound ?c ?p ?co ?r))) ) :effect (and (not(exist_component ?c))
(not(contains_component ?s ?c))) )
( :action remove_connector ....)
( :action detach :parameters ( ?c - component ?co - connector ?p - port ?r - role)
:precondition (bound ?c ?p ?co ?r) :effect (not(bound ?c ?p ?co ?r)))
( :action attach :parameters ( ?s - system ?c - component ?co - connector ?p -
port ?r - role) :precondition (and (contains_component ?s ?c) (contains_connector
?s ?co) (not(bound ?c ?p ?co ?r)) (exists ( ?t - component_type) (imply
(has_component_type ?c ?t) (type_has_port ?t ?p)) ) (exists ( ?t - connector_type)
(imply (has_connector_type ?co ?t) (type_has_role ?t ?r)) ) (forall ( ?cc - component
?pp - port) (imply(contains_component ?s ?cc) (not(bound ?cc ?pp ?co ?r))) )
(forall ( ?con - connector ?rr - role) (imply(contains_connector ?s ?con)(not(bound
?c ?p ?con ?rr))) ) )
:effect (bound ?c ?p ?co ?r)))
( :action add_type_port :parameters ( ?t - component_type ?p - port)
:precondition (and(not(type_has_port ?t ?p)) (forall ( ?c - component)
(imply(has_component_type ?c ?t)(not(exist_component ?c)))) ) :effect
(type_has_port ?t ?p) )
( :action add_type_role ....)
( :action remove_type_port :parameters ( ?t - component_type ?p - port)
:precondition (and(type_has_port ?t ?p) (forall ( ?c - component)
(imply(has_component_type ?c ?t)(not(exist_component ?c)))) ) :effect (not
(type_has_port ?t ?p)) )
( :action remove_type_role ...)
```

Fig. 1. ArchReconfigurable domain

```
(define (problem Migration)
( :domain ArchReconfigurable)
( :objects ClientServer1 ClientServer2 - system
Client Server - component
Conn - connector
Client_type - component_type
Server_type - component_type
Conn_type - connector_type
request service - port
requestor servicer - role )
( :init (exist_component Client) (exist_component Server) (exist_connector
Conn) (contains_component ClientServer1 Client) (contains_component
ClientServer1 Server) (contains_connector ClientServer1 Conn)
(has_component_type Client Client_type) (has_component_type Server
Server_type) (has_connector_type Conn Conn_type) (type_has_port Client_type
request) (type_has_port Server_type service) (type_has_role Conn_type
requestor) (type_has_role Conn_type servicer) (bound Client request Conn
requestor) (bound Server service Conn servicer) ) ( :goal (and (exist_component
Client) (exist_component Server) (exist_connector Conn)
(contains_component ClientServer2 Client) (contains_component
ClientServer2 Server) (contains_connector ClientServer2 Conn) (bound Client
request Conn requestor) (bound Server service Conn servicer) )) )
```

Fig. 2. Migration problem of client-server system

3 Modeling of Software Architecture Evolution in PDDL

Barnes and co-authors propose a method allowing to modelize reconfiguration of software architecture problem with PDDL [1]. Each architectural element type (such as component types, connector types, components, connectors, ports, roles and systems) is translated to an object type. Relations between the elements (including attachments) is translated into facts due to predicates. Each operator of reconfiguration is translated by a PDDL action or trajectory constraints. Furthermore, each architectural element is translated into an object. Finally, the original architecture is translated to the initial state "init" and the target architecture is translated into the goal state "goal".

We propose a PDDL domain for software architecture evolution. To achieve this, we reuse and enriche the work presented in [8]. The work [8] considers that any type of component (connector respectively) can admit only one instance. Furthermore, we promote pre/post-conditions of PDDL actions using the "new" possibilities offered by the PDDL language such as the negation operator and operators of quantification forall and exists. A PDDL description contains two parts: the domain description and the problem description.

The ArchReconfigurable domain factored a class of reconfiguration software architecture problems: migration of a software architecture, substitution of a component, adding and removing components (See Fig. 1). For example, the problem of migration of a software architecture from an environment to another

can be modeled in PDDL by using our ArchReconfigurable domain. For this, and for simplicity reasons, we propose to study the migration of a client-server architecture from a system to an another (See Fig. 2). Figure 3 gives a solution plan for the description of planning problem. It is probable, due to the characteristics of the problem, that in some cases the problem will be better adapted to some planners and not to others. We have tested our program by IPC-14 planners such as NuCeLaR.

```
(DETACH CLIENT CONN REQUEST REQUESTOR)
(DETACH SERVER CONN SERVICE SERVICER)
(REMOVE_CONNECTOR CONN CLIENTSERVER1)
(CREATE_CONNECTOR CONN CLIENTSERVER2)
(REMOVE_COMPONENT CLIENT CLIENTSERVER1)
(CREATE_COMPONENT CLIENT CLIENTSERVER2)
(ATTACH CLIENTSERVER2 CLIENT CONN REQUEST REQUESTOR)
(REMOVE_COMPONENT SERVER CLIENTSERVER1)
(CREATE_COMPONENT SERVER CLIENTSERVER2)
(ATTACH CLIENTSERVER2 SERVER CONN SERVICE SERVICER)
```

Fig. 3. Plan Migration of client-server system

4 From PDDL to Event-B

4.1 Modelling in Event-B

Event-B formal method [5] allows a formal development process based on the concept of refinement with mathematical proof. The language of Event-B offers possible ways to formulate both data and treatment. Data is described by a set-theoretical language and a first order predicate logic. Whereas treatments are described by a simple action language. Set-theoretical language of Event-B includes: predicate, set, relation and function with their operators. There are five actions in Event-B: deterministic assignment ($:=$), non-deterministic assignment ($:\in$), non-deterministic assignment that satisfies a predicate ($:|$), parallel assignment ($\|$) and the null action (skip).

In order to gather again data and treatments, the Event-B language supports two syntactic constructions: a CONTEXT and a MACHINE. The potential relationships between contexts and machines of an Event-B model are: SEES, EXTENDS and REFINES. A machine can be "refined" by another one. A context can be "extended" by another one. Both relationships SEES and REFINES are transitive. Moreover, a machine can "see" one or several contexts. An Event-B context contain the static parts of the model, namely carrier sets, constants, axioms, and eventual theorems. Whereas an Event-B machine contain the dynamic parts of the model, namely variables, invariants, eventual theorems, eventual variants, and events. A model can contain contexts only, machines only, or both.

An event is composed of a guard and a body. When the guard is satisfied, the event can be enabled. When the guards of several events are satisfied at

the same time, the choice of the event is non-deterministic. Any event in Event-B modelizes a labelled transition system. However, there is a mandatory event named INITIALISATION which initializes the Event-B machine.

Refinement allows us to build models gradually, which reflects more precised models. The refined machine is called the abstract machine whereas the refinement is called the concrete machine. Each abstract event (event of the abstract machine) can be refined by one or more concrete events (event of the concrete machine). Furthermore, the invariants are accumulated during refinements.

4.2 The PDDL2EventB Tool

We have conceived, realized and tested a tool called PDDL2EventB. Our PDDL2EventB plugin, integrable into the Eclipse platform, built components of two kind CONTEXT and MACHINE from a file that contains a "domain" construction and zero or more "problem" construction. Our PDDL2EventB tool generates as output a textual representation of Event-B system accepted by Camille editor of Event-B [5]. To achieve this, we have successfully used the MDE approach and available plugins of EMF Eclipse and oAW platform (Xtext and Xtend). The description of ArchReconfigurable domain (See Fig. 1) and Migration problem (See Fig. 2) is translating respectively in abstract Event-B model and refined Event-B model. For example, the abstract model generated is presented as follows:

CONTEXT C_ArchReconfigurable
SETS

component_type connector_type system component connector port role

MACHINE M_ArchReconfigurable
SEES C_ArchReconfigurable
VARIABLES

exist_component contains_component has_component_type
exist_connector contains_connector has_connector_type
type_has_port type_has_role bound

INVARIANTS

inv1 : $exist_component \in component \rightarrow BOOL$
inv2: $contains_component \in system \times component \rightarrow BOOL$
inv3: $exist_connector \in connector \rightarrow BOOL$
inv4 : $contains_connector \in system \times connector \rightarrow BOOL$
inv5 : $has_component_type \in component \times component_type \rightarrow BOOL$
inv6: $has_connector_type \in connector \times connector_type \rightarrow BOOL$
inv7 : $type_has_port \in component_type \times port \rightarrow BOOL$
inv8: $type_has_role \in connector_type \times role \rightarrow BOOL$
inv9: $bound \in component \times port \times connector \times role \rightarrow BOOL$

EVENTS
Event $evt_create_component \mathrel{\widehat{=}}$
any

$c\ s$

where

$grd1 : c \in component$
$grd2 : s \in system$
$grd3 : exist_component(c) = FALSE$
$grd4 : \exists t \cdot t \in component_type \land has_component_type(c \mapsto t) = TRUE$

then

$act1 : exist_component(c) := TRUE$
$act2 : contains_component(s \mapsto c) := TRUE$

end
.....

Element Name	Total	Auto	Manual	Reviewed	Undischarged
M_ArchR...	**56**	**56**	**0**	**0**	**0**
INITIALISA...	18	18	0	0	0
inv_exist_c...	3	3	0	0	0
inv_exist_c...	3	3	0	0	0
inv_contai...	3	3	0	0	0
inv_contai...	3	3	0	0	0
inv_has_co...	1	1	0	0	0
inv_has_co...	1	1	0	0	0
inv_type_h...	3	3	0	0	0
inv_type_h...	3	3	0	0	0
inv_bound	3	3	0	0	0
evt_create...	4	4	0	0	0
evt_create...	4	4	0	0	0
evt_remov...	4	4	0	0	0
evt_remov...	4	4	0	0	0
evt_detach	2	2	0	0	0
evt_attach	8	8	0	0	0
evt_add_ty...	3	3	0	0	0
evt_add_ty...	3	3	0	0	0
evt_remov...	3	3	0	0	0
evt_remov...	3	3	0	0	0

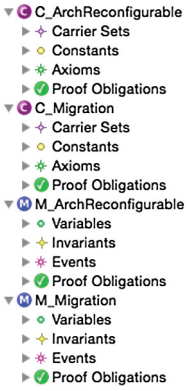


Fig. 4. Proof obligation rules discharged

In fact, a PDDL "domain" is translated into an abstract model described by CONTEXT and MACHINE Event-B that respectively have the static and the dynamic parts of the domain. The C_ArchReconfigurable context includes translations of the types and constants from the ArchReconfigurable domain. The Event-B machine M_ArchReconfigurable includes translations dynamic predicates and actions from the ArchReconfigurable domain. Also a PDDL problem is translated into a refined Event-B model with a C_Migration context and M_Migration machine. The C_Migration concrete context that extends C_ArchReconfigurable includes the translation of objects from Migration problem. The M_Migration concrete machines that refines M_ArchReconfigurable.

In the M_Migration machine, the initial state "init" is translated by INITIALISATION event that refines the M_ArchReconfigurable abstract machine and introduce a new event "GOAL" in order to translate the state "goal" of the planning problem.

Note that, our tool allows to translate domain construction containing types (atomic or hierarchical), typed constants, predicates and actions. It also allows to translate problem construction. The formulas accepted by PDDL2EventB tool are: the assertion, the negation operator, the equality operator, the implication operator and quantification operators (forall and exists). The activity of proof and validation are performed through the Rodin platform [5].

5 Consistency of Event-B Models Generated

The PDDL2EventB tool produces Event-B models from PDDL descriptions. We can make formal reasoning on the Event-B models generated. These reasonings concern the abstract Event-B model as well as the refined Event-B model. Such reasoning aimed at formal verification of several types of properties. These include safety properties, deadlock-free and liveness properties.

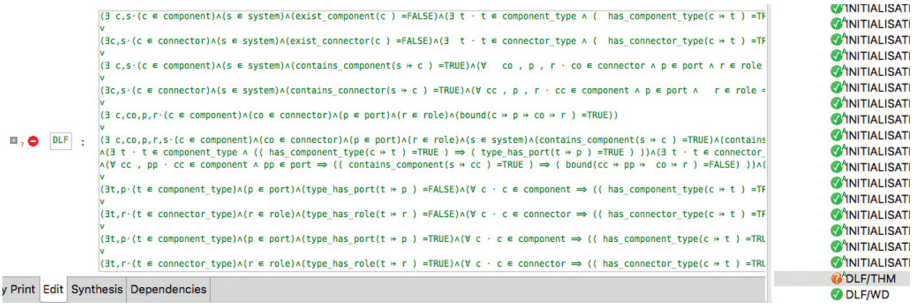


Fig. 5. DLF theorem

5.1 Reasonings on the Obtained Abstract Model

The Event-B abstract model produced by our PDDL2EventB tool corresponds to the domain PDDL. It comprises Event-B context and Event-B machine. According to our experiments the Event-B abstract model meets the safety properties: the invariant is established by the INITIALISATION event and it is preserved by the others events. For example, the Event-B model stemming from the ArchReconfigurable domain (See Fig. 1) is presented in Sect. 4 and it's discharge proof obligations are given in Fig. 4. In general, the deadlock-free property is not formally verified on Event-B abstract models produced by our PDDL2Event-B tool, particularly for models with multiple events (greater than 10). For example, the theorem expressing the property deadlock-free of the abstract model stemed from ArchReconfigurable domain is not automatically proven (See Fig. 5). But using

ProB and reducing the size of abstract sets, the ProB model checker does not detect deadlocks. Moreover, through ProB, we verified that all events can be visited. Finally, liveness properties specific to each PDDL domain can be described using the great possibilities of LTL^e language (extended Linear Temporal Logic) [6] and formally verified by ProB.

5.2 Strengthening Invariant

The abstract model in Event-B stemed from the PDDL domain includes the invariant relating to typing properties of Event-B variables corresponding to the dynamic predicate of PDDL domain concerned. Such invariant can be strengthened by properties describing constraints within or between variables. These constraints could be established by animation of Event-B models using ProB. Also, they can be obtained by further analyzing the specifications related to treated PDDL domain. The addition of these constraints can promote mathematical proof of deadlock-free and the verification and the validation of Event-B events corresponding to PDDL actions.

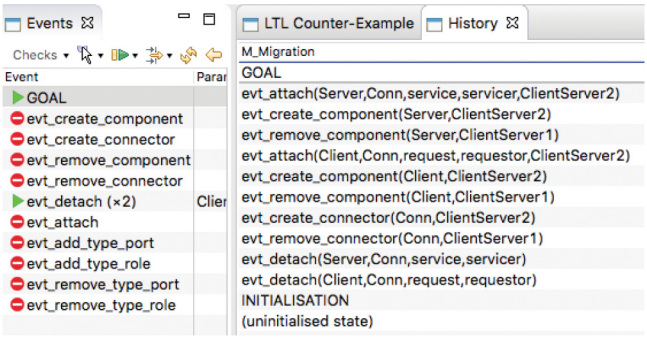


Fig. 6. Animation: migration of client-server architecture with ProB

5.3 Reasoning About Refined Model Obtained

The refined model generated by our PDDL2EventB tool is expected to produce an automatic refinement of the abstract model. According to the experiments and to the use of the internal and the external provers of the Rodin platform, the refinement relation between the two models produced by PDDL2EventB is formally proven. In fact, the problems encountered are related to the determination of the INITIALIZATION event. Through the use of ProB animator, we can automatically generate plan solution or validate plan solution. For example the Fig. 6 validates plan solution related to the problem of migration of a software architecture (See Fig. 3). Unlike the VAL tool associated with PDDL, the animation with ProB allows locating errors of an incorrect plan solution.

6 Conclusion

We propose, in this work, to study the opening of PDDL description on the Event-B formal method. We experimented our transformation approach from PDDL to Event-B on the Software Architecture Evolution domain. The domain was checked and validated in Event-B. In fact, the modeling activity aims to establish formal Event-B model. As for the activity of proof, it aims the correctness of these models. The formal Event-B method offers correctness criteria such as: the establishing of the invariant by the INITIALISATION event, the preserving of the invariant by other events of the machine and the correctness of the refinement. An Event-B model proved is not necessarily valid. The Rodin platform providing some validation tools such as: animator and ProB model checker [9]. We illustrate the use of ProB model checker to validate a priori correctness of Event-B models generated.

As perspective for this work, we could extend PDDL2EventB tool by new facilities related to numeric and temporal representations of new versions of PDDL.

References

1. Barnes, J.M., Pandey, A., Garlan, D.: Automated planning for software architecture evolution. In: 2013 IEEE/ACM 28th International Conference on Automated Software Engineering (ASE), pp. 213–223. IEEE (2013)
2. McDermott, D., Ghallab, M., Howe, A., Knoblock, C., Ram, A., Veloso, M., Weld, D., Wilkins, D.: Pddl-the planning domain definition language (1998)
3. www.icaps-conference.org (2016)
4. Howey, R., Long, D., Fox, M.: Val: automatic plan validation, continuous effects and mixed initiative planning using PDDL. In: 16th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2004, pp. 294–301. IEEE (2004)
5. Abrial, J.-R.: Modeling in Event-B: System and Software Engineering. Cambridge University Press, Cambridge (2010)
6. Plagge, D., Leuschel, M.: Seven at one stroke: LTL model checking for high-level specifications in B, Z, CSP, and more. Int. J. Softw. Tools Technol. Transf. **12**(1), 9–21 (2010)
7. Abrial, J.-R., Butler, M., Hallerstede, S., Hoang, T.S., Mehta, F., Voisin, L.: Rodin: an open toolset for modelling and reasoning in Event-B. Int. J. Softw. Tools Technol. Transf. **12**(6), 447–466 (2010)
8. Méhus, J.-E., Batista, T., Buisson, J.: ACME vs PDDL: support for dynamic reconfiguration of software architectures. arXiv preprint arXiv:1206.0122 (2012)
9. Voisin, L., Abrial, J.R.: The rodin platform has turned ten. In: Ait Ameur, Y., Schewe, K.D. (eds.) ABZ 2014. LNCS, vol. 8477, pp. 1–8. Springer, Heidelberg (2014). doi:10.1007/978-3-662-43652-3_1

Trust Intrusion Detection System Based on Location for Wireless Sensor Network

Hela Maddar(✉), Wafa Kammoun, and Habib Youssef

PRINCE Research Laboratory, Sousse University,
Hammam Sousse, Sousse, Tunisia
maddar.hela.si@gmail.com, wafa_kammoun@yahoo.fr,
habib.youssef@fsm.rnu.tn

Abstract. Despite the rapid development of WSNs and its increasing integrations in different areas, they remain very vulnerable to attacks because of their limitation of resources and their brittle use. Several trusted works have been developed for this purpose. Despite the robustness of the same trust model from literature, they do not take into account the most attacks that threaten wireless sensor networks. Similarly, they do not take into account the different layers of the OSI model. View all these limitations, we have created a new model, a very effective one, which takes into account most of the attacks that can take place in different layers of the OSI model.

Keywords: WSNS · Trust · IDS · Direct trust · Indirect trust · Risk · TDOA · Watchdog

1 Introduction

The trust management is essential when we consider the wireless sensor networks, because of their distributed nature and their non-renewable batteries. Strong physical constraints make very difficult the implementation of the trust management mechanisms for this type of network [1]. Despite the robustness of some works already studied, a lack of confidence rate is always noticed. Likewise the majority of existing studies in the literature are a single layer which does not allow the support of most malicious attacks which make inefficient network. Attacks can target multiple layers of the OSI model. For that we create a trust system which takes into account the different attacks that can take place on the network at different network layers to form an efficient model very trusty that resists against the various attacks threatening wireless sensor networks. This paper is organized as follows: Sect. 2 presents the related work, Sects. 3 and 4 describe our proposed scheme, Sect. 5 presents simulation results, and finally Sect. 6 concludes the paper.

2 Related Work

For a long time, trust has been the focus of researchers. It started in social sciences where trust between humans was studied. From 2004, trusts and reputation systems have been adapted for WSNs [2]. These systems must be light enough to ensure proper

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_82

operation without harming the functionality of the systems [3]. The first trust management system used for WSNs is the Beta reputation system that has been proposed by Josang & Ismail in 2002 [4]. Second work of trust management used for WSNs is RFSN (Reputation-based Framework for high integrity Sensor Networks) [5]. In this work, Ganeriwal and Srivatava used the work Josang & Ismail presented in 2002 [4] based on the probability beta function in the creation of their trust models for wireless sensor networks. It's the first model of trust and reputation that is exclusively developed for wireless sensor networks [2]. Authors in ATSAN (Agent-based Trust Model in Wireless Sensor Networks) [2] propose an agent model based on trust to detect malicious nodes. ATSAN runs in the middleware of each agent node that uses the watchdog scheme to monitor the behavior of nodes in the radio range. Unlike RFSN [5] using a single watchdog for all nodes. Some researches use a security mechanism to ensure confidence like PLUS (Parameterized and Localized Trust Management Scheme) [6]. Another model specially developed for WSN is the GTMS (Based Trust Management Group) [7]. In this model, only one trust value is calculated for each group. If this value goes under the limit, the whole group will be considered untrusted. Many researches focus on data transmission evaluation like DFDI (Distinguish Forged Illegal Data of nodes from data of legal innocent nodes) [8]. Recently, some hybrid models integrate both confidence of data and communication like in [9, 10].

Despite the performance that the existing models present like [9–11] they do not take into account the majority of attacks that can take place on the network. A malicious node can be seen as honest and flood the network with false recommendations. This is due because the most of trusted model take into account only one layer of the OSI model. Also they are unable to ensure the true identity of nodes which can caused great damage to any trust system. View all these limitations and inadequacies that WSNs are suffering, we tried to create a very strong and effective trust model that resist against the majority of attacks threatening WSNs. It ensure verification of the true location of node using a geographical location mechanism and the right data transfer using watchdog mechanism. In the section below we detail our new effective trust model.

3 Description of Our Model

3.1 Hierarchical Detection Model

In our model, we work in a hierarchical network. The collection station is considered trusted. Our intrusion detection system fact on 3 levels: At the first level sensor nodes are IDS agents which are responsible for monitoring their neighbors and send all malignancies detected to their cluster head 'CH'. To do this, a set of rules have been proposed to detect any malice from neighbors (presented on Sect. 4). In the second level, the cluster heads are responsible for assessing the level of trust of nodes in their area that are based on received recommendations and to send a full report to the SB which must be aware of any changes in the network. At the highest level, each CH must monitor its CHs neighbors and transfer any malice detected to the base station basing on same of proposed rules. The base station collect the recommendations

received from different CH and classified their behaviors. In the following, we describe the different metrics that will be used to assess the behavior of nodes. To note, the base station is responsible to communicate the necessary data to different nodes concerning their neighbors. The communication with the base station is always assumed secure.

3.2 Metric Used to Monitor the Behavior of Nodes

Conformity of Radio Power. On network, the signal strength may change same time due severe conditions like humidity or hot weather. In our model, we can check compliance of radio power value. Responsible node should compare the value of RSII received with initial value of signal power recorded on its database. If the value of RSII is higher than initial value, the communication will be rejected and the behavior will be considered abnormal. The received value must satisfy this rule:

$$RSIIi(N) \leq RSII_{initial} \tag{1}$$

Mechanism Watchdog. Each packet is broadcasted in such way that is received by the relevant receiver and by some neighbors (within range of the radio signal). A node can activate its IDS (Intrusion Detection System) agent and monitor packets neighbors.

Geographic Location of Sensor Nodes. In our model we implement TDOA (Time Difference Of Arrival) method [12] that is based on the difference between the arrival time of a transmitter signal to two receivers. By combining information obtained from several pairs of sensors, we can determinate the exact position of nodes. The TDOA between the elements of a pair of receiver's i, j is expressed as the formula below:

$$TDOA_{ij} = t_i - t_j = \frac{1}{c}\left(d_i - d_j\right) = \frac{d_{ij}}{c} \tag{2}$$

Where c is the speed of wave propagation in a given site and d is distance of each path.

Packets Successfully Transmitted Rate. The number of message transmitted by a node during its transmission slot should contain messages received from its neighbors and its own message which includes its observations. Any difference involve a detect of malice. Note, the comparison of number of messages received and transmitted is performed using the watchdog mechanism.

Nonce Value. For each communication between two nodes A and B, two nonce values are accorded. For each received of new message from a node, the receiver node must maintain its nonce updates by incrementing its value by one.

Checking The Amount of Energy. A node with a depleted energy can be exploited by the fake nodes. A check of the remaining amount of energy is required. A node with a 5% lower energy should send a request to its CH to declare its crucial energy state. Subsequently the cluster head informs the neighboring nodes to remove the node concerned from their trust bases.

Time Slot. Our model is based on the multiple access scheme in time allocation or TDMA (Time Division Multiple Access) which is used to divide time into intervals (time-slot) allocated to each node. Thus, only one node has the right to access the channel. This node must transmit its data during its intervals of time.

Information Table. Each node incorporates in its base an information table which comprises the different metrics mentioned above like different identities of its neighboring nodes. The value of the radio power strength equivalent to different neighbors, the nonce value corresponding, number of packets received by the node i at time t, the number of packets transmitted by this node i, the true position of the various neighbors and the estimated position calculated by TDOA geographic location method, trust value of its neighbor node. Finally, the rate of energy remaining. A cluster head must have an information table which integrate all nodes in its coverage area.

3.3 Establishment of the Trust

In our model, the establishment of trust is based on direct observation, indirect observation, updating, aging and risk. The calculation of trust is based on direct and indirect observation. These observations affect the decision of allowing a node to enter or refusing to communicate with another node. Direct observation is based on a single node observation about its neighbor's nodes. Indirect observation is based on trust recommendation received from neighbors nodes recorded on cluster head database. Aging and update are factors that ensure efficiency and robustness of our model. The risk in our work is an important factor that can detect false alarms about legitimate nodes.

3.4 Trust Value Fixed by Our Model

Initially, each node is trusted and has a trust value equal to 1. In our model, a confidence threshold is set to 0.5. A node is considered trusted if its trust value is superior or equal to 0.5. The trust value is the same for each success transmission. Each time the node observes an untrusted neighbor node, it performs directly a trusted decrease of the confidence value of the node by 0.1. It's the direct observation. After minimizing the trust value of this malicious node, the neighbor node sends a message to its cluster head to update its confidence table regarding this malicious node. If the confidence value record in the cluster head database will be inferior or equal to 0.4, this malicious node will be excluded from the network. Each time a cluster head detect an untrusted node, it removes it from its table and sends a broadcast message to all nodes to exclude it. Then the cluster head sends a message to inform the sink about the change.

3.5 Trust Calculation Process

To update trust value in case of maliciousness detected by a neighboring node, a direct trust value should be updated. The direct trust value is defined by the following mathematical expression at a given moment t [11]:

$$T_i(j,t) = T_i(j,t-1) - x \tag{3}$$

Where x = 0 if there is no malice, x = 0.1 in cases of malice detection and x = T_i (j, t − 1) in cases of lack of energy or a risk equal to 1. The expression T_i (j, t − 1) represents the direct trust value of node j in base of node i at t − 1. This value can be changed after a transaction at time t. If a transaction is performed without any failure the confidence value remains the same and x is set to 0. The value of indirect trust is defined by the following mathematical expression [11]:

$$T_{CHi}(j,t) = T_{CHi}(j,t-1) - \sum_{i'=1}^{n'} x \tag{4}$$

Where $T_{CHi}(j,t)$ is the node j trust value recorded in the cluster head database. It represents the value of indirect observation, which will be subsequently transferred to the nodes that want to interact with j at time t and is not in direct communication with it. The value n represents all the nodes in direct interaction with j and sent a negative recommendation to cluster head in a specific time t.

4 Attacks Detection Rules

Selective Forwarding Attack. In selective forwarding attack, a node stops transmitting all messages that must be transferred. A malicious node seeks through this nearly normal behavior to provide as honest node on network. To detect such attacks, we compare the number of messages sent and received by the same node using watchdog mechanism.

```
If received paquets Node_ID - Send paquets Node_ID +1>0
Send_vote_message_alert (Node_ID, attack type)
End
```

Fabrication Messages Attack. On network, malicious nodes can seek to produce forge transmitted messages. To limit this type of attack we use watchdog mechanism. A check of the number of received and transmitted message is always performed.

```
If received paquets Node_ID - Send paquets Node_ID +1<0
Send_vote_message_alert (Node_ID, attack type)
End
```

Sinkhole Attack. A node that seeks to provide as sink node changes its radio power and impersonates the collection station. To limit this kind of attack, a verification of the true position of the sender is performed by neighboring nodes using the TDOA algorithm. The estimation value will be compared with the true position of the collecting station. Similarly, a check of the RSII value is required.

```
If Estimation Posttion Node_ID.xd <> True Position Sink
Node_ID.xd && Estimation Position Node_ID.yd <> True
Position Sink Node_ID.yd
If RSII Node_ID <> RSII Sink_Node_ID
Send_vote_message_alert(Node_ID, attack type)
End …
```

Blackhole Attack. The attacker tries to falsify routing information in order to force the passage of data by itself. Thereafter, it stops transfer of data creating a sort of impasse in the network. To detect such attacks, we check at first compliance of radio power. If there is a radio power Compliance, a verification of number of messages transferred by the node in question is required, if the number of forwarded message is less than a certain threshold equal to 90%, a malice will be detected. This check is performed using watchdog mechanism. In the third place, a verification of node position is performed using the TDOA approach. The order of verification is not randomly chosen, in fact, this order allows us to minimize consumption of energy and minimize the error rate that can be caused by distances calculations.

```
If RSII Node_ID > RSII initial Node_ID
If Paquets Transferred Node_ID < threshold paquet
transferred //(at fixed time slot, with threshold=90%)
If True Position Node_ID.xd<> Estimation Position
Node_ID.xd&& True Position Node_ID.yd <> Estimation
Position Node_ID.yd
Send_vote_message_alert(Node_ID, attack type)
End …
```

Sybil Attack. It's multiple identity attack. In this type of attack the attacker often changes its identity by stealing every time an identity of innocent node. To detect this attack we compare at each receiving of new message the location corresponding to the identity received with the location calculated by neighboring nodes.

```
If True Position Node_ID.xd <> Estimation Position
Node_ID.xd && True Position Node_ID.yd <> Estimation
Position Node_ID.yd
Send_vote_message_alert(Node_ID, attack type)
End
```

Usurpation Attack. To detect usurpation attack, we check firstly if node was sending their data in their right time slot. Secondly, we verify the location corresponding to the receiver node if it is equal to location calculated by the neighbor's node.

```
If Time_send_message _Node_ID <>Time_slot_Node_ID
If True Position Node_ID.xd <> Estimation Position
Node_ID.xd&& True Position Node_ID.yd<> Estimation
Position Node_ID.yd
Send_vote_message_alert(Node_ID, attack type)
End …
```

Hello Flood Attack. This type of attack uses a powerful radio signal at long range which seeks to flood part of the network causing false lists of neighbors. To detect this attack a verification of RSII value will be performed.

```
If RSII Node_ID > RSII initial Node_ID
Send_vote_message_alert(Node_ID, attack type)
End
```

Replication Attack. Through replication attack, a node seeking to deplete a node resources by the replication of old messages. To detect this attack, a check of the value of nonce is performed after each delivery of new message.

```
If nonce value msg (t) <= nonce value msg (t-1)
Send_vote_message_alert(Node_ID, attack type)
End
```

Desynchronization Attack. An attacker tries using a desynchronizing attack to break connection between two nodes communicating on the network by sending false information to force them to go out from synchronization. To limit this kind of attack, a check of behavior of source announcing false recommendation is made. If the number of wrong recommendation reported from node is superior or equal than 3 and the

number of bad recommendations marked by all neighboring nodes is equal to 0 for a specific node, a desynchronisation attack will be directly detected and a risk value will be incremented by one knowing that initially all risk values are initialized to 0.

Detection of False Alarms. Each detection of desynchronization attack, a risk value will be increased by one. If the risk value reaches 1 for a network node. The trust value of this node will be reset to 0.

Collusion Attack. In collusion attack, an adversary sends its signal when it listens to a transmission from legitimate nodes to cause interference. Thanks to TDMA (Time Diffusion Multiple Access), nodes should send their data during their time slot. Otherwise a collusion attack will be detected.

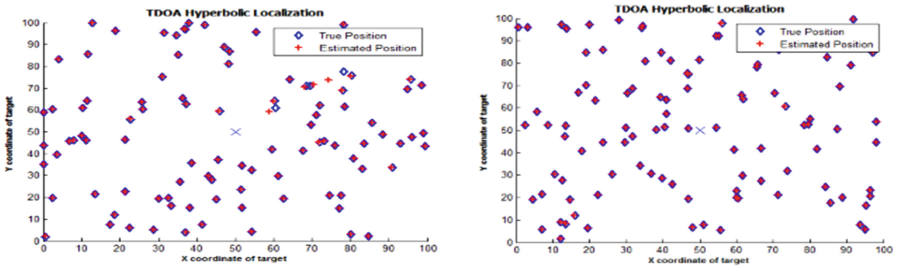
5 Experimentation

To evaluate the features of our model, we used the MATLAB simulator. We created a 100 nodes topology on an area of 100*100 m with energy equal to 0.1 J. We used the LEACH routing protocol [13]. The objective of our simulations is to study the effect of each attack on the network. To evaluate the performance of our model several metrics are analyzed: detection attacks rate, the false positive rate and the rate of false negatives.

5.1 Checking of the Geographical Location of Nodes Through TDOA

Attacks such as Sinkhole attack, Sybil attack, etc. require verification of the identity of the node which is not possible in normal WSNs. For that we implemented a geographic location model to ensure the real identity of nodes.

From the above Fig. 1a and b, we can clearly determine that from a number of synchronized neighbor equal to 5 we can get the correctly location of nodes with a percentage equal to 100% and 0 risk error.



(a) Case of four synchronized neighbors. (b) Case of five synchronized neighbors.

Fig. 1. Cases of four and five synchronized neighbors.

5.2 Hello Flood Attack Detection

In our work, this attack has been implemented as a node that generates a high signal strength and replicates old connections of messages in order to deplete the energy of receiving nodes. To prove the robustness and efficiency of our model, we have disposed our network to a varied number of Hello flood attacks (See Fig. 2).

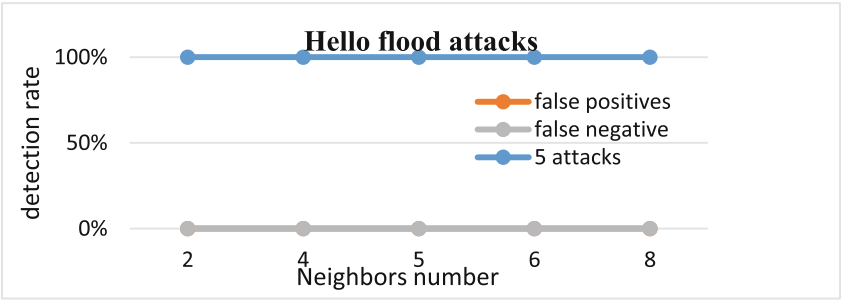


Fig. 2. Our network faced to five hello flood attacks.

We can clearly determine from the above figure that the hello attacks are very well managed by our model with a detection rate equal to 100%.

5.3 Scenario Composed of Several Simultaneous Attacks

To ensure more capacity of our model, we have confronted our network to 20 attacks of 4 types: five sinkhole, five Sybil, five hello flood and five replication attacks.

From Fig. 3, we can clearly determinate the high efficiency of the network when the number of synchronized neighbor exceeds five with a detection rate equal to 100%. If the number of synchronized neighbors is less than 5, we also find good results with an acceptance rate superior or equal to 90% with very light rate of false positive and false negative. Which proves the strongest robustness of our model.

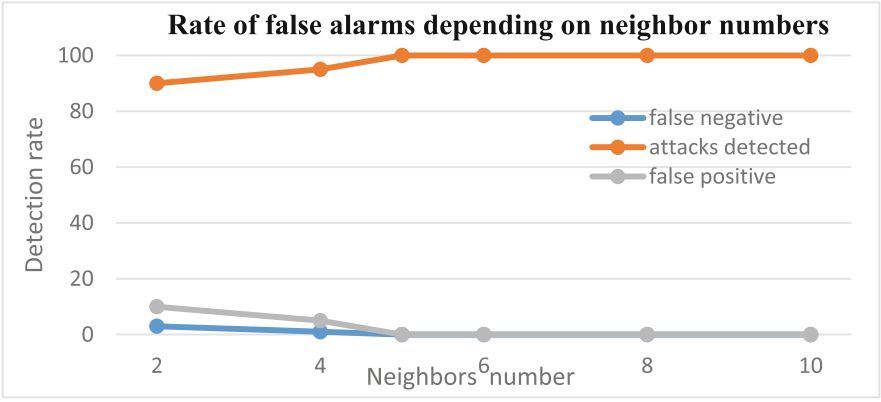


Fig. 3. Our network faced to twenty different attack.

6 Conclusion

In our work, we have created a powerful new model that takes into account all layers of the OSI network model to deal with most attacks that threaten WSNs. In our model, we implemented the TDOA geolocation method to ensure nodes location. Likewise we used the mechanism watchdog to ensure the right transfer of data. Despite the robustness of our model and its effectiveness against existing models, it is an expensive energy consumption model. In future works, we seek to integrate the concept of cloud computation to provide more resources to our model.

References

1. Di Pietro, R., Guarino, S., Verde, N., Domingo-Ferrer, J.: Security in wireless ad-hoc networks: a survey. Comput. Commun. **51**, 1–20 (2014). Elsevier
2. Chen, H., Wu, H., Zhou, X., Gao, C.: Agent-based trust model in wireless sensor networks. In: Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. IEEE (2007)
3. Ozdemir, S.: Functional reputation based reliable data aggregation and transmission for wireless sensor networks. Computer Engineering Department, Elsevier (2008)
4. Jøsang, O., Ismail, R.: The beta reputation system. In: The 15th Bled Electronic Commerce Conference, Bled, Slovenia (2002)
5. Ganeriwal, C., Balzanoand, L.K., Srivastava, M.B.: Reputation-based frame work for high integrity sensor networks. In: Proceedings of the 2nd ACM Workshop on Security of Adhoc and Sensor Networks (2004)
6. Yao, Z., Kimand, D., Doh, Y.: PLUS: parameterized and localized trust management scheme for sensor network security. In: IEEE International Conference on Mobile Adhoc and Sensor Systems (2008)
7. Deng, H., Sun, X., Wang, B., Cao, Y.: Selective forwarding attack detection using water marking WSNs. In: ISECS International Colloquium on Computing, Communication, Control, and Management (2009)
8. Hur, J., Lee, Y.,Yoon, H., Choiand, D., Jin, S.: Trust evaluation model for wireless sensor networks. In: The 7th International Conference on Advanced Communication Technology, pp. 491–496 (2005)
9. Na, Y., Liping, G., Chunxue, W.: A light-weighted data trust model in WSN. Int. J. Grid Distrib. Comput. **7**(2), 33–40 (2014)
10. Kumar, M., Alex, L., et al.: A hybrid trust based secure model for wireless sensor network. Int. J. Emerg. Trends Technol. Comput. Sci. 141–147 (2014)
11. Maddar, H., Kammoun, W., Cheikhrouhou,O., Youssef, H.: Lightweight trust model with high longevity for wireless sensor networks. In: Proceedings of the 2nd International Conference on Information Systems Security and Privacy, pp. 548–553 (2016). ISBN 978-989-758-167-0
12. Stanoev, A., Filiposka, S., et al.: Cooperative method for wireless sensor network localization. Ad Hoc Netw. **40**, 61–72 (2016)
13. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: Energy efficient communication protocol for wireless microsensor networks. In: Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, p. 10 (2000)

Temporal Patterns Visualization for Knowledge Acquisition in Dynamic Decision-Making Environment

Jihed Elouni[1(✉)], Hela Ltifi[2], Mounir Ben Ayed[2], and Mohamed Masmoudi[1]

[1] METS Micro Electro Thermal Systems, National School of Engineers (ENIS), University of Sfax, BP W, 3038 Sfax, Tunisia
auni.jihed@gmail.com, mohamed.masmoudi@enis.rnu.tn

[2] REGIM: REsearch Groups in Intelligent Machines, National School of Engineers (ENIS), University of Sfax, BP W, 3038 Sfax, Tunisia
{hela.ltifi,mounir.benayed}@ieee.org

Abstract. To support decision-making, several visual data mining methods were developed to facilitate visual analysis and extraction of hidden and interesting patterns, particularly from temporal data. Such patterns allow decision-makers to timely and quickly recognize, gain insight and interpret the knowledge in their temporal visual representation. In this paper, a visualization method is proposed to articulate the knowledge produced by these temporal patterns. This method was applied to generate a patterns visualization tool to fight against nosocomial infections in the intensive care units to prove its viability.

Keywords: Decision support systems · Visual data mining · Patterns · Knowledge visualization · Nosocomial infections

1 Introduction

The analysis of the large amount of temporal data currently stored in databases and extracting hidden patterns represent a challenge for dynamic decision-making, which requires effective and user-friendly solutions [5].

Data mining techniques and tools are more and more applied in many domains particularly in complex and dynamic environments for decision-making tasks. This solution is called intelligent Decision Support Systems (iDSS). Generated patterns have different forms according to the applied data mining technique (such as probabilities, association rules, etc.) [7]. To be interpreted and comprehended, these patterns must be represented in a visual form. Visual representation of extracted patterns allows articulating the knowledge behind them. In this context, comes our contribution that addresses the knowledge visualization from automatic temporal extracted patterns.

In this work, we present a method to support such a view of knowledge. It should allow better treat, analyze, extract and interpret knowledge insight in the temporal patterns representation for assisting in decision-making activities.

As an application we have developed a knowledge visualization tool for the fight against Nosocomial Infections (NI) in Intensive Care Unit (ICU) of the Teaching Habib

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_83

Bourguiba Hospital of Sfax, Tunisia. The extracted temporal patterns are a set of NI acquiring probabilities that are daily calculated by the Dynamic Bayesian Networks (DBN) technique.

This article is organized into 4 sections. In Sect. 2, we present our theoretical context. In Sect. 3, our visualization model is proposed. In Sect. 4, we discuss the application of our proposal. Section 5 presents the evaluation of its effectiveness and usability. Our conclusions and further works end this paper.

2 Research Context

In this section, we briefly present the concepts related to our research context, including decision-making based on data mining and visualization as well as the knowledge visualization.

2.1 Visual Intelligent DSS in Dynamic Situations

Decision support systems (DSS) are special class of information system whose role is supporting decision-making activities by generating best possible solutions to the decision problems [7]. However, with the large amount of temporal data and dynamic environment, generated decisions cannot be sufficiently consistent and manage uncertainty in their results. In this context, many research works highlighted the integration of data mining and visualization technologies to overcome such difficulty that faces decision-makers [6].

Integrating data mining technology into the decision support process aims to help in the analysis of large amounts of temporal data. It trends to extract interesting hidden

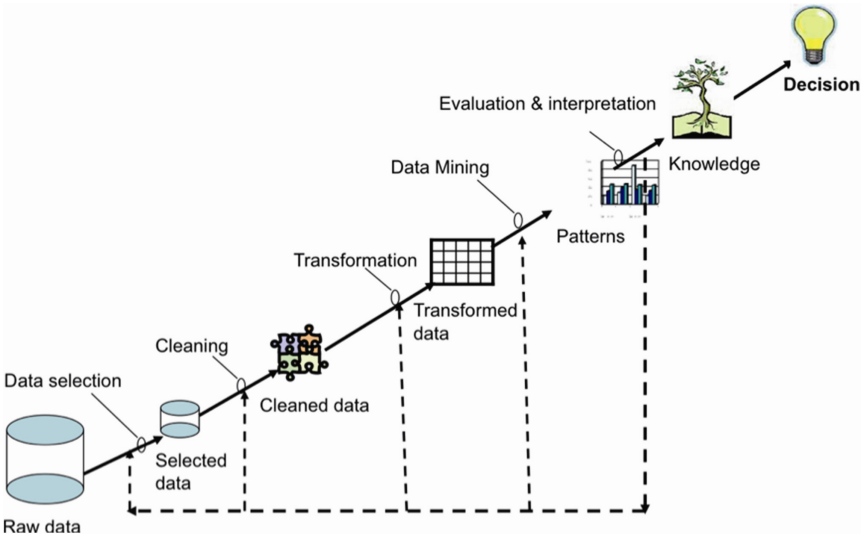


Fig. 1. Data mining process [7]

relationships among variables. Data mining processes data from different perspectives into useful knowledge for decision-making, and becomes an important component in designing **i**ntelligent Decision Support Systems (iDSS). It passes by four phases (cf. Fig. 1) [7]: (1) data preparation consisting of data selection, cleaning and transformation, (2) data mining algorithm application, (3) evaluation and interpretation of the extracted patterns and (4) knowledge integration.

Since the human brain can trait a small amount of information, it can recognize a limited number of different visual objects instantly [5, 11]. Consequently, the conversion of numerical data to images is an important visualization process to be computationally done. For this, information visualization has received increased attentions in developing visualization techniques for exploring databases. The emerging results from information visualization community can be an important contribution to iDSS community if providing visual techniques enabling iDSS to use the available information and to provide more relevant knowledge source for decision-making [6, 11].

The contribution is the visual data mining incorporation for supporting decision-making via means of intelligent and visualization techniques [6, 12]. Such featured viDSS (**v**isual **i**ntelligent DSS) would be adaptable and applicable for solving an important range of problems, and can be used for many fields such as education, medical care, and banking decisions.

As presented above viDSS: (1) generates interesting patterns, and (2) uses visualization techniques to understand data, follow the data mining algorithm application and interpret these patterns. In this context comes our contribution that concerns in particular the patterns visualization for knowledge acquisition. We talk in particular about the knowledge visualization framework.

2.2 The Knowledge Visualization

Applying information visualization techniques to represent extracted hidden models is considered as insufficient for knowledge acquisition by decision-makers [4]. In this context, we are interested in the knowledge visualization framework of Burkhard [1, 2].

Many definitions of knowledge visualization exist in the literature. For example Eppler [3, 4] define it as all (interactive) graphic means that can be used to develop or convey insights, experiences, methods, or skills for knowledge creation. This definition implies that the main purpose of knowledge visualization is to support the process of creating and sharing knowledge with others. Knowledge transfer is then a central process in the knowledge visualization. According to Burkhard [1, 2], for a successful transfer and an effective creation of knowledge through visualizations, four relevant perspectives should be considered (cf. Fig. 1): the function type, the knowledge type, the recipient, and the visualization type. The first three perspectives weigh on the user (i.e. decision-maker) while the fourth perspective focuses on the information visualization techniques.

Knowledge visualization is more complex and difficult to manage that the information visualization. Differences between the two domains are presented by Table 1.

Table 1. Knowledge visualization vs information visualization

	Information visualization	Knowledge visualization
Objectives	New insights is explored by computer applications	Visual representations are used to enhance the transfer and the creation of knowledge
Content	Explicit data like facts and numbers	Knowledge types like experiences, insights, social structures
Perspectives	**Function type**: concerns the aim that should be achieved (i.e. coordination, attention, recall, motivation, elaboration and new insight) **Knowledge type**: concerns the type of knowledge to transfer (i.e. Know-what, know-how, know-why, know-where, know-who) **Recipient type**: concerns the target group (i.e. individual, group, organization and Network)	**Visualization type**: concerns the types of visualization to use and develop (i.e. Sketch, Diagram, Image, Map, Object, Interactive Visualization and Story). *Knowledge visualization perspective = information visualization perspectives + visualization type*
Contribution	Innovation-oriented: create technical methods	Solution-oriented: apply new and traditional visualization problems to solve predominant problems
Phases	1. Information architecture 2. Design 3. Interaction	1. Cognition 2. Perception 3. Communication

Our contribution aims to propose a method that articulates knowledge from data mining extracted patterns. Such method allows assisting decision-makers in acquiring, transferring and integrating knowledge for dynamic decision-making.

3 Proposal

The proposed method aims to better analyze, understand and interpret the temporal data mining patterns in order to efficiently acquire and transfer knowledge between decision-makers. The objective of this method is to design and develop knowledge visualization tools for the purpose of dynamic decision-making. The proposed method is based on the three main phases of knowledge visualization: cognition, perception and communication (cf. Fig. 2).

As presented by Fig. 2, we propose an enrichment of the three steps of the knowledge visualization.

The first phase focuses on the decision-maker **cognition**. It begins when he/she must take real-time decisions and he/she received automatically generated patterns from temporal data (by the data mining algorithm). It consists of sending a query to visually represent these patterns. The developed information visualization technique will be applied to the temporal patterns in order to generate a visual representation associated to a timeline. Several choices are available to the decision maker such as browsing,

managing, analyzing and structuring the relevant models from which associated knowledge will also be visualized interactively. The purpose of these actions is to reduce the cognitive load of the decision makers where interacting with the patterns representation.

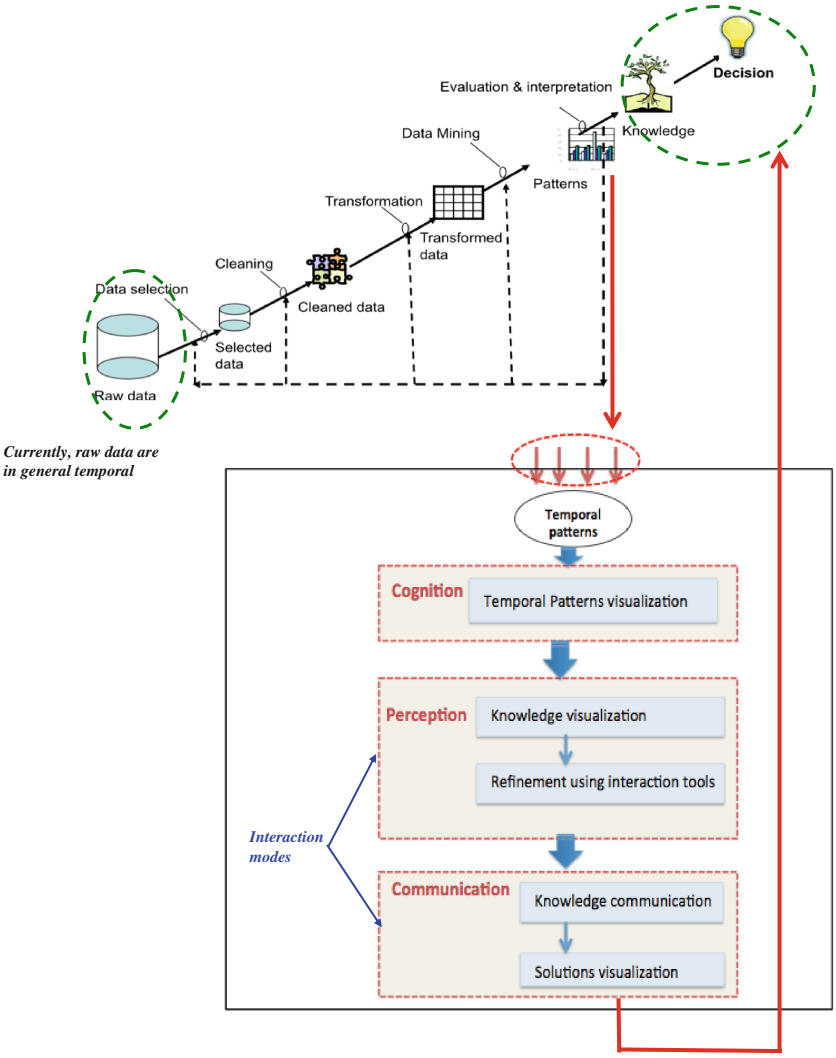


Fig. 2. Temporal patterns visualization method

The second phase concerns the **perception** of these visualized patterns, especially because of its temporal aspect. It consists of an interactive visual analytics that helps to better understand, learn and interpret visual information and associated historical relationships. Our method offers via the refinement step to the decision-maker the ability to apply the modes of interaction directly on the visualization. These modes include

techniques like filtering, zooming, labeling, etc. [10]. Such refinement allows refining the patterns display until finding the trusted and comprehended knowledge to be integrated for decision-making. It is a visual acquisition of knowledge from patterns. We are talking about visual communicating already building knowledge to decision-makers. This is the first step of the **communication** stage. Based on the communicated knowledge, the tool integrates it to provide possible solutions to the initial decision problem. A visual representation of the generated recommendations allows decision-makers to compare and interpret them in order to better chose the best one. Interactions modes are also applied to the solutions recommendations visualization for the important cognitive effect they provide.

Our method was applied to the temporal database of the Intensive Care Unit (ICU) of the Habib Bourguiba Teaching Hospital for knowledge visualization in order to fight against nosocomial infections.

4 Application

The impact of viDSS is in supporting decision making via means of intelligent and visualization techniques. Our contribution is to integrate to the viDSS a knowledge visualization tool. Its application can prove our method viability. In this context, we have applied it to a medical case for the fight against Nosocomial Infections (NI) in the Intensive Care Unit of the Teaching hospital Habib Bourguiba, Sfax, Tunisia. These infections are hospital acquired infections occurred up to 48 h after hospital admission. The patterns to be visualized in this case study are generated by a DBN algorithm that handles temporal patient data.

Physicians are the users of our tool. They should daily obtain knowledge on the patterns predicting the NI appearance. This prediction is based on temporal variables, which are performed medical acts, infectious examination and antibiotics catch. The Developed DBN was presented in [7]. It produces probabilities on the NI occurrence. On this system, we have developed and integrated our knowledge visualization tool. To do so, we have applied our method as follows:

1. *Cognition phase*: the developed tool (visible in Fig. 3) requests the system to provide the NI probabilities at a specific hospitalization day to interactively visualize the patterns. Our objective is to ameliorate the decision-making on the hospitalized patient state by using an appropriate interactive representation to visualize knowledge on the NI occurrence risk.

We choose to work with the bubble chart visualization based on the following Knowledge visualization perspectives:

- Function type: Elaboration of temporal decisions
- Knowledge Type: Know what is the risk of NI occurrence
- Recipient Type: Individual decision-makers (ICU physicians)
- Visualization type: Interactive visualization, which is the interactive bubble chart representation.

The principle of this visualization is to compare and present relationships between labeled circles. It helps to: (1) automatically generate trend lines and interactive bubbles, where the chosen color indicates the risk of NI acquisition (green for a probability <50%, yellow for a probability >50% and red of infected patient state), (2) adjust them by the decision-makers, and (2) consequently update associated data.

2. *Perception phase*: the interactive bubble chart representation is an abstract graphic to interpret and analyze the relationships and trends between all motifs during time. The user can interact with the visualization tool with several ways to generate new insights:

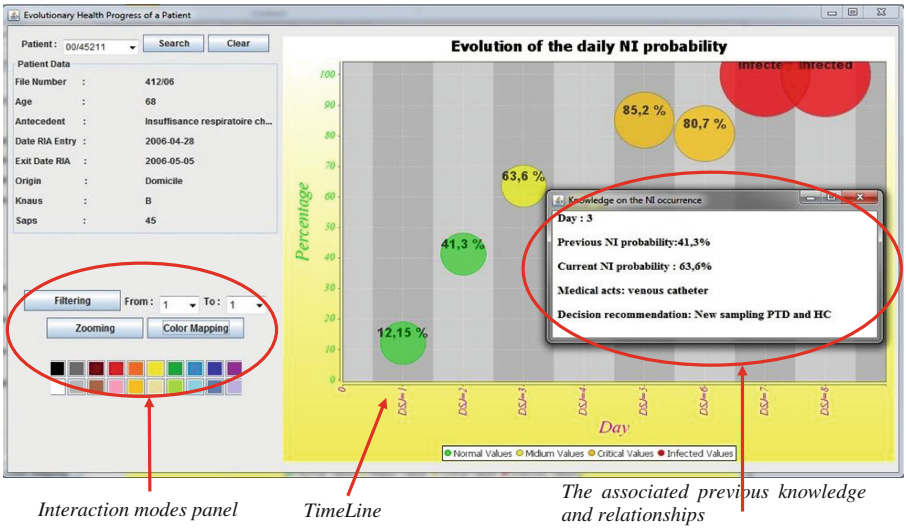


Fig. 3. Developed knowledge visualization tool

(1) To acquire knowledge related to the probabilities and the previous stored knowledge in the knowledge base,
(2) Collect and retrieve the knowledge in relation to the patterns and previous knowledge stored in the knowledge base, and
(3) To visually interpret the differences or similarities between the related measures over time.

3. *Communication phase*: when the user choose the correspondent date in the timeline, the NI occurrence probabilities are visualized as interactive bubbles and the decision-maker can interact with them. It can perform:
   (1) Interactive filtering on the patterns and their relationships in all the displayed area to facilitate research, and improve the visualization.
   (2) Zooming to switch from an overview to take different positions of viewing ad more significant information and/or investigation.
   (3) Color mapping for personalizing the bubbles' colors.

In the following section, we propose to evaluate the developed knowledge visualization tool under the effectiveness and usability dimensions.

5 Effectiveness and Usability Evaluation

5.1 Effectiveness Evaluation

Rogers et al. [9], defines effectiveness by how good a tool is at providing what it claims to do. As our prototype application is intended to improve physicians' understanding of extracted patterns and decision-making. Consequently, the tool application is considered effective when it can visually generate good prediction results. This outcome can be assessed by comparing NI predicted values with actual values before and after integrating the developed tool in the existing DSS for the fight against NI. Participants from the two groups attended the tests individually. The first group is composed of five ICU physicians while the second one includes five Health Informatics (HI) professionals.

The two groups of participants were evaluating these two versions of DSS (iDSS vs viDSS) using a set of trials on a test base containing 58 patients in order to compare the predictive ability of these two versions. According to the results presented in Table 2, the classification rate was correct to 74% ((9 + 34)/58) in the first version of the DSS for the fight against nosocomial infections (based on the classification results of the DBN algorithm), which is interesting. However, after integrating our visual prototype we found an improved classification rate corresponding to 86% ((12 + 38)/58), which were appreciated by the ICU physicians and the Health informatics professionals.

Table 2. The prediction results provided by the DBN

		First version of DSS [7]		Our prototype	
		Actual value (confirmed by experiment)			
		Positive	Negative	Positive	Negative
Predicted value (by the test)	Positive	9	8	12	5
	Negative	7	34	3	38

5.2 Usability Evaluation

To evaluate our knowledge visualization tool, it's necessary to define its ability to satisfy the users' needs. To know the performance of our proposed tool, we choose to apply the evaluation method of [8]. 15 min after a training session on the developed visualization tool, some relevant questions related to the criteria of the tool were proposed to the participants (ICU physicians and HI professionals) of the two groups. The participants attribute to each question a value from 1 to 10. The evaluation results are presented in Table 3.

Table 3. The usability evaluation results

Criteria	Participants scores									
	ICU physicians					HI professionals				
	PH 1	PH 2	PH 3	PH 4	PH 5	HI1	HI2	HI3	HI4	HI5
Emancipation	**7.33**	**7.66**	**7.66**	**7**	**7.16**	**7**	**8**	**7.33**	**7**	**7**
Satisfaction	7	8	8	6	7	7	8	8	6	7
Flexibility	8	8	7	7	7	7	8	7	7	7
Learning	7	7	8	8	7,5	7	8	7	8	7
Information	**7**	**7.16**	**7.33**	**7**	**7**	**7.33**	**7**	**7.33**	**6.83**	**6.66**
Accessibility	7	8	8	8.5	7.5	8	8	7	8	7
Accuracy	7	7	7	6.5	7.5	7	7	8	6.5	7
Search ability	7	6,5	7	6	6	7	6	7	6	6
Technologies	**7.33**	**7**	**7.33**	**7**	**7**	**8**	**7.33**	**7.66**	**7.33**	**7**
Learnability	8	7	7	7	7	8	7	7	7	7
Efficiency	6	7	7	8	6	8	8	8	8	7
Error proneness	8	7	8	6	8	8	7	8	7	7

The usability evaluation of the visualization tool has shown that users from the two groups have, in general, appreciated using our tool (since evaluation values are between 6 and 8.5).

One of key usability metrics is the Time Constraints. Such factor is considered as important for evaluating the interaction tasks of our visual tool. To do so, all users run the tool interaction tasks for three days. Average Time of completion for each task is presented in Table 4.

Table 4. Time constraints of the tool' interaction tasks

	Zooming	Filtering	Labeling	Color mapping
First day	4 trials	4 trials	5 trials	2 trials
Second day	4 trials	4 trials	4 trials	3 trials
Third day	4 trials	3 trials	3 trials	2 trials
Trials amount	**12**	**11**	**12**	**7**
Total time task (in seconds)	18	18	18	18

While the calculated total time, presented in Table 4, was the same for the four interaction tasks, users performed the zooming and labeling tasks 12 times (*faster tasks*: each trial lasted 1.5 s.) vs. 11 times for filtering (*medium task*: each trial lasted 1.7 s. per trial in the two first days and 1.5 s. per trial in the third day) vs. 7 times for color mapping (*slower task*: each trial lasted 3 s. per trial in the two first days and 1.5 s. per trial in the last day). We found that users spent more time with interaction tasks during the two first days. After they learned them, all users have taken the minimum time per trial in the third day.

6 Conclusion

In this paper, we are particularly interested in the decision support systems based on data mining and visualization technologies. The main advantage of such kind system is its contribution to a better understanding of the decision process in complex dynamic environments. However, it does not take into account the knowledge acquisition from automatically generated patterns. In this context, comes our contribution. It consists of proposing a knowledge visualization method taking into account the perspectives of the Burkhard framework.

The introduced method contains three main phases: cognition, perception and communication. These phases are enriched by a set of successive steps to enhance the visualization results and to guide user comprehend patterns and acquire knowledge to better make appropriate and right decisions. Our method was applied for the daily fight against NI in the ICU. To calculate NI occurrence probability we have used the DBN classification results. The extracted patterns are interactively visualized using interactive bubble chart representation to obtain related knowledge. The effectiveness and usability evaluation of our visualization tool gives good results and reflected the feasibility of our proposed model. In future work, we plan to apply the proposed method for visualizing other kinds of patterns (such as association rules) and to design other visualization techniques to represent patterns and build knowledge. We plan also to integrate cognitive aspects in our proposed method.

References

1. Burkhard, R.A.: Learning from architects: the difference between knowledge visualization and information visualization. In: International Conference on Information Visualization (IV 2004). IEEE Computer Society, Los Alamitos, pp. 519–524 (2004)
2. Burkhard, R.A.: Towards a framework and a model for knowledge visualization: synergies between information and knowledge visualization. In: Tergan, S.-O., Keller, T. (eds.) Knowledge and Information Visualization. LNCS, vol. 3426, pp. 238–255. Springer, Heidelberg (2005). doi:10.1007/11510154_13
3. Eppler, M.J., Burkard, R.A.: Knowledge visualization. Towards a new discipline and its fields of application. In: Schwartz, D.G. (ed.) Encyclopedia of Knowledge Management. Idea Group, Hershey (2004)
4. Eppler, M.J., Burkhard, R.A.: Visual representations in knowledge management: framework and cases. J. Knowl. Manag. **4**(11), 112–122 (2007)
5. Fekete, J.-D., Wijk, J.J., Stasko, J.T., North, C.: The value of information visualization. In: Kerren, A., Stasko, J.T., Fekete, J.-D., North, C. (eds.) Information Visualization. LNCS, vol. 4950, pp. 1–18. Springer, Heidelberg (2008). doi:10.1007/978-3-540-70956-5_1
6. Ltifi H., Lepreux S., Ben Ayed M., Alimi, M.A: Survey of information visualization techniques for exploitation in KDD. In: The 7th ACS/IEEE International Conference on Computer Systems and Applications, AICCSA-2009, Morocco, pp. 218–225 (2009)
7. Ltifi, H., Kolski, C., Ben Ayed, M., Alimi, M.A.: Human-centered design approach for developing dynamic decision support system based on knowledge discovery in databases. J. Dec. Syst. **22**(2), 69–96 (2013)

8. Palmius, J.: Criteria for measuring and comparing information systems. In: Proceedings of the 30th Information Systems Research Seminar in Scandinavia IRIS (2007)
9. Rogers, Y., Sharp, H., Preece, J.: Interaction Design: Beyond Human-Computer Interaction, 2nd edn. Wiley, New York (2007)
10. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages. IEEE, Los Alamitos, pp. 336–343 (1996)
11. Mazeika, A., Simoff, S.J., Böhlen, M.H. (eds.): Visual Data Mining, Theory, Techniques and Tools for Visual Analytics, vol. 4404. Springer, Heidelberg (2008)
12. Ya'acob, S., Mohamad Ali, N., Mat Nayan, N.: Understanding big picture and its challenges: experts and decision makers perspectives. In: Zaman, H.B., Robinson, P., Olivier, P., Shih, T.K., Velastin, S. (eds.) IVIC 2013. LNCS, vol. 8237, pp. 311–322. Springer, Heidelberg (2013). doi:10.1007/978-3-319-02958-0_29

Cloudlets Architecture for Wireless Sensor Network

Hela Maddar(✉), Wafa Kammoun, and Habib Youssef

PRINCE Research Laboratory, Sousse University,
Hammam Sousse, Sousse, Tunisia
maddar.hela.si@gmail.com, wafa_kammoun@yahoo.fr,
habib.youssef@fsm.rnu.tn

Abstract. The use of sensor networks is increasing more and more in many areas, such as scientific, medical, domestic, military etc. However, such networks are still suffering from a power autonomy limitation that restricts the network lifetime. Energy is a vital component for WSNs. Namely, energy and material resources remain a major challenge for these types of network. The lack of capacity and the search for solutions to these limitations remain the axis of the most curious research that vitally affects the life of such networks. Some new technologies are experiencing a considerable success these days, which is the cloud technology. The cloud is seen today as a major development in the computer world. The integration of this new technology scare offers a high storage and computing power capabilities. In this work, we seek to create a new architecture for wireless sensor network that can be used for different types of applications which cloud computing and cloudlets for wireless sensor networks to form a scalable network with high durability and high material strength.

Keywords: Cloud computing · Cloudlets · Wireless sensor network

1 Introduction

Nowadays WSNs have growing needs standpoint memory storage capacity and computing. The limited resources of these types of networks require the scientific community to explore new effective solutions. At the onset of cloud technology, many researchers have set themselves on computing power and virtual storage which is unlimited. Having seen the great contributions that these new technologies present, we tried to create a new architecture for wireless sensor networks integrating these new technologies. Our new architecture presents a very remarkable contribution standpoint longevity, scalability and resource management. This new architecture can be used for various types of applications such as medical, environmental, industrial, etc. It represents a very large input on the world of wireless sensor networks.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_84

2 What's a Cloud Computing and Why Do We Use It

Cloud computing is the use of more computing power or storage of remote systems through a network. According to the National Institute definition of Standards and Technology (NIST) [1] "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction". The cloud is an important change in computer systems paradigm. The concept cloud appear in 2010. It allows the use of data in all areas. Connectivity anytime, anywhere and on any device. It allows you to search and analyze information in near real time. It enables immediate flexibility of user requirements. Use on request and accessibilite from any geographical point of the network. This allows more efficient use of resources and optimizes energy consumption.

According to NIST, there are three categories of services that can be offered by the cloud [2]. They are IaaS, PaaS, and SaaS. We describe these services respectively below: IaaS (Infrastructure as a Service): This is the lowest level of service. It consists of virtual machines on which the consumer can install an operating system and applications. This service is similar to classic hosting services data processing centers.

PaaS (Platform as a Service): This type of service is located above the previous. The operating system and infrastructure tools are under the supplier's responsibility. The consumer control application and can add their own tools. The situation is analogical to that of web hosting, or consumer leases the server operation. The difference is that this type of service offer great elasticity and capacity to be adapted to demand. Example of PaaS are Google App Engine, Microsoft windows azure.

SaaS (software as a Service): This type of service is located above the service 'PaaS,. It consists of a set of application accessible on demand without purchased without installation or update. An example is Gmail, social networks like facebook and twitter, etc.

3 What's a Cloudlets and Why Do We Use It

In order to increase the processing power nodes, memory and longevity of the network, it is necessary to use a more powerful device with improved processing speed. This has become a vitality. It is a valid reason for introducing the cloudlets [2]. These are little cloud that represent in our work an intermediate layers located between the cloud and the different cluster head on network. In general, we can define the cloudlets as components with decentralized infrastructure where resources can be drawn from mobile computers proximities. In our case, the interaction of cloudlets is with all cluster head nodes of the network and the cloud that represents our collection station.

To better explain and better show the difference between cloudlets and cloud computing, we conducted a comparative table to show the difference between them (Table 1):

Table 1. Difference between cloud computing and cloudlets

	Cloudlets	Cloud computing
State Management	Soft Self-managed, it is not directed to professional management	Soft/hard Professional administration
Environment	Data Center: data center in a box	Machine room
Property	Decentralized by local	Centralized
Network	LAN latency, bandwidh	Internet latency, banwidh
Sharing	Limited number of users	Large number of users

4 State of the Art

The cloud computing technology was recently considered by researchers in the field of wireless sensor networks because of the large contribution which can present to such field. An example of this work is [3] done by Chui Yaw and all. In this paper the authors propose a model to monitor the environmental conditions of an agricultural site. The cloud is an intermediary between the network and the client to ease data retrieval and to provide sufficient storage space for a number of sensor node that can reach 600,000 nodes. In [4], the authors propose a framework for sensor networks that integrates mobile cloud computing to ensure better durability, safety and rehabilitated. In this model, each cluster is responsible for the collection of data and their shipments to a gateway node that is responsible for sending data to a cloud gateway which is an intermediary with the cloud. The cloud is charged by the encryption and decryption of data transmitted and received by the gateway nodes.

A new paradigm connected to the cloud recently appeared is the cloudlets. A cloudlets can be seen as a small cloud. Internally, a cloudlets brings like a set of multi-core computers together. Usman Shaukat and all in [5] presented a large study on cloud computing and cloudlets by presenting their advantages, their limitations, and the usefulness of their integrations in local wireless networks and how they can be very beneficial. A further very recent architecture integrates cloudets for wireless sensor networks developed by Aymen and all in [6], this architecture is done for health care applications. In this architecture the role of cloudlets is the collections of data patients. The authors in this work have announced their data security fears since the data transmitted to the cloud represent private patient data regarding their health status. We can just emerged from this that the integration of cloudlets in the world of wireless sensor networks has not seen the light until today for an adaptable architecture to different types of application like medical applications, environmental, industrial, etc. For this we have decided to create a new architecture that can be applicable to any types of applications.

In the next section, we present our new architecture that integrates both cloud computing and cloudlets for better resource management and better longevity of network.

5 Our Proposed Model

Our new network architecture integrate cloud computing and cloudlets. This integration solves a vital problem for wireless sensor network standpoint durability and capacity of hardware resources. Our architecture is 4 third based on a hierarchical network.

In the highest level we find the collection station. It represents the heart of our architecture. Our collect station is a cloud computing. It has a global view of the entire system. It is the data collection center.

At this central node are connected cloudlets, they act as a bridge between the collection station that is our cloud and the clusters heads. The role of cloudlets is to collect data received from different cluster heads, aggregate them, then perform their transfer to the base station.

At the third level, we find the clusters heads. These areas managers are generally responsible for the collection of data from different nodes of their areas, then the transfer of data collected to the nearest cloudlets.

At the last level, we find the sensor nodes. These nodes are responsible for observing their surroundings, the collection of observed data and then the transfer of their data report to the next level which is the clusters heads level following on their transfer the process set by the routing protocol used.

Through this new architecture we seek to increase the lifetime of the network, its capacity storage and calculations. And subsequently forming a high-durability and high-capacity network which may include metrics safety without fear strong requested powers. Below the general architecture of our new network (Fig. 1).

Our new architecture is designed for dense networks with a large number of sensor nodes. The cloudlets will be distributed in a symmetrical way on the network coverage area. In our architecture we proposed to implement four cloudlets. This number can varied based on the nature of the application and the number of node deployed.

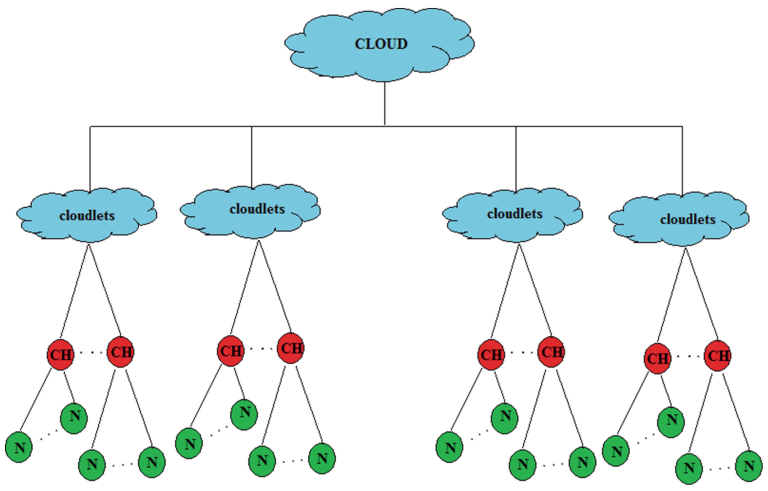


Fig. 1. Our 4-tier architectures proposed.

6 Simulation and Performance Evaluation

In our simulations, we tried to take into account a critical point for WSNs. In a real network, a collection station can be placed in a remote area network. Which can pose a real problem for most networks, this requires a high energy power and a lot of expense to route the data to the final station. In our case, this problem is no longer posed seen the strong power of cloud computing and cloudlets which are loaded for the routing of data to the final point. For this, we tried to fix the base station to different locations to see the bring of our contribution.

6.1 Simulation Environment

To evaluate the functionality of our study, we used the MATLAB simulator. Our experimental model includes 200 nodes randomly distributed sensors on a square area of 100 × 100 m^2. To test the effectiveness of our contribution, we used the LEACH protocol [7] as a routing protocol in our simulations (Table 2).

Table 2. Parameters of the experimental model

Parameter	Value
Number of round (round Simulation)	300
Number of node	200
The location of the first base station SB1	100*100
The location of the second base station SB2	200*200
The initial energy for advanced nodes	0.2j
The initial energy for normal nodes	0.1 j
The location of the first Cloudlet 1	25*25
The location of the second Cloudlet 2	75*25
The location of the third Cloudlet 3	25*75
The location of the fourth Cloudlet 4	75*75
Transmitting and receiving send energy Elec ETX = = ERX	50 nJ

Below the different parameter values taken during our experiment (Fig. 2):

6.2 Topology Based on the New Architecture

The figure above describes the new topology of the LEACH protocol during the simulation with a base station located on 200 × 200 m^2. In this figure the yellow color represents the dead nodes. The red circles represent cloudlets. Black signs represent advanced nodes and green circle represent normal nodes. The blue signs on green or black elements represent the Cluster Head in the area where they were quoting.

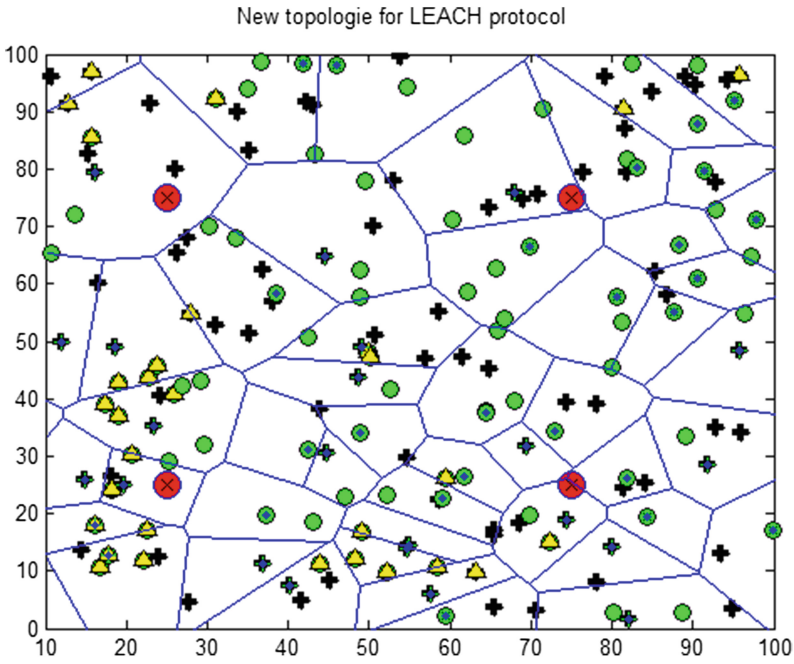


Fig. 2. Cloud LEACH topology.

6.3 Evaluation of Lifetime of Different Protocols

To compare the lifetime of the network between the Cloud LEACH architecture and normal LEACH protocol, we measured the residual energy of the network for each protocol (see Fig. 3).

This figure shows the energy consumption rate by the protocols Cloud LEACH and LEACH compared to the number of round. From this graph we can clearly notice that the cloud LEACH protocol in the different positions of the base station gives better results compared to LEACH protocol. We treated two scenarios, firstly we treat case of a base station located on 100*100 m^2. Taking the example of round number 300, during this round, the remaining energy for Cloud LEACH protocol is equal to 2 J and for LEACH it is just equal to 0,4 J. Tacking now the second scenario, if the base station is placed on 200 * 200 m^2, the remaining energy is nearly equal to 2,5 J for cloud LEACH. For LEACH there is no amount of energy remaining. We can clearly conclude that the integration of Cloud and cloudlets on wireless sensor network is become a necessity view the excellent life prolongation that they provide.

6.4 Number of Alive Nodes for Each Protocol Over Time

Using this curve, we seek to determine the number of alive nodes for Cloud LEACH and LEACH protocol in different rounds (Fig. 4).

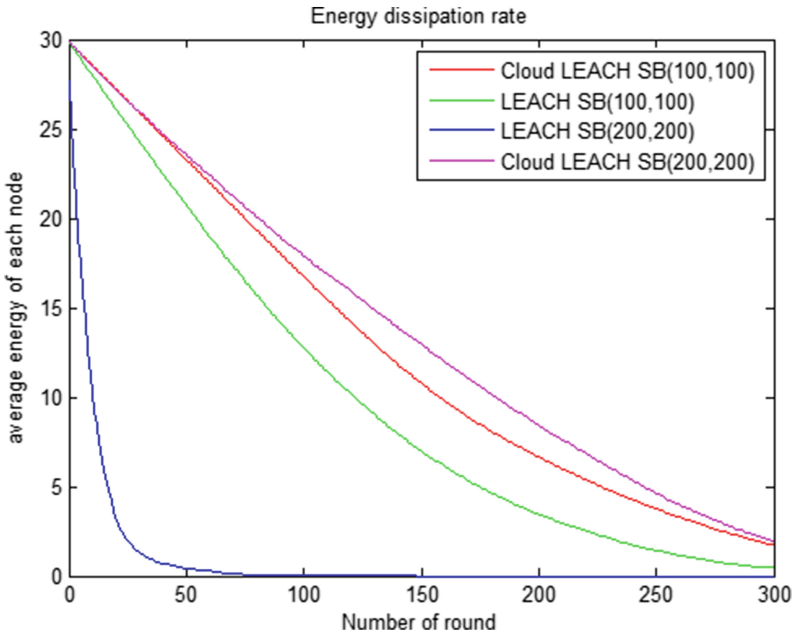


Fig. 3. Energy dissipation rate of LEACH and Cloud LEACH.

From the above curve, we can clearly see the big difference between the numbers of alive nodes during different iterations. Cloud LEACH protocol always gives us the best result. For more detail we treat at first place LEACH and cloud LEACH protocol for a location of the base station equal to 100 × 100 m^2, For a round number equal to 300, LEACH has an equivalent number of live nodes equal to 22 nodes, for the same round Cloud LEACH has a number of live nodes equal to 57. So we got a difference of 35 alive node. Secondly, we compare LEACH and cloud LEACH protocol for a location of the base station equivalent to 200 × 200 m^2. We note that from the round number 72 we have a number of alive node near to 0 for LEACH. But during this tour, Cloud LEACH has a great number of live nodes equal to 185 node.

6.5 Number of Dead Nodes for Each Protocol Over Time

The figure below shows a comparison of the numbers of dead nodes by each protocol during different rounds (Fig. 5).

This figure confirms our previous description. The most efficient protocol is the Cloud LEACH for both scenarios studied.

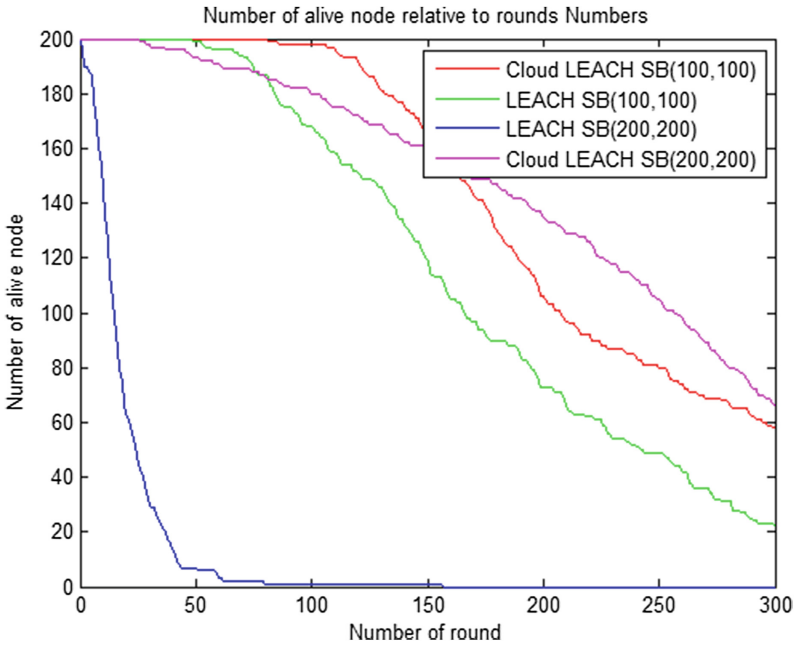


Fig. 4. Number of live node compared with the number of turns.

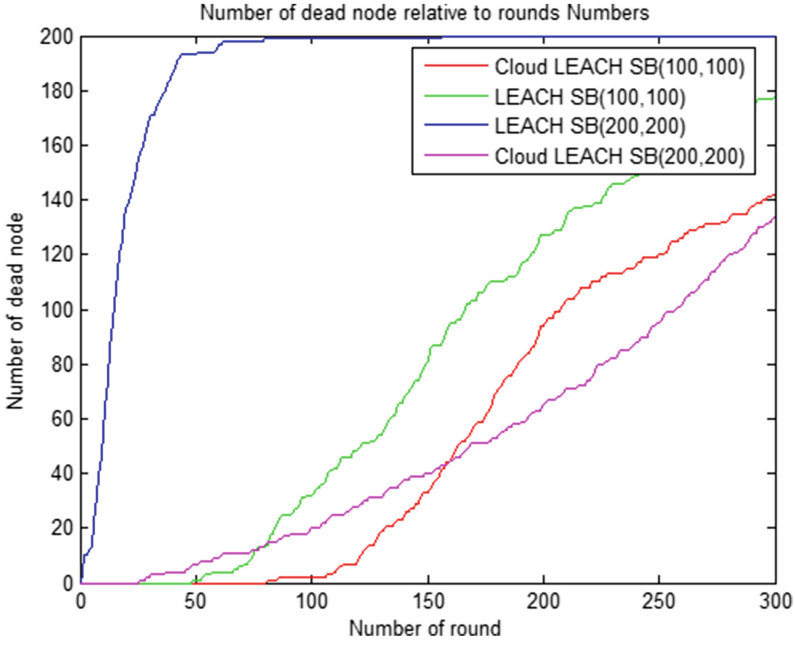


Fig. 5. Number of dead nodes.

6.6 Number of Packets Transmitted to the Base Station

Furthermore the lifetime of the network, another measure to assess the efficacy of a routing protocol is the number of packets sent to the base station during the different rounds (Fig. 6).

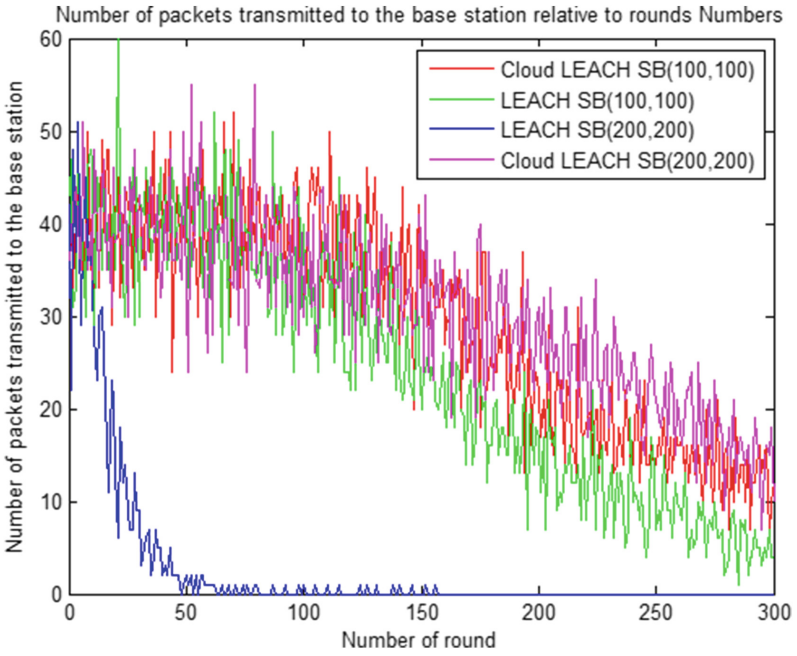


Fig. 6. Number of data packets transmitted to the base station relative to numbers of turns.

The figure above shows the number of packets sent to the SB during the different rounds. Given the number of dead nodes during various iterations over the network, a decline in the number of packets sent is very logical. The protocol that guarantees the minimum regression of packets sent is the most efficient. If we consider the scenario of LEACH and Cloud LEACH for a base station located on 100*100 m^2, we note that the red curve that represents Cloud LEACH excels the green curve that represents LEACH protocol. Similarly, if we take the second scenario where the location of the base station is equivalent to 200×200 m^2, we note that the pink curve that represents Cloud LEACH excels the blue curve which represents LEACH.

6.7 Number of Packets Transmitted by the Nodes to Clusters Heads

The figure below shows a comparison of the numbers of packets transmitted by different network nodes to the cluster nodes Head during the different rounds (Fig. 7).

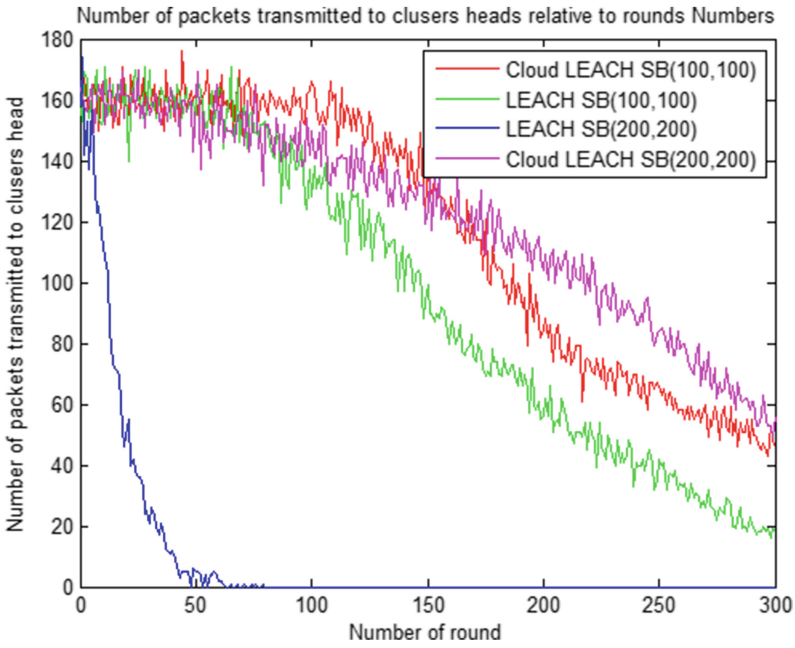


Fig. 7. Number of data packets transmitted to the clusters Heads compared to the number of turns.

From the figure above, we deduce clearly the effectiveness of cloud Leach protocol compared to LEACH. If we consider the scenario of LEACH and Cloud LEACH for a base station located on 100*100 m^2. We note that the red curve which represents Cloud LEACH excels the green curve which represents LEACH. Similarly, if we take the second scenario, where the location of the base station is equivalent to 200×200 m^2, we can clearly note that the pink curve which represents Cloud LEACH remarkably excels the blue curve which represents LEACH. From this, we can clearly deduce the best networking offered by Cloud LEACH.

7 Conclusion

In this work, we have proposed a new architecture for wireless sensor network for hierarchical topology which includes cloud and cloudlets technology, then we have applied our new architecture on LEACH protocol. To test our new contributions, we conducted a comparative study between the new Cloud LEACH and LEACH protocols for two different locations of the base station. We demonstrated the enormous contribution of the integration of Cloud Computing and cloudlets for sensor networks. This new architecture increases the lifetime of the network and subsequently ensures a very good durability of the network. Also, it allows us to be able to minimize the packet loss rate and improve the quantity of packet sent over the network.

Despite these remarkable benefits, ensuring the security of succulent data over the network remains a very important point to be considered. Also, ensuring the vehicle safety of the transmitted data is mandatory. This sensitive point will be considered in our next researches using trust and security mechanism.

References

1. Bohn, R.B., Messina, J., Liu, F., Tong, J., Mao, J.: NIST cloud computing reference architecture. In: Proceeding IEEE World Congress Services, pp. 594–596 (2011)
2. Bahtovski, A., Gusev, M.: Cloudlets challenges. In: 24th DAAAM International Symposium on Intelligent Manufacturing and Automation (2013)
3. Chung, W.Y., Yu, P.S., Huang, C.J.: Cloud computing system based on wireless sensor network. In: Proceeding of Federated Conference Computer Science and Information Systems (FedCSIS), pp. 877–880 (2013)
4. Zhu, C., Leung, V.C.M., Yang, L.T., Hu, X., Shu, L.: Collaborative location-based sleep scheduling to integrate wireless sensor networks with mobile cloud computing. In: Proceeding of IEEE Globecom Workshops (GC Wkshps), pp. 452–457 (2013)
5. Shaukat, U., Ahmed, E., Anwar, Z., Xia, F.: Cloudlet deployment in local wireless networks: Motivation, architectures, applications, and open challenges. J. Network Comput. Appl., 18–40 (2016)
6. Amraoui, A.E., Sethom, K.: Cloudlet softwarization for pervasive healthcare. In: Proceeding 30th International Conference Advanced Information Networking and Applications Workshops (WAINA), pp. 628–632 (2016)
7. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: Proceedings of the 33rd Annual Hawaii International Conference. IEEE (2000)

Data Fusion Classification Method Based on Multi Agents System

Elhoucine Ben Boussada[1(✉)], Mounir Ben Ayed[2], and Adel M. Alimi[1]

[1] Computer Engineering and Applied Mathematics Department, National Engineering School Sfax, Sfax, Tunisia
houcine.boussada@gmail.com, adel.alimi@ieee.org
[2] Computer Science and Communication Department, Faculty of Sciences Sfax, Sfax, Tunisia
mounirbenayed@ieee.org

Abstract. Computer analysis of electrocardiogram (ECG) data has proven to be an important method to detect cardiac arrhythmias, so that can be of great assistance to the experts in detecting cardiac abnormalities. In this study, we propose to develop a system to aid in the diagnosis of anomalies cardiac signals. This system is based on data fusion and architected by using the multi-agents system. Therefore, the proposed system helps doctors to quickly and precisely diagnose a heart disease by examining only the class of the ECG beats. In order to achieve the goal of real-time classification, the data used are divided into two datasets: the training set for the unsupervised learning of the classifier and the testing set for the real-time classification. This system is tested on a MIT-BIH arrhythmia database.

Keywords: Data fusion · ECG classification · Artificial intelligence · Multi-agent system · Neural network · MIT-BIH

1 Introduction

The analysis of electrocardiogram (ECG) has an important role in diagnosing the heart diseases. Subsequently the automatic detection of cardiac arrhythmias beats in an ECG provides clinicians information about the patient health. In this context, several works have been developed in order to analysis and classify the ECG signals. An overview of the ECG classification in the literature reveals the existence of many interesting researches based on different methods. The bibliographic study of some recent researches shows that we can distinguish two techniques for the ECG classification. The first type is based on one method of classification and the second type is based on the combination of two or more methods of classification.

For the first type, there are three well known classifier algorithms: support vector machine (SVM) [2], Bayesian expert system [1] and multilayer Perceptrons neural network (MLP) [2, 3]. The comparison of these different studies shows that the MLP with back propagation gives the most important recognition rate. But the studies of Edurado José and all [3] based on Optimum-path forest (OPF) shown that the MLP

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_85

classifier is the fastest one for the testing task, while the learning task is best performed by the SVM classifiers. And prove that the OPF classifier combined with the Bayesian classifier shown the best balance for classifying the arrhythmic classes.

For the second type, R. Joy Martis [4], prove that the combination of Independent Component Analysis (ICA) and Probabilistic Neural Network (PNN) classifier, with spread value of 0.03, performed better than the Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

Generally a complete system for ECG classification can be divided in to four subsequent categories;

- Preprocessing: consist in detecting and attenuating frequencies of the ECG signal.
- Segmentation: perform a signal normalization and QRS enhancement by delimiting the QRS complex.
- Feature extraction: this step is the key point for the final classification performance, so that many computational methods has used in this step which the variation of wavelet transform the most popular.
- Classification: the final and important step, large number of approaches used for a better accuracy, sensitivity and specificity in order to prove the best classifier.

To improve the classification results many author use the combining methodology. Combining is an approach to improve the performance in classification particularly for difficult problem and those that are high precision. We start with this idea and we said why do not make a multi-agents system with the many combining methodology. Indeed we proposed in the first part of this paper an evolutionary agent-based system for ECG classification. Each agent is composed of the best combination classifier. The second part proposes and discusses Neural Network for ECG Arrhythmia Classification.

2 Cooperative Intelligent Agents System for ECG Arrhythmia Classification

2.1 Combining Methodology

The main purpose of this work is to classify five types of beat classes of arrhythmia which are: Normal beat (N), Supra Ventricular Ectopic Beat (SVEB), Ventricular Ectopic Beat (VEB), Fusion beat (F), and Unknown beat (Q). In this study we follow different recommendations gives by the Association for Advancement of Medical Instrumentation (AAMI) [1], which has developed a standard (ANSI/AAMI/ISO EC57, 1998-R2008) for testing and reporting performance results of computational techniques aiming at arrhythmia classification. The AAMI also recommends the use of the MIT-BIH arrhythmia data base for performance evaluation of arrhythmia systems. The MIT-BIH arrhythmia database contains annotations for both timing information and beat class information manually reviewed by independent experts. So that in our experiment we use this data base (MIT-BI).

As seen in the introduction a complete system for ECG classification can be divided in to four subsequent categories here in this paper we are interested for the forth step (i.e., the classification) but the first three steps have significant importance.

2.2 ECG Pre-processing

An ECG signal consists of three basic waves: P, QRS and T. These waves correspond to a trial depolarization (P wave), ventricular depolarization (QRS waves), and ventricular repolarization (T wave). In the most of literature who treat the problem of ECG classification, authors has based on QRS complex like [4–7], or on RR interval like [8]. Figure 1 shows an example of normal ECG signal.

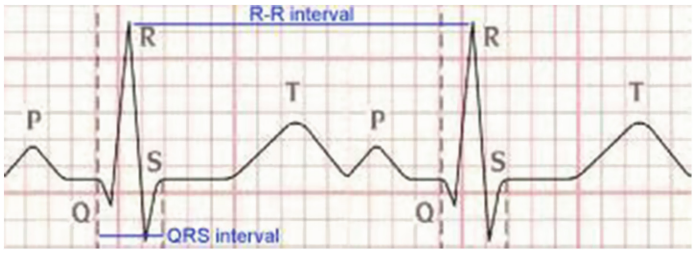


Fig. 1. A normal ECG signal beat

The ECG records from MIT-BIH have generally noise; they present a base line wander and high frequency noise. The main objectives of preprocessing stage are the filtering of high frequency noise and the enhancement of signal quality to obtain appropriate features. The most popular method used in this way is Discrete Wavelet Transform (DWT), it's a linear operation that decomposes a signal in to components that appear at different scales. In this work we used DWT to treat the step of ECG Pre-processing.

2.3 Feature Extraction

In this section there are many approaches widely used in the literature. In [9] Chazal et al. (2004), is based on features from RR interval, QRS complex and the T-wave length. The author also used the information about the presence or absence of P-wave. The algorithm used for fiducially point extraction in Chazal et al. is the same proposed in Laguna et al. however Yu and Chen (2007), used statistical techniques approach directly on heartbeat samples, and extract 10 statistical features, the authors also used the RR interval, and based in this step on wavelet decomposition. In our work we used features from both the RR interval and QRS complex because they have great discriminatory capabilities for heart beat classes.

The ECG signals for training and testing are taken from MIT-BIH arrhythmia database which contained 48 two lead of 30 min recording. 100 beats were chosen for each arrhythmia (SVEB, VEB, F, and Q) and normal (N), ECG divided in to two groups of: training data (50 beats), and testing data (50 beats) see Table 1. Every ECG signal has QRS interval RR interval, in this study we consider RR interval as a beat.

Table 1. Number of testing and training data

	N	SVEB	VEB	F	Q	Total
Training data	50	50	50	50	50	200
Testing data	50	50	50	50	50	200
MIT-BIH data file	101-105-107-108-109-111-118-119-200 201-202-203-207-214-212-223-233-234					

2.4 ECG System Classifier

This section describes the general idea of the proposed system. In addition the result of classification step is the most important output that measures the accuracy, the sensitivity and specificity of the system. Figure 2 shows the proposed system for ECG classification. The detail of Multi-Agents system is explained in the following section.

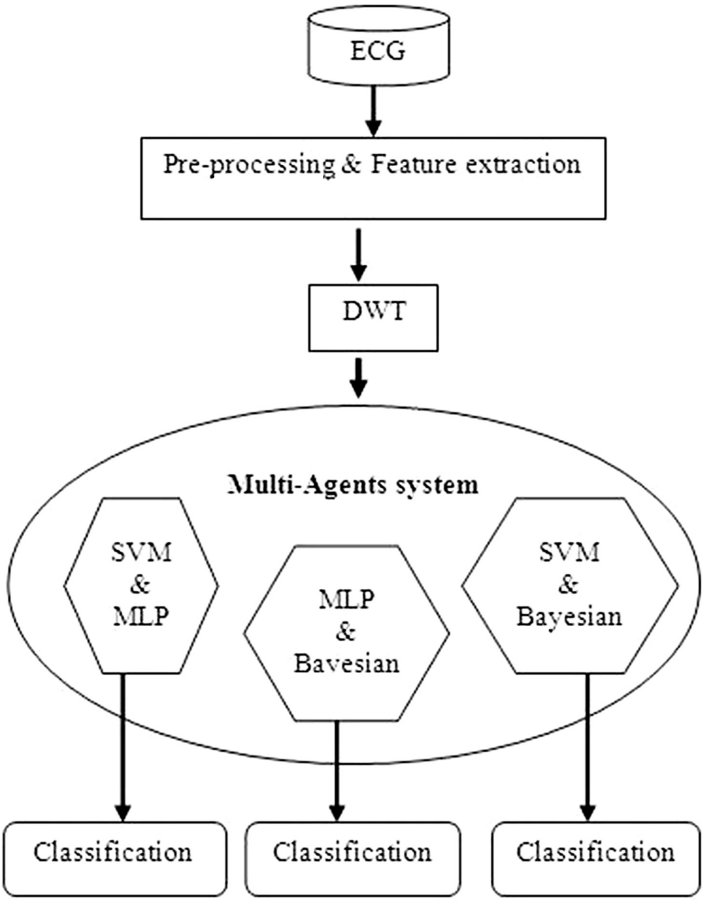


Fig. 2. Multi-agents system for ECG classification

2.5 The System Architecture

The combination of different classifiers explores the idea that different classifiers may offer complementary information about an ECG signal to be classified. To explore this idea we propose, multi-agents system architecture, composed with intelligent agents that cooperate for best results. The system size used in the experiment have a coordinator agent (CAg), and three agents, each agent has two classifiers in the classification module. Three types of classifiers will be used, which are, MLP, SVM and Bayesian classifier. These classifiers are used in different combination, so that we have SM-Ag, MB-Ag and SB-Ag, who use respectively the combination of SVM and MLP classifier, MLP and Bayesian classifier and finally SVM and Bayesian classifier. Also we have a Supervisor Agent (Sag), See Fig. 3.

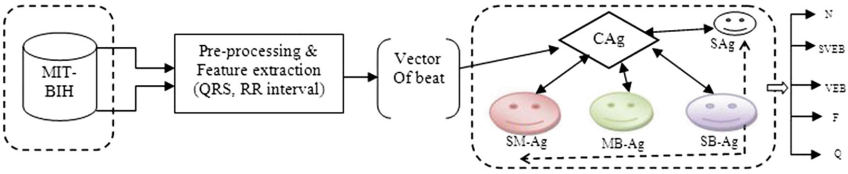


Fig. 3. The classification stages

Our proposed classification system architecture is based on Aalaadin model, which core concepts are Agent, Group and Role. An Agent is defined as an active communicating entity with role. A Group is defined as a set of agent, and finally a role is defined as an abstract representation of an agent service within a group. The role encapsulates the way an agent should act within a group.

The multi-agent system contains three types of agents:

- Coordinate Agent: one agent (CAg), the main function of CAg is to manage the input data set (vector of beat) by affecting a key which matches Sag and the classifier agents.
- Classifier Agents: this type contains three agents, SM-Ag, MB-Ag, and SB-Ag. The main function of classifier agent is to classify an ECG signal and communicate the result to Sag. This result was provided with three coefficients affected at the agent at the test phase.
- Supervisor Agent: this type contain one agent (Sag), which receive the key of any vector and starts counting the time of the execution by the three classifier agents. This time is stopping when it's receiving the results of the classifiers. And according to a rule base it will decide the final class of the vector.

In order to evaluate the different combining classifier, we present the three measures used: Accuracy, Sensitivity and Specificity. These three measures have calculated during the test step using testing data.

Accuracy is defined as the ratio of the Number of ECG Beats Correctly Classified (NB2C), and the Number of Total ECG beats (TB).

$$Accuracy = \frac{TB\,2C}{TB} \tag{1}$$

Sensitivity measured the performance of classification for each class, so it's defined as the ratio of Number of Correctly Classified ECG beats of one Class (NB3C), and the Total ECG Beats classified as that Class (TBC), including the Number of ECB Missed Classification ECG beats (NBMC).

$$Sensitivity = \frac{NB\,3C}{(TBC + NBMC)} \tag{2}$$

Specificity is defined as the ration the Number of ECG Beats Correctly Classified (NB2C) and all ECG beats of a specific Class i.e. TBC and NBMC.

$$Specificity = \frac{NB\,2C}{(TBC + NBMC)} \tag{3}$$

The scheduling process consists of five steps;

1. CAg receives a vector of ECG beat and affects a key for each vector.
2. The CAg request the vector of ECG beat with his key to the three agent's classifier; also inform (by message) the SAg that the key number of ECG beat vector has transmitted to the agent's classifier.
3. Each agent's classifier begins the classification process and requests his result, to SAg, when finished.
4. When SAg receive a request from CAg, it records the key and the time, which stopped when it receive the result from each agents classifier.
5. Finally the decision is gives by the SAg according to a rule base.

The rule base consists as a comparison of three parameters. Which are: Accuracy (1), Sensitivity (2), and Specificity (3). Also SAg provides a forth parameter witch matching the scheduling time of the classifier.

We note here that our final system decision was done on the basis of a rule base. In a further work we can apply in this step an evolutionary optimization method like Particle Swarm Optimization (PSO) which used in several work and shows its importance for the divergence in the most reliable decision as example gives by [7, 10].

2.6 The Experiments Results

The proposed system consist of two main stages: the first, the pre-processing and feature extraction applied DWT to get QRS and RR interval. Then these two kinds of features form an ensemble feature vectors. The second main stage is the scheduling process already described in the previous paragraph (Table 2).

Table 2. Classification results

	SM-Ag	MB-Ag	SB-Ag
Accuracy (%)	98.17	96.06	96.21
Specificity (%)	99.07	94.15	92.45
Sensitivity (%)	98.05	95.24	96.19

In this first part we have proposed a new method for the design of classifier system to automatic ECG analyzing. The main idea of the proposed system is combining different classifier in multi-agents system architecture. However this approach offers complementary information about the ECG state. DWT was used, at the pre-processing and feature extraction step, to extract QRS and RR interval. This feature vectors serves as an input for the system. For the simulation we used the ECG records from MIT-BIH data base. In this study five classes were used. The result gives high accuracy, specificity and sensitivity of SM-ag.

3 Conclusion

In this paper we proposed a multi-data source fusion agent based method for ECG classification. The benchmark MIT-BIH arrhythmia database was used for evaluating the classifier. It is found that this system have an acceptable performance with can be increased by the application of the PSO algorithm.

Although the proposed system can be more developed and compared with others classifier for the same data, all of which are the subject to our future work.

References

1. Luz, E.J.S., Nunes, T.M., de Albuquerque, V.H., Papa, J.P., Menotti, D.: ECG arrhythmia classification based on optimum-path forest. Expert Syst. Appl. **40**, 3561–3573 (2013)
2. Moavenian, M., Khorrami, H.: A qualitative comparison of artificial neural networks and support vector machines in ECG arrhythmias classification. Expert Syst. Appl. **37**, 3088–3093 (2010)
3. Javadi, M., Ebrahimpour, R., Sajedin, A., Faridi, S., Zakernejad, S.: Improving ECG classification accuracy using an ensemble of neural network modules. PLoS ONE **6**(10), e24386 (2011)
4. Martis, R.J., et al.: ECG beat classification using PCA, LDA, ICA and discrete wavelet transform. Biomed. Sig. Process. Control (2013). http://dx.doi.org/10.1016/j.bspc.2013.01.005
5. Javadi, M., et al.: Classification of ECG arrhythmia by a modular neural network based on mixture of experts and negatively correlated learning. Biomed. Sig. Process. Control (2012). http://dx.doi.org/10.1016/j.bspc.2012.10.005
6. Ben Boussada, E., Ben Ayed, M., Alimi, A.: An intelligent data fusion agent-based model for ECG classification. In: International Conference on Computer Related Knowledge, ICCRK, Sousse, Tunisia. IEEE (2012)

7. Ben Boussada, E., Ben Ayed, M., Alimi, A.: Multi-data source fusion agent based method for ECG classification. In: International Conference on Computer Applications Technology, ICCAT, Sousse Tunisia. IEEE (2013)
8. Ka, A.K.: ECG beats classification using waveform similarity and RR interval (2011). http://arxiv.org/abs/1101.1836v1
9. Chazal, P., O'Dwyer, M., Reilly, R.B.: Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans. Biomed. Eng. **51**(7), 1196–1206 (2004)
10. Ben Boussada, E., Ben Ayed, M., Alimi, A.: ECG arrhythmias classification using data fusion and particle swarm optimization. J. Intell. Comput. (2013)

Community Detection in Bipartite Networks Using a Noisy Extremal Optimization Algorithm

Noémi Gaskó(✉), Rodica Ioana Lung, and Mihai Alexandru Suciu

Babeş-Bolyai University, Cluj, Romania
gaskonomi@cs.ubbcluj.ro

Abstract. The network community structure detection problem consists in finding groups of nodes that are more connected to each other than to the rest of the network. While many methods have been designed to deal with this problem for general networks, there are few methods that deal with bipartite ones. In this paper we explore the behavior of an optimization method designed for identifying the community structure of unweighted networks when dealing with bipartite networks. We find that by using the specific Barber modularity, results are comparable with those reported in literature.

Keywords: Community detection · Bipartite networks · Extremal optimization

1 Introduction

Network community structure detection has emerged in the last years as a new research area due to the important insights it can offer into the properties of a network [8]. The community structure detection problem consists in finding groups of nodes within a network that are more linked to each other than to other nodes. If all nodes in a network belong to such a group, we can say that the network presents a community structure. The concept of community, as expressed here, is intuitively derived from the concept of community related to social networks. While easy to grasp, it is difficult to formally define it without limiting its applicability; currently there does not exist a consensus regarding the formal definition of a community.

Regarding bipartite networks, there is a general assumption that using methods designed for general ones will not be efficient due to the particular bipartite structure [2,3,13]. However, there are many real-world applications that require the analysis of a bipartite network: in IP networks [9], bank-firm credit networks [15], software package refactoring [18], protein interaction networks [19], acupoint compatibility [20]. etc. Similar to the case of general networks, a substantial amount of work has been done regarding the use of optimization methods for uncovering the community structure in bipartite networks, either directly

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_86

optimizing a modularity function [3] or indirectly using an optimization method combined with statistical approaches [10].

In this paper we explore the direct use of an optimization method called Noisy Extremal Optimization (NoisyEO), designed for community structure detection in unweighted, undirected networks, to identify communities in bipartite networks by maximizing the modularity [16]. Previous experience has shown that for these networks, the performance of NoisyEO does not depend on the fitness function used [21], so we adapt NoisyEO to optimize the Barber modularity [3] and compare results. We show that NoisyEO is a robust method also in the case of bipartite networks, with results comparable with those provided by other existing methods.

2 Community Structure Detection - Method

Due to the lack of a formal definition for the community structure, several alternatives have emerged, among which some of the most popular are those that convert the problem into an optimization one, in which the maximum of a fitness function is supposed to correspond to the community structure. The popularity of this approach comes from the fact that optimization of a fitness function can be performed by using search heuristics, particularly nature inspired heuristics, that are supposed to be highly adaptable and scalable.

For a general graph one of the most popular fitness function is the *modularity* introduced by Newman and Girvan in [16,17]:

$$Q_N = \sum_{C \in \mathcal{C}} \left(\frac{m_C}{m} - \left(\frac{D_C}{2m} \right)^2 \right), \tag{1}$$

where m is the number of edges, m_C is the number of edges connecting vertices in community C, and D_C is the sum of the degrees of all vertices in community C. A higher modularity value indicates a better solution. The modularity Q_N basically computes the difference between the community structure and the corresponding sets of nodes in a random network that does not present a community structure.

A bipartite network is a graph $G = (R, B, E)$, where R and B are two disjoint types of nodes, $R \cap B = \emptyset$, and there is no edge between any node from R to a node from B. Because the modularity Q_N is computed with respect to a random network, for bipartite networks Barber modified the formula [3] to take into account a *random bipartite* network:

$$Q_B = \sum_{C \in \mathcal{C}} \left(\frac{m_C}{m} - \frac{R_C \times B_C}{m^2} \right), \tag{2}$$

where m is the number of edges, m_C is the number of edges connecting vertices in community C, R_C is the sum of the degrees of vertices from R in community C, B_C is the sum of the degrees of vertices from B in community C.

In what follows we are exploring the use of both modularity measures within an extremal optimization algorithm designed for solving the community structure detection problem in unweighted and undirected networks, called Noisy

Extremal Optimization (NoisyEO [14]). NoisyEO has proven to be a robust optimization technique, mainly because of the extremal optimization component [21] which maximizes the fitness function of each node computed as the contribution of the node to the fitness of its community.

The fitness $f(C)$ of community C, defined in [11], is:

$$f(C) = \frac{m_C}{(D_C)^\alpha}, \tag{3}$$

where α is a parameter controlling the community size. The node's contribution to a community is then:

$$u_i(C_1, ..., C_n) = f(C_i) - f(C_i \backslash \{i\}),$$

where C_i represents the community of player i, and $C_i \backslash \{i\}$ is the same community without node i.

The extremal optimization algorithm evolves pairs of individuals searching the space independently. An individual is encoded as a vector with n components, each component representing a network node that is assigned a community number. Each iteration the current individual is evaluated and the nodes having the worst payoffs are randomly reinitialized. The second individual preserves the best solution found by the current individual. The full description of the method can be found in [14].

In order to decide whether the current individual is better than the best solution found so far any fitness function can be used. In this paper, we will test both modularity measures (Newman and Barber) within NoisyEO in order to explore its behavior when dealing with bipartite networks.

3 Numerical Experiments

In the absence of some clearly defined benchmarks for testing algorithms for bipartite networks, comparable with those for general networks [12], we have considered the following four real world benchmarks datasets:

- Southern woman dataset [5] - a bipartite network with 32 nodes and 89 edges;
- Club membership network [7] - 40 nodes and 95 edges;
- Corporate leadership [4] - 44 nodes and 99 edges;
- American revolution [1] - 141 nodes and 160 edges

Because for these networks the real community structure is not known, we report the values of both modularity measures. For each network we perform two sets of experiments: in one the Newman modularity is maximized (corresponding values for the Barber modularity are also presented), and in the other one the Barber modularity is maximized (and corresponding values for the Newman modularity are presented). Results are further compared with those reported by other methods in the literature, where available.

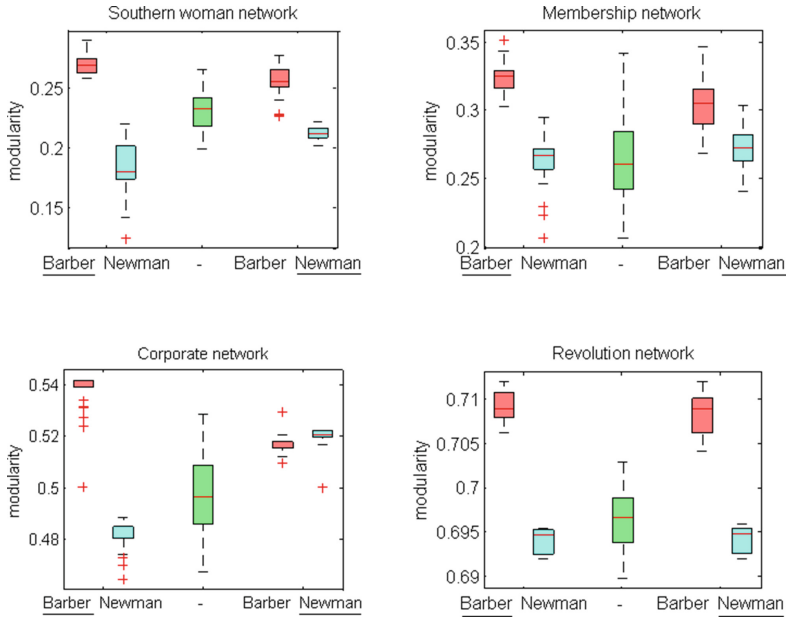


Fig. 1. Results obtained for each network over 30 independent runs. The modularity that is maximized is underlined; for each run both modularities are represented. Thus, for example, in the top-left, the first two boxplots represent the results reported by NoisyEO when maximizing the Barber modularity and the values of the Newman modularity for the reported solutions. The green boxplot in the middle represents the Barber modularity values for the results reported by NoisyEO when no modularity function is used.

Parameter settings. To perform numerical experiments, the following parameters were used (similar to those in [14]):

- population size: 30;
- number of generations: 10000;
- expected minimum and maximum number of communities: 2 and 8;
- number of iterations between shifts: 45;

Results. The results reported by NoisyEO are presented in Fig. 1 as boxplots of modularity values obtained on 30 independent runs for each variant. For each pair of boxplots the modularity that is maximized is underlined. The green boxplot in the middle represents the values of the Barber modularity obtained when no actual modularity function was used (each iteration the current individual replaces the best individual regardless of its modularity value).

While values represented in boxplots in Fig. 1 are very close, a Wilcoxon sum rank test comparing Barber modularities obtained for each network shows that the differences are significant for the southern women, membership and corporate

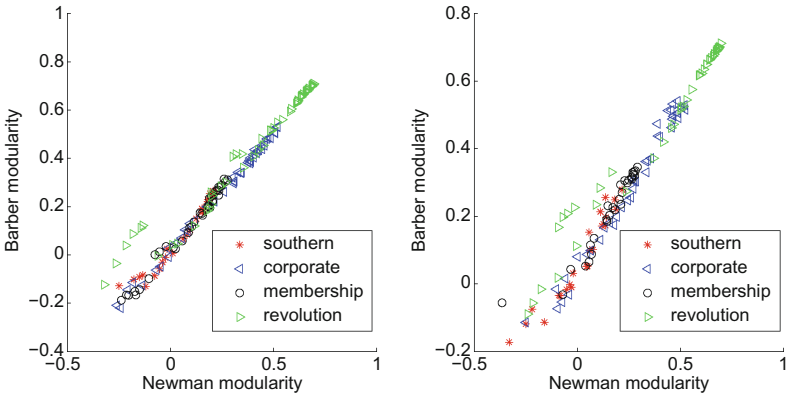
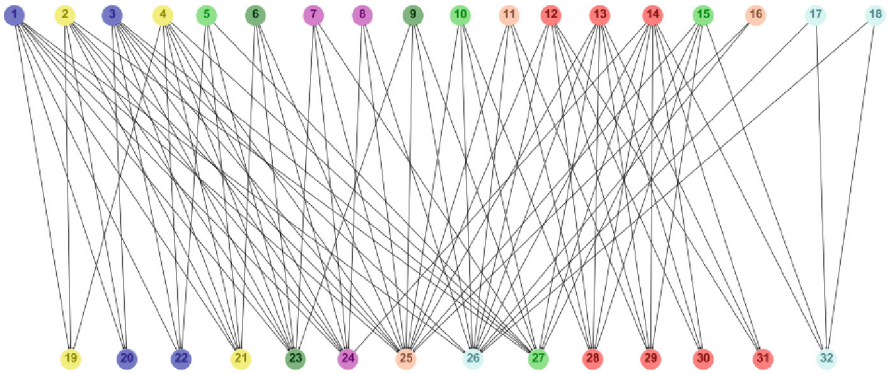


Fig. 2. Values of Barber versus Newman modularities obtained in a single run for the four real-world datasets (left - maximizing the Newman modularity; right - maximizing the Barber modularity).

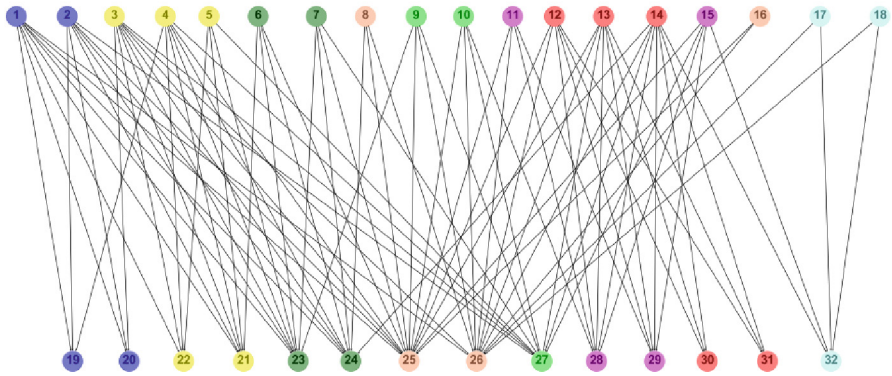
networks. Although not needed, these results confirm that, when dealing with bipartite networks the results reported when using the Barber modularity are better (with respect to this measure). Regarding the performance of NoisyEO, the results reported when using no modularity function are worse than those obtained when maximizing the modularity, which is not the case for general networks [21].

We further studied the connection between the two modularities. Figure 2 represents the values of the two modularities for individuals in the NoisyEO population when (left) maximizing the Newman modularity and (right) maximizing the Barber modularity for the four networks. A strong correlation is observed (and confirmed with correlation values over 0.85 in all cases). As expected the two modularities are also close. Since the interpretation of results is very difficult in the absence of the real community structure, and since it is known that the Newman modularity fails to represent the real community structure in some networks, in the presence of such a strong correlation maybe the two measures could be used in parallel until results could be objectively evaluated.

In the case of the southern woman network, most algorithms report the number of communities they identify: 2, 3 or 4 [2]. The BRIM algorithm [3] reports for this network also 4 communities with a modularity equal to 0.34554. Doreian et al. [6] reports a modularity value equal to 0.29390 and 3 communities. Communities detected by NoisyEO are represented in Fig. 3 showing few differences between covers, obtaining 8 communities. Regarding the other networks, we can only find interpretation of the detected communities and sometimes representations in a descriptive manner, without numerical results.



Newman modularity = 0.2056



Barber modularity = 0.2835

Fig. 3. The souther women network: communities detected by NoisyEO when maximizing the Newman modularity (top) and Barber modularity (bottom) (8 communities in both cases).

4 Conclusions

In this paper we are exploring the use of a community structure detection algorithm designed for unweighted and undirected networks for finding the structure of bipartite networks. By simply replacing the Newman modularity with the Barber modularity we find promising results, but to improve the method more attention to the specificities of bipartite networks has to be paid. The problem also requires attention: benchmarks with known structures are needed to evaluate the performance of different methods and allow quantitative comparisons, as a higher modularity may not always represent a better structure.

Acknowledgments. This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS - UEFISCDI, project number PN-II-RU-TE-2014-4-2332.

References

1. American revolution network dataset–KONECT (2016)
2. Alzahrani, T., Horadam, K.J.: Community detection in bipartite networks: algorithms and case studies. Underst. Complex Syst. **73**, 25–50 (2016)
3. Barber, M.J.: Modularity and community detection in bipartite networks. Phys. Rev. E **76**(6), 066102 (2007)
4. Barnes, R., Burkett, T.: Structural redundancy and multiplicity in corporate networks. Connections **30**(2), 4–20 (2010)
5. Davis, A., Gardner, B.B., Gardner, M.R., Warner, W.L.: Deep South: A Sociological Anthropological Study of Caste and Class. University of Chicago Press, Chicago (1941)
6. Doreian, P., Batagelj, V., Ferligoj, A.: Generalized blockmodeling of two-mode network data. Soc. Netw. **26**(1), 29–53 (2004)
7. Faust, K.: Centrality in affiliation networks. Soc. Netw. **19**(2), 157–191 (1997)
8. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
9. Jakalan, A., Gong, J., Su, Q., Hu, X., Abdelgder, A.M.S.: Social relationship discovery of IP addresses in the managed IP networks by observing traffic at network boundary. Comput. Netw. **100**, 12–27 (2016)
10. Kheirkhahzadeh, M., Lancichinetti, A., Rosvall, M.: Efficient community detection of network flows for varying Markov times and bipartite networks. Phys. Rev. E - Stat. Nonlin. Soft Matter Phys. **93**(3), 032309 (2016)
11. Lancichinetti, A., Fortunato, S., Kertész, J.: Detecting the overlapping and hierarchical community structure in complex networks. New J. Phys. **11**(3), 033015 (2009)
12. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E **78**, 046110 (2008). http://link.aps.org/doi/10.1103/PhysRevE.78.046110
13. Li, Z., Wang, R.S., Zhang, S., Zhang, X.S.: Quantitative function and algorithm for community detection in bipartite networks. Inf. Sci. **367–368**, 874–889 (2016)
14. Lung, R.I., Suciu, M., Gaskó, N.: Noisy extremal optimization. Soft Comput. 1–18 (2015). http://dx.doi.org/10.1007/s00500-015-1858-3
15. Marotta, L., Miccichè, S., Fujiwara, Y., Iyetomi, H., Aoyama, H., Gallegati, M., Mantegna, R.N.: Bank-firm credit network in Japan: an analysis of a bipartite network. PLoS ONE **10**(5), e0123079 (2015)
16. Newman, M.E.J.: Modularity and community structure in networks. Proc. Nat. Acad. Sci. **103**(23), 8577–8582 (2006)
17. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 026113 (2004)
18. Pan, W., Li, B., Jiang, B., Liu, K.: Recode: software Package refactoring via community detection in bipartite software networks. Adv. Complex Syst. **8**(2), 1450006 (2014). http://www.scopus.com/inward/record.url?eid=2-s2.0-84900352221&partnerID=40&md5=dd098d7838876fa64cb78c8e5c4d4bd1

19. Pang, Y., Bai, L., Bu, K.: An energy model for detecting community in PPI networks. In: Chen, Q., Hameurlain, A., Toumani, F., Wagner, R., Decker, H. (eds.) DEXA 2015. LNCS, vol. 9261, pp. 119–127. Springer, Heidelberg (2015). doi:10.1007/978-3-319-22849-5_9
20. Ren, Y.L., Sun, T.X., Gao, Y., Zheng, J.L., Shu, H.P., Liang, F.R.: Our considerations about research on regularities of acupoint combination based on bipartite network community structure partition. Zhen ci yan jiu = Acupuncture Research/(Zhongguo yi xue ke xue yuan Yi xue qing bao yan jiu suo bian ji) **39**, 148–152 (2014)
21. Suciu, M., Lung, R.I., Gaskó, N.: Game theory, extremal optimization, and community structure detection in complex networks. In: Proceedings of the Genetic and Evolutionary Computation Conference 2016, GECCO 2016, pp. 405–412. ACM, New York (2016). http://doi.acm.org/10.1145/2908812.2908878

An Approach for Incorporating the Usability Optimization Process into the Model Transformation

Marwa Hentati[1(✉)], Lassaad Ben Ammar[1], Abdelwaheb Trabelsi[2], and Adel Mahfoudhi[3]

[1] CES Laboratory, National School of Engineering, University of Sfax, Sfax, Tunisia
marwahentati@yahoo.fr, lassaad.ben-ammar@ceslab.org
[2] College of Computation and Informatics, Saudi Electronic University, Dammam, Saudi Arabia
Abdelwaheb@gmail.com
[3] College of Computers and Information Technology, Taif University, Taif, Saudi Arabia
a.mahfoudhi@tu.edu.sa

Abstract. Model transformation plays an important role in the model-driven engineering (MDE) approach which has the capability to generate the system code from a conceptual model. In fact, it describes the process of converting one model into another of the same system. Considering a source model, there may be several ways to transform it into target models with the desired usability attributes. This paper proposed an original approach that considers the interactive systems generation process as a usability optimization problem by the MDE approach. The aims were to generate all possible concrete user interfaces from a given abstract user interface. Then, the usability optimization process selects the model transformation were the desired usability attributes are fulfilled.

Keywords: Model transformation · MDE · Usability optimization

1 Introduction

Nowadays, the Model-driven Engineering (MDE) framework is widely adopted in the Human-Computer Interaction (HCI) domain. This framework is proved to be quite appropriate [5]. In fact, this approach tends to produce user interfaces through the definition of models and their transformation into a less abstract level to the code. From a source model, there may be several ways to transform it into target models. Although they may be equivalent from the functional viewpoint, target models may differ in their usability attributes. In this context, the focus is generally placed on data and functional modeling, disregarding usability aspects. Therefore, there is a need to extend the MDE process in order to promote usability aspects as a first class entity in the development process.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_87

The present paper aims to delineate an approach that addresses these weaknesses by extending the MDE process in order to optimize the usability of the user interfaces. The proposed approach is intended to optimize the usability from the conceptual model. Therefore, the selection of the adequate alternative transformation depends on the usability properties, which are able to convey. Hence, the proposition is structured in three main stages: presenting all the possible transformation associated with each abstract construct, quantifying the UI usability value by means of a fuzzy system and selecting the adequate alternative transformation able to generate the optimal usability UI.

We structure the remainder of this paper as follows. Section 2 presents a state of the art on approaches pertaining to incorporate usability into model transformation and some others to incorporate optimization process into model transformation. Section 3 provides a description of the different stages of the proposed approach. In order to show the usefulness of our proposal, a case study is presented in Sect. 4. Finally, Sect. 5 presents the conclusion and the future research work.

2 Literature Review

2.1 Approaches for Incorporating Usability into Model Transformation

There are currently several research studies that have dealt with usability in MDE environment such as in [13], the authors have addressed the usability features related to the system functionality, which may involve important changes in the system architecture. They use the term FUFs (Functional Usability Features) to indicate this type of usability features. The conceptual model can be seen as the input of a model compiler that can generate the software application automatically (or semi-automatically). In [14] a set of the well-known FUFs are gathered in a usability model that is included from an early stage of a holistic MDD method. The authors in [2] have suggested an approach which integrate usability issues as a part of a user interface development process from the conceptual models. The proposed approach is structured in two main contributions: The first one is called usability-driven model transformation. It aims to ensure that the generation of a user interface which maximizes the required level of usability [1]. The second contribution aims to perform the usability evaluation from the conceptual models by defining metrics for conceptual primitives that constitute these models [3]. In [7], the authors have proposed an architecture to perform the quality of the model-driven transformation. The aim was to dene a set of artifacts and a process for specifying and executing model transformations. The selection of the alternative transformations is guided by quality attributes. Although the previously mentioned works are useful in the interplay between the usability and MDD paradigm, they lack precise details about the selection of the alternative transformation. In fact, all of them try to select the alternative transformation without addressing the problem of the choice optimization which

maximizes the desired usability properties. This problem is the object of some initiatives in the research field among which we quote [6,15,16].

2.2 Approaches for Incorporating Optimization Process into Model Transformation

In [6], the authors implemented the SUPPLE system that automatically generates a multi-target user interface using the functional interface specification as input. The user interface generation is treated as an optimization problem that adapted to a person's abilities, devices, preferences, and tasks. Therefore, the SUPPLE system renders each element from the functional interface specification into an appropriate concrete widget according to a defined matching function. In the same context, [15] proposed to optimize the usability of the graphical user interface (GUI). Authors propose a new iterative process for creating a (high-level) interaction design in the context of the model-driven generation of GUIs. The main goal is to transform each element from the user interface specification model to an optimal GUI component by means of a structural screen model. The optimization function has been formulated on the basis of the manipulation and navigation time. Face to the increasing use of small screen devices, authors in [16] suggested an approach for extending and incrementing the Discourse-based communication models transformation process. This approach introduces straight-forward optimization techniques into the MDD of GUIs to reduce usability problems. Hence, the proposed approach allows the optimization of the generated GUIs for small screen devices in such a way as to utilize the given space and to minimize navigation and scrolling. After reviewing the aforementioned studies, it becomes clear that optimizing the usability of user interfaces in the MDE process is still an immature area. In fact, a limited usability aspect may not ensure the usability of the final system. It is within this context that our research work lies to introduce a framework addressing the problem of usability optimization in an MDE-compliant method.

3 Optimizing Usability in Model-Driven Development Framework

The main target of this paper is the definition of an approach to optimize the usability of user interfaces at the conceptual model which covers the AUI (Abstract User Interface) and the CUI (Concrete User Interface) models which represent an abstraction of the user interface. This approach is divided into three main stages: (1) definition of all the possible transformation associated with each abstract construct (2) definition of an early usability evaluation using a fuzzy logic system (3) Selecting the alternative transformation able to generate a user interface with an optimal usability. In the first stage, all the possible rule transformation of each CUI construct is established from a given AUI. Each one of them maximizes at least one usability property. Then, the second stage performs an early usability evaluation process taking as input the CUI established

in stage one. We opted to the fuzzy logic system for quantifying the usability of UI at the early stage of the development process. In the third stage, the user interface with the optimal value of usability will be selected as a result of the optimization process in a particular context of use.

3.1 Model-Driven Engineering: The Cameleon Framework

Recent studies indicate that the adoption of model-driven engineering (MDE) has increased. It offers a promising approach for the development of software application. It promotes the use of models and model transformations as the primary artifact to build and maintain software systems. MDE tools have many benefits for a software developing company: reduced development time, reduced costs, higher customer satisfaction, higher quality and hence improved competitiveness. MDE paradigm is currently being adopted in the human-computer interaction (HCI) domain to support user interface (UI) specification and engineering activities [11]. In this context, a renowned approach is the CAMELEON reference framework which provides a unifying framework developed by the EU-funded CAMELEON project [5]. Furthermore, it describes a framework that serves as a reference for classifying user interfaces that support multiple targets, or multiple contexts of use on the basis of a model-based approach. CAMELEON structures the user interface development life cycle into four levels of abstraction which are: the tasks and concepts (T&C) level, the abstract user interface level (AUI), the Concrete User Interface level (CUI) and the final user interface (FUI).

3.2 Model-Driven Transformation with Alternatives

Model transformation is a key mechanism when building software systems with MDE approach. It favors iterative refinement of models by a model transformation. The model transformation consists of a set of transformation rule which describes all the possible alternative transformations for a given domain. The elaboration of the transformation rules identifies the corresponding construct in the CUI model for each construct from the AUI model. There may be more than one construct in the CUI model that can be associated with the construct from the AUI model. Thus, from the functional perspective, all possible transformation associated with an abstract construct are equivalent. However, they may differ from the non-functional perspective and they do not satisfy the same usability attributes. To better explain this stage, we take as an example the abstract component InputElement illustrated in Fig. 1.

3.3 The Early Usability Evaluation

The model transformation already defined allows producing more than one transformation rule associated with an abstract construct. Thus, it is required to provide a usability model that contains the desired usability attributes which should be maximized in the CUI model.

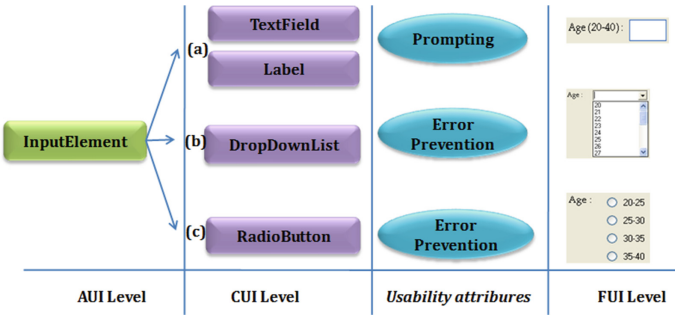


Fig. 1. Exemple of different model transformations for the input data element

The Proposed Usability Model. Since our objective is to evaluate usability at an early stage, we will focus on the product dimension. Therefore, we opt to use a usability model which extends the one presented in the ISO/IEC 9126-1 standard and [12]. This view is more deal with the characteristics of the product itself and can be used to evaluate the intermediate artifacts produced in an MDE development process. The entire model including all the sub-characteristics, measurable attributes, and their associated metrics is presented in [9,10]. The proposed usability model decomposes usability into four sub-characteristics that are defined as follow: ***Learnability***: the capability of the software product to enable the user to learn its application. ***Understandability***: the capability of the software product to enable the user to understand whether the system is suitable, and how it can be used for particular tasks and conditions of use. ***Operability***: the capability of the software product to enable the user to operate and control it. ***Attractiveness***: the capability of the software product to be attractive to the user. Since the sub-characteristics have been described abstractly, we have analyzed some usability guidelines presented in the literature in order to extract and adapt more detailed usability attributes. Table 1 shows the decomposition of the learnability sub-characteristic into three measurable attributes and its generic description.

Table 1. Decomposition of the learnability sub-characteristic

Sub-characteristic	Attribute	Generic description
Learnability	Prompting	Quantifing in terms of the proportion of label that displays supplementary information
	Predictability	Predicting the future actions by the number of the word in a label
	Informative feedback	Measuring in terms of actions with feedback response

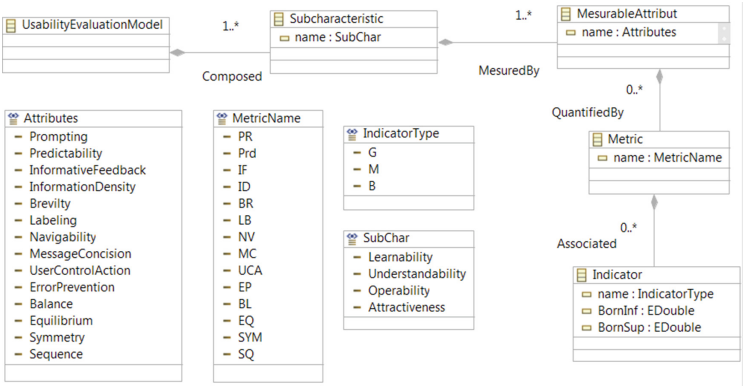


Fig. 2. Usability meta-model

In addition, the definition of the meta-model is a prerequisite in order to formalize the approach and to use the model in a productive way. Therefore, we propose the usability meta-model illustrated in Fig. 2. The proposed usability meta-model is composed of four classes: *Sub-characteristic:* a set of general properties of a software product by which its usability can be evaluated. It may be refined into multiple levels of measurable attributes. *Measurable Attribute:* a measurable usability property to which a metric can be assigned. *Metric:* a procedure for examining a software application to assign a value to each usability attributes. *Indicator:* a qualitative value assigned to the metric value used to classify the generated metrics and give it meaning.

A Fuzzy-System for the Usability Measurement. The aim of the usability measurement is to put metrics and attributes together in order to obtain a single usability value. The qualitative values obtained from the previously defined metrics seem to be insufficient to solve the problem of the usability measurement. Thus, a mechanism for combining the usability attribute is needed. The great challenge of combining the usability attributes relies on the fact that each one of them was an imprecisely defined concept and has fuzzy aspects [4]. Hence, the usability concept cannot be readily quantified into a numerical value. For these reasons, we opted for the fuzzy-systems theory as a mechanism for quantifying the UI usability. The usability attributes are usually characterized by a set of linguistic terms such as Good, Medium and Bad which were previously identified as indicators. The fuzzy-term set was G, M, B and the universe of discourse (UoD) was divided into three domains corresponding to the three different fuzzy sets. The entire usability measurement using the fuzzy mechanism is presented in [8].

Selecting the Optimal CUI. The entire CUI generated in stage one has a usability value. The higher one will be selected as the optimal CUI. It should be noted that the present stage is actually performed (semi)-automatically.

The selecting stage will be an exhaustive task with the increasing number of CUI possible combinations. That is why we aim to automate this stage by implementing the optimization process in the future.

4 An Illustrative Case Study

The main goal of this section is to illustrate the feasibility of our proposal. The object of the case study is a Tourist Guide System (TGS). The system will run on terminals of customers (laptop, mobile phone, PDA, etc.). Since the (TGS) was large, we focused our interests on the generation of the CUI for the Search itinerary task. We suppose to have the AUI presented in Fig. 3 as a result of the transformation of the Search itinerary task. Therefore, the result of the transformation is an XML file which is in accordance with the AUI meta-model (left part of Fig. 3). An editor with the Graphical Modelling Framework (GMF) of Eclipse was developed in order to better clear up the user interface layout. The sketch of the user interface presented by the editor is shown in the right part of Fig. 3.

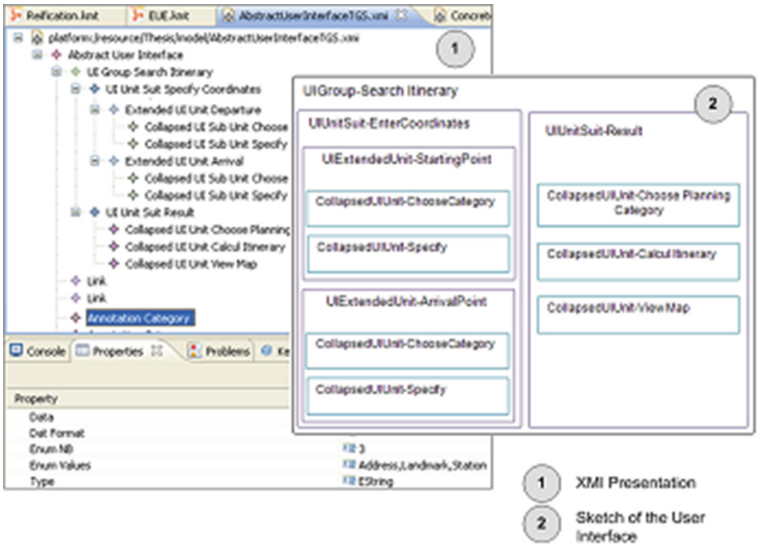


Fig. 3. The abstract user interface model for TGS

The AUI contains an abstract container *UIGroup* called Search itinerary which gives access to two abstract containers *UIUnitSuit* called Enter Coordinates and Result. The Enter Coordinates container gives access to specify the starting and the destination points. The tourist should choose the category (Address, Landmark, and Station) before specifying the starting or destination

point. The validation of the coordinates allows tourist to choose the planning (Pedestrian, Cyclable, Vehicle, Metro, Train, and Bus). To better explain our proposal, we started the case study by a CUI model transformation which did not take into account our proposed usability optimization in a specific context of use. The context is the following: a laptop as an interactive device (normal screen size), a tourist with a low level of experience of using the TGS. Although the deducted CUI filled their functional objectives, it did not satisfy our proposed usability optimization process. Each input element was transformed into *UIFieldEdit* in the CUI model. Since novice users should be protected against errors while entering data, the input elements with limited values should be displayed with list box elements such as dropdown list and radio button. Moreover, novice users need to be well informed and guided by the presence of supplementary information displayed in the labels *UIStaticField* to guide the user during the interaction with the system. The execution of a proposed approach leads to the generation of an optimal CUI which contains concrete elements that fulfil the desired usability attributes. Figure 4 illustrates such transformation taking into account the computing device and the target population characteristics. The second transformation to be conducted takes an iPAQ Hx 2490 Pocket PC as platform and a customer with a low level of experience of using the TGS as a target population. Hence, the migration to such platform raises a new redistribution of the user interface elements. The small screen size (240×320) is not sufficient to display all information. In Fig. 5, the CUI with an optimal usability for small screen group a limited number of concrete components, whose maximum number of concepts can be manipulated (5 in the case of iPAQHx2490 Pocket PC). To ensure the information density, the user interface elements are redistributed on several windows. The redistribution of interface elements on several windows will bring more steps to reach the goal by the aims of navigability elements (eg. Next, Return).

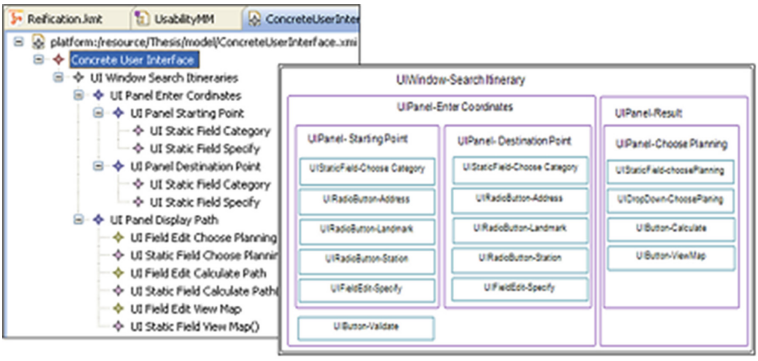


Fig. 4. An optimal concrete user interface for normal screen size

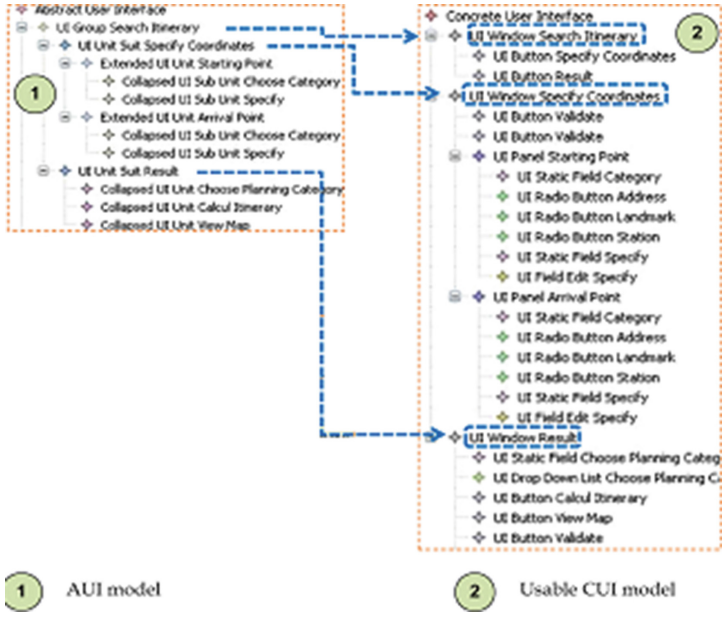


Fig. 5. An optimal concrete user interface for small screen size

5 Conclusions

This paper presents an approach for optimizing the usability aspects as a part of an MDE process. The proposed approach extends the Cameleon reference framework by integrating the usability optimization process from the conceptual models. This proposition gathers three main stages: the first one aims to define all possible CUI that can be generated from a given AUI. Then, a usability optimization process was performed by means of metrics in order to quantify it. These metrics was interpreted by defining a mechanism of indicator taking into account the target population and the computing device characteristics. The qualitative values obtained from the defined metrics were insufficient to measure the final usability values. For this reason and because the usability is a fuzzy aspect, we opted to the fuzzy system for measuring the UI usability.

References

1. Ammar, L.B., Mahfoudhi, A.: Early usability evaluation in model driven framework. In: Proceedings of the 15th International Conference on Enterprise Information Systems, ICEIS 2013, Angers, France, 4–7 July 2013, vol. 3, pp. 23–30 (2013)
2. Ammar, L.B., Trabelsi, A., Mahfoudhi, A.: Incorporating usability requirements into model transformation technologies. Requirements Eng. **20**(4), 465–479 (2015)

3. Ammar, L.B., Trabelsi, A., Mahfoudhi, A.: A model-driven approach for usability engineering of interactive systems. Softw. Qual. J. **24**(2), 301–335 (2016)
4. Chang, E., Dillon, T.S.: A usability-evaluation metric based on a soft-computing approach. IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. **36**(2), 356–372 (2006)
5. Favre, J.-M.: Towards a basic theory to model model driven engineering. In: 3rd Workshop in Software Model Engineering, WiSME, pp. 262–271. Citeseer (2004)
6. Gajos, K.Z., Weld, D.S., Wobbrock, J.O.: Automatically generating personalized user interfaces with supple. Artif. Intell. **174**(12), 910–950 (2010)
7. Gonzalez-Huerta, J., Blanes, D., Insfran, E., Abrahão, S.: Towards an architecture for ensuring product quality in model-driven software development. In: Proceedings of the 11th International Conference on Product Focused Software, pp. 28–31. ACM (2010)
8. Hentati, M., Ammar, L.B., Trabelsi, A., Mahfoudhi, A.: A fuzzy-logic system for the user interface usability measurement. In: 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 133–138. IEEE (2016)
9. Hentati, M., Ammar, L.B., Trabelsi, A., Mahfoudhi, A.: Model-driven engineering for optimizing the usability of user interfaces. In: ICEIS - Proceedings of the 18th International Conference on Enterprise Information Systems, Rome, Italy, 25–28 April 2016, vol. 2, pp. 459–466 (2016)
10. Hentati, M., Trabelsi, A., Ammar, L.B., Mahfoudhi, A.: Towards optimizing the usability of user interface generated with model-driven development process. In: 8th International Conference on Human System Interaction (HSI), pp. 206–212. IEEE (2015)
11. Hussmann, H., Meixner, G., Zuehlke, D.: Model-Driven Development of Advanced User Interfaces, vol. 340. Springer, Heidelberg (2011)
12. Ngo, D.C.L., Samsudin, A., Abdullah, R.: Aesthetic measures for assessing graphic screens. J. Inf. Sci. Eng **16**(1), 97–116 (2000)
13. Panach, J.I., Aquino, N., Pastor, Ó.: A proposal for modelling usability in a holistic MDD method. Sci. Comput. Program. **86**, 74–88 (2014)
14. Panach Navarrete, J.I., Juristo Juzgado, N., Pastor López, O.: Including functional usability features in a model-driven development method. In: Computer Science and Information Systems, vol. 10, pp. 999–1024. ComSIS Consortium (2013)
15. Petter, A., Behring, A., Zlatkov, M., Steinmetz, J., Mhlhuser, M.: Modeling usability in model-transformations. In: Proceedings of the 1st International Workshop on Non-functional System Properties in Domain Specific Modeling Languages, NFPinDSML, vol. 394, pp. 1613–0073 (2008)
16. Raneburger, D., Popp, R., Kavaldjian, S., Kaindl, H., Falb, J.: Optimized GUI generation for small screens. In: Hussmann, H., Meixner, G., Zuehlke, D. (eds.) Model-Driven Development of Advanced User Interfaces, vol. 340, pp. 107–122. Springer, Heidelberg (2011)

An ILS Heuristic for the Waste Collection Vehicle Routing Problem with Time Windows

Alba A. Campos(✉) and José Elias C. Arroyo

Campus Universitário, Universidade Federal de Viçosa,
Viçosa, MG 36570-900, Brazil
alba.campos@ufv.br, jarroyo@dpi.ufv.br

Abstract. This study addresses the Waste Collection Vehicle Routing Problem with Time Windows (WCVRPTW) with multiple disposal sites and drivers' lunch breaks. The problem consists of routing garbage trucks to collect customers waste within given time window in order to minimize the total travel cost. The waste is driven to disposal sites, multiple trips to disposal sites are allowed per day for the vehicles, and the vehicles must be empty when returning to the depot. To find near-optimal routes for the waste collecting vehicles, we propose a heuristic algorithm based on the meta-heuristics Iterated Local Search (ILS) and Variable Neighborhood Descent (VND). The performance of the proposed algorithm is evaluated by comparing its results to CPLEX commercial solver, and a variable neighbourhood tabu search (VNTS) algorithm addressed in the literature. The results obtained, on benchmark problems from the literature, are highly competitive and a number of new improved solutions are reported.

Keywords: Waste collection · Vehicle routing · Iterated Local Search · Meta-heuristics

1 Introduction

The waste collection vehicle routing problem with time windows (WCVRPTW), studied in this paper, was first introduced by Kim et al. [7]. The WCVRPTW can be characterized as a variation of vehicle routing problems with time windows and intermediate facilities. In this problem, the location of each node (depot, customers and disposal facilities) is known and each node has a time window which restricts the times at which a node is available to be visited by a waste collecting vehicle. Vehicles start from the depot and visit the customers to collect their waste demand. When a vehicle is full it must go to the nearest disposal site to unload the waste. After being emptied, the vehicles repeat the process (collecting and unloading) and so the vehicles can make multiple visits to the disposal sites per day. The vehicle drivers are assumed to take one hour lunch break.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_88

The vehicles return to the depot at the end of the working day. The objective of the problem is to determine feasible routes for the vehicles in order to minimize the total travel cost (or distance).

Since the vehicle routing problems with time windows is a NP-hard problem [3], the WCVRPTW studied in this paper is also NP-hard.

Kim et al. [7] studied the WCVRPTW and considered the following objectives: number of vehicles, total travel time, route compactness and workload balancing. They solved the problem by using an extension of Solomon's well-known insertion algorithm [10] and a capacitated clustering-based algorithm. Kim et al. [7] also proposed some real-world benchmark problems to test the algorithms. Ombuki-Berman et al. [9] considered the same problem by using a multi-objective genetic algorithm on the set of benchmark problems obtained by Kim et al. [7]. Benjamin and Beasley [1] proposed a greedy constructive algorithm to produce initial solutions and developed three meta-heuristic algorithms: tabu search (TS), variable neighborhood search (VNS) and variable neighbourhood tabu search (VNTS). They used the benchmark problems proposed by Kim et al. [7] and improved the results. Benjamin and Beasley [2] improved the effectivity of the algorithms TS, VNS and VNTS [1] by using a disposal facility positioning (DFP) procedure to evaluate routes. Buhrkal et al. [4] presented a mixed-integer linear programming (MILP) model for the WCVRPTW and proposed an adaptive large neighborhood search (ALNS) algorithm for solving the problem. They tested their ALNS algorithm on benchmark problems proposed by Kim et al. [7] and on real-life instances provided by a waste collection company. The ALNS algorithm performed better than the algorithms proposed by Benjamin and Beasley [1]. Recently, Han et al. [5] analyzed the major contribution about Waste Collection Vehicle Routing Problem in literature and analyzed different methods and techniques to solve the problem.

Waste collection is an important logistic activity within any city. Population growth has caused problems in the urban environment affecting the quality of million people lives. One of the problems arising from population growth is the accelerated production of municipal solid waste, especially in recent decades. According to a study by the United Nations, 20% to 30% of municipal budgets is committed to the collection and disposal of waste.

Motivated by the computational complexity and the practical relevance of the problem, in this paper we propose an Iterated Local Search (ILS) heuristic coupled with Variable Neighborhood Descent local search method to find good quality solutions. The effectiveness of the proposed heuristic algorithm is evaluated and compared by computational experiments on the benchmark problems proposed by Kim et al. [7].

This paper is organized as follows. Section 2 presents the problem definition and the MIPL formulation. Section 3 deals with the proposed algorithm, describing the constructive and local search procedures, as well as the perturbation mechanism adopted. Section 4 contains the results obtained and a comparison with the work found in the literature. Section 5 presents the concluding remarks of this paper.

2 Problem Statement and Mathematical Formulation

The WCVRPTW is formally defined as follows. There is a graph $G = (V, A)$, where V is the set of nodes, and $A = \{(i, j)|\forall i, j \in V, i \neq j\}$ is the set of arcs. $V = \{0\} \cup V^f \cup V^c$, where 0 is the depot node, $V^f = \{1, ..., m\}$ is the set of m disposal sites, and $V^c = \{m + 1, ..., m + n\}$ is the set of n customers. There is a set of nk vehicles $K = \{1, ..., nk\}$. Let t_{ij} and c_{ij} be the travel time and cost associated with arc (i, j), respectively. Each node $i \in V$ has a service time s_i and a time window $[a_i,b_i]$. The parameter q_i represents the amount picked up at a customer $i \in V^c$. C is the capacity of the vehicles, S is the number of customers to visit on each route, and R is the route capacity (total amount collected at customers for each route). The vehicle drivers have a lunch break with duration s^u starting in the interval $[a^u, b^u]$. The objective of the WCVRPTW is to find a set of routes for the vehicles in order to minimize total travel cost.

Figure 1 shows an example of a vehicle route in which 25 customers are served and 3 three disposal sites are used. Note that a vehicle starts from the depot and visit the customers and the disposal sites and returning the depot.

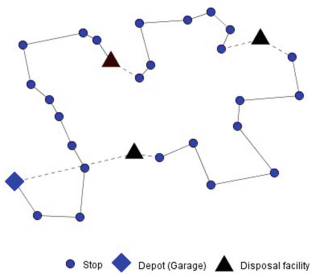


Fig. 1. Example of a vehicle route with multiples disposal sites.

The mixed-integer linear programming (MILP) model of the problem is originally from Buhrkal et al. [4]. The following decision variables are used within the model. $x_{ijk} = 1$ if the vehicle $k \in K$ uses arc $(i, j) \in A$, otherwise zero. d_{ik} represents the accumulative demand at node $i \in V$ for vehicle $k \in K$. w_{ik} represents the start time of service at node $i \in V$ for vehicle $k \in K$. The binary variable y_{ijk} indicates whether lunch was taken between visiting node i and j by vehicle k. The depot is split in a start and an end depot $(0, 0')$ and M is a big parameter. The mathematical model is defined according to Eqs. (1)–(20).

The goal of the model is to minimize the sum of the total the travel cost, expressed by the objective function (1). Constraints (2) ensure that all the vehicles must leave and return to depot (3). Constraints (4) ensure that each customer is visited exactly once. Constraints (5) ensure that inflow and outflow must be equal except for the depot nodes. Constraints (6) and (7) represent the time windows and service time. Constraints (8) ensure that the vehicles must be empty when start the route and when they return to the depot. Constraints (9)

accumulate demand for all nodes except the disposal sites. Vehicle capacity is given by constraint (10). Each vehicle driver must take exactly one lunch break (11). Constraints (12) ensure that a lunch break happens only between two nodes i and j if they are connected. Constraints (7) are modified to (13), which ensures that both possible break durations are taken into account. Constraints (14) e (15) ensure that lunch break happens only within time window. The number of customers serviced is limited in (16) and the route amount in (17). Constraints (18), (19) e (20) impose binary conditions, nonnegative integer constraint, and nonnegative conditions on the variable set.

$$Min \sum_{(i,j)\in A} \sum_{k\in K} c_{ij} x_{ijk} \tag{1}$$

$$\sum_{j\in V} x_{0jk} = 1, \qquad \forall k \in K \tag{2}$$

$$\sum_{i\in V} x_{i0'k} = 1, \qquad \forall k \in K \tag{3}$$

$$\sum_{i\in V} \sum_{k\in K} x_{ijk} = 1, \qquad \forall j \in V_c \tag{4}$$

$$\sum_{i\in V} x_{ijk} = \sum_{i\in V} x_{jik}, \qquad \forall j \in V_c \cup V_f, k \in K \tag{5}$$

$$a_i \leq w_{ik} \leq b_i, \qquad \forall i \in V, k \in K \tag{6}$$

$$w_{ik} + s_i + t_{ij} \leq w_{jk} + (1 - x_{ijk})M, \qquad \forall (i,j) \in A, k \in K \tag{7}$$

$$\sum_{i\in(0,0')} d_{ik} = 0, \qquad \forall k \in K \tag{8}$$

$$d_{ik} + q_i \leq d_{jk} + (1 - x_{ijk})M, \qquad \forall i \in V \setminus V_f, j \in V, k \in K \tag{9}$$

$$d_{ik} \leq C, \qquad \forall i \in V, k \in K \tag{10}$$

$$\sum_{(i,j)\in A} y_{ijk} = 1, \qquad \forall k \in K \tag{11}$$

$$y_{ijk} \leq x_{ijk}, \qquad \forall (i,j) \in A, k \in K \tag{12}$$

$$w_{ik} + s_i + y_{ijk} s^u + t_{ij} \leq w_{jk} + (1 - x_{ijk})M, \qquad \forall (i,j) \in A, k \in K \tag{13}$$

$$a^u + s^u + t_{ij} \leq w_{jk} + (1 - y_{ijk})M, \qquad \forall (i,j) \in A, k \in K \tag{14}$$

$$w_{ik} + s_i + t_{i,j} \leq b^u + (1 - y_{ijk})M, \qquad \forall (i,j) \in A, k \in K \tag{15}$$

$$\sum_{(i)\in V} \sum_{j\in V^c} x_{ijk} \leq S, \qquad \forall k \in K \tag{16}$$

$$\sum_{i\in V^f} d_{ik} \leq R, \qquad \forall k \in K \tag{17}$$

$$x_{ijk} \in \{0,1\}, \qquad \forall (i,j) \in A, k \in K \tag{18}$$

$$y_{ijk} \in \{0,1\}, \qquad \forall (i,j) \in A, k \in K \tag{19}$$

$$d_{ik} \geq 0, \qquad \forall i \in V, k \in K \tag{20}$$

3 Proposed Heuristic Algorithm

In this paper, to solve the WCVRPTW we proposed a heuristic, called ILS-VND, which is a hybridization of both the Iterated Local Search (ILS) meta-heuristic and the Variable Neighborhood Descent (VND) local search. ILS [8] consists of an iterative process that combines a perturbation phase and a local search phase. The VND heuristic [6] is used in the local search phase to improve the initial solution and the perturbed solutions. The pseudo-code of the proposed algorithm, called ILS-VND, is described in Algorithm 1. The ILS-VND starts with an INICIAL_SOLUTION procedure (line 2) in which an initial solution is generated by a constructive heuristic. This solution is improved by a VND procedure (line 3). The iterations of the algorithm are computed from line 4 to 10 until a stopping is satisfied. In each iteration the current solution is perturbed by the PERTURBATION procedure (line 5) and the perturbed solution is improved by a VND local search procedure (line 6) which uses three neighborhood structures. If the best solution is improved, it is updated (acceptance criterion) (lines 7–9). When d reaches a maximum value (dmax), it is set to its lowest value (lines 10 and 11). The best obtained solution is returned by the Algorithm (line 11). In the next subsections, we describe the procedures of our algorithm.

3.1 Initial Solution

The initial solution was constructed by using the greedy algorithm proposed by Benjamin and Beasley [1] which is described as follows.

1. Start a new vehicle route and initialize the various counters we need to keep track of the use made of the vehicle.
2. Try to insert lunch break.
3. Visit the feasible nearer customer without wait time. Repeat steps 2 and 3 until not find more feasible customers to be routed. If the vehicle uses at least 80% of the capacity and the disposal site is closer than the customer considered, go to step 4.
4. If vehicle is no empty visit nearest feasible disposal site.
5. The vehicle waits for a customer time window to open and then go to step 2. If no more viable customers can be inserted in the route, go to step 1.

3.2 VND Local Search

In the ILS-VND two VND local search procedures are used: VND1 and VND2. VND1 and VND2 are used to improve the initial solution and a perturbed solution, respectively. The VND1 local search is based on neighborhood structures based on exchange of two customers, who may or may not be on the same route. The VND2 local search uses the following neighborhood structures:

– M^1 (customer insertion): This neighborhood contains solutions obtained by removing a customer from a route and inserting it in another position of the same route.

– M^2 (disposal site insertion): This neighborhood is formed by removing a disposal site from a route and including it in another positions of the same route.
– M^3 (disposal sites exchange): Neighbor solutions are obtained by exchanging two disposal sites in a route.

In the VND local search procedures we use the concept of customer neighbors [1]. A customer neighbor set of size nc for customer i, denoted by $N(i, nc)$, is composed of the nc customers that are closest to customer i, but with compatible time windows. A customer j is defined to have a compatible time window with customer i (and hence potentially belongs to $N(i, nc)$) if it is possible to visit i at some time in its time window, service i and then go directly onto j to service j without waiting for its time window to open.

The steps of the VND1 local search are presented in Algorithm 2. The algorithm receives as input parameter a set K^* of values for the customer neighbor sizes. These values are used to limit the customer neighbor sizes. Then the array Γ is initialized with set of values of K^*. In the line 2, variable nc receives the lowest value of the set Γ (line 5). Then this value is removed of the set. In the next line the algorithm executes a local search, where all the customers are exchange. If the new route is feasible and the new solution is better than the initial solution, s^* receives a new solution and Γ is initialized with K^* (line 11). Otherwise, the exchange of customers is undone (line 14) and a new search is performed. At the end of the algorithm returns the best solution found (line 19).

Algorithm 1. ILS-VND(Max_Iter, K^*)

```
1  begin
2      s_0 ← INICIAL_SOLUTION();
3      s* ← VND1(s_0, K*);
4      while Max_Iter do
5          s' ← PERTURBATION(s*);
6          s'' ← VND2 (s');
7          if f(s'') < f(s*) then
8              s* ← s'';
9          end
10     end
11     return s*;
12 end
```

The VND2 local search is presented in Algorithm 3. The algorithm uses the set of neighborhoods $\{M^1, M^2, M^3\}$ (line 2), and systematically modifies the neighborhood structures that belong to this set in a deterministic way. In M^1 neighborhood all customers of a route are inserted in the positions before and after of the customers in $N(i, nc)$ on the same route. In M^2 each disposal site is inserted in each position of the route, and in M^3 disposal sites on a route are exchanged. The initial neighborhood M^1 ($k = 1$) is set in line 3. The loop in lines

Algorithm 2. VND1(s, K^*)

```
1  begin
2      Γ ← K*; (set of sizes)
3      while Γ ≠ ∅ do
4          nc ← min{k|k ∈ Γ};
5          Γ ← Γ − {nc};
6          for all customers i do
7              for all customers j ∈ N(i, nc) do
8                  s' ← solution obtained by exchanging i and j;
9                  Evaluate solution s';
10                 if f(s') < f(s) and s' is feasible then
11                     s* ← s'; Γ ← K*;
12                 end
13                 else
14                     Undo the exchange between the customers
15                 end
16             end
17         end
18     end
19     return s*;
20 end
```

4–12 investigates one neighborhood at a time, until a local optimum with respect to current neighborhood is found. In case an improving solution is found, the current solution is updated and the search resumes from the first neighborhood $k = 1$ (line 7). Otherwise, the order of the neighborhood is increased by one in line 10. The current solution s^* that is now locally optimal with respect to all neighborhoods is returned in line 13.

3.3 Pertubation Mechanism

The goal of the perturbation method in an ILS heuristic is escape from a local optimal solution. In this work, we used a perturbation mechanism based on random customers exchanges. This heuristic considers that each perturbation makes three exchanges of neighbors customers.

4 Computational Results

The proposed ILS-VND algorithm was tested on ten real benchmark problems generated by Kim et al. [7]. The test problems are available for download at[1].

The ILS-VND algorithm was coded in C++ programming language and executed in a PC Intel Core i7, CPU (3.4 GHz x 8), with 32 GB of RAM memory and

[1] https://sites.google.com/site/logisticslaboratory/research/research-areas/waste_collection_vrptw_benchmark.

Algorithm 3. VND2(s)

```
1  begin
2  |  Let {M^1, M^2, M^3} be the set of neighborhoods structures;
3  |  k ← 1; {current neighborhood}
4  |  while k ≤ 3 do
5  |  |  Find the best neighbor s' ∈ M^k of s;
6  |  |  if (f(s') ¡ f(s)) then
7  |  |  |  s ← s'; k ← 1;
8  |  |  end
9  |  |  else
10 |  |  |  k ← k + 1;
11 |  |  end
12 |  end
13 |  return s;
14 end
```

operating system Ubuntu Linux 15.10, 64 bits. The parameters of the ILS-VND algorithm were defined experimentally. In the VND1 local search, the following customer neighborhood sizes were user: 5, 10, 25, 50 ($K^* = \{5, 10, 25, 50\}$). The stopping condition of the algorithm is set to 20 iterations ($Max_Iter = 20$).

We compare the ILS-VND algorithm against three algorithms: the clustering-based algorithm (CBA) of Kim et al. [7], and the algorithms VNS and VNTS proposed by Benjamin and Beasley [2]. It is important to mention that, these algorithms were also tested on the benchmark problems generated by Kim et al. [7]. The performance of each algorithm is measured by the *Relative Percentage Deviation (RPD)* that is computed as: $RPD\% = \frac{f_{algorithm} - f_{best}}{f_{best}} \times 100$, where

Table 1. RPDs of the algorithms: results for the problems of Kim et al. [7].

Problem: ($\|V^c\|$,$\|V^f\|$)	CBA	VNS	VNSTS	ILS-VND
P1: (99, 2)	30.70	16.90	0.00	53.20
P2: (275,1)	33.80	17.80	15.30	0.00
P3: (330, 4)	9.80	9.20	0.00	4.30
P4: (442, 1)	19.50	22.30	9.40	0.00
P5: (784, 19)	21.20	19.10	1.10	0.00
P6: (1048, 2)	26.30	19.90	13.10	0.00
P7: (1347, 3)	58.20	39.30	33.10	0.00
P8: (1596, 2)	20.90	16.90	0.00	10.90
P9: (1927, 4)	34.40	21.70	7.80	0.00
P10: (2092, 7)	25.40	19.60	6.58	0.00
Average	28.02	20.27	8.64	6.84

f_{best} is the minimum cost value (objective function) obtained among all the algorithms, and $f_{algorithm}$ is the cost obtained with a given algorithm.

The obtained RPDs for the algorithms CBA, VNS, VNTS and ILS-VND are reported in Table 1. For each problem, the number of customer $|V^c|$ and number of disposal sites $|V^f|$ are presented in the first column. We can see that the proposed ILS-VND algorithm obtains the best solution for 8 out of 10 problems. ILS-VND is better than VNTS on 8 problems. Also, we can observe that the algorithm VNTS is better than VNS which in turn is better than CBA.

Table 2. Results of VNTS and ILS: number of vehicles (nv), total travel costs and CPU times (s).

Problem	VNTS			ILS-VND		
	nv	Total travel cost	CPU time (s)	nv	Total travel cost	CPU time (s)
P1	3	156.90	16.00	3	240.50	0.00
P2	3	454.50	172.00	2	394.00	0.04
P3	6	186.70	219.00	6	194.80	0.04
P4	11	79.70	273.00	10	72.80	0.12
P5	5	641.80	2152.00	4	634.50	0.96
P6	17	2123.80	637.00	17	1876.80	2.07
P7	8	874.70	2698.00	8	656.80	3.83
P8	13	1206.10	2556.00	12	1338.30	5.26
P9	16	1127.70	6541.00	16	1037.50	18.41
P10	16	1558.10	6069.00	16	1461.90	17.21
Average		841.00	2133.30		790.79	4.79

Since VNTS and ILS-VND are the two best algorithms, in Table 2 we compare the relative values of the objective function (total travel cost) determined by these algorithms. In this Table we also present the number of vehicles (nv) used in each solution, and the CPU time (seconds) consumed by each algorithm. As can be seen in this Table 2, VNTS is better than ILS-VND only on three problems (P1, P3 and P8), and it spends the most CPU time on all the problems. The solutions found by the two algorithms use the same number of vehicles, except on problems P2, P4, P5 and P8, where the solutions of ILS-VND use one vehicle less.

In this paper, we also generate 10 small-size problems in order to compare our algorithm ILS-VND with the mathematical MILP model presented in Sect. 2. The CPLEX solver is applied to solve the MILP model with a threshold CPU time of 3600 s for each small problem. That is, if after the established time no optimal solution is obtained, the best current solution (upper bound) is returned by CPLEX. Table 3, shows the obtained experimental results for small-size problems. Column 1 gives the size of the problems (number of customer $|V^c|$ and

Table 3. Results for small-size problems.

Problems ($\|V^c\|$,$\|V^f\|$)	CPLEX			ILS-VND		
	nv	Total travel cost	CPU time (s)	nv	Total travel cost	CPU time (s)
p1: (20, 2)	1	91.00	16.00	1	132.60	0.01
p2: (20, 1)	1	47.10	3600.00	1	47.10	0.01
p3: (20, 4)	1	23.60	0.43	1	23.90	0.01
p4: (20, 1)	1	98.00	3600.00	1	100.00	0.01
p5: (20, 2)	1	49.20	3600.00	1	55.70	0.01
p6: (20, 3)	1	29.30	273.10	1	35.50	0.01
p7: (20, 2)	1	101.70	27.63	1	103.60	0.01
p8: (20, 4)	1	150.10	3.36	1	160.60	0.02
p9: (20, 7)	1	23.70	0.51	1	23.90	0.01
p10: (17, 4)	1	50.40	3600.00	1	50.40	0.01
Average		66.41	1472.10		73.33	0.01

number of disposal sites $|V^f|$). For each algorithm we show the number of vehicles (nv), the total travel cost and the consumed CPU time (seconds). The CPLEX found the optimal solution for 6 problems (p1, p3, p6, p7, p8 and p9). We can see that the solutions found by ILS-VND are very near to the CPLEX solutions. For all the small problems, the proposed algorithm spends the least CPU times, on average 0.01 s. We also can see that the solutions of CPLEX and ILS-VND only use one vehicle.

5 Conclusions

In this paper we have studied a real life waste collection vehicle routing problem with time windows using benchmark problems from Kim et al. [7]. To solve this problem, we proposed a hybrid algorithm based on the meta-heuristics Iterated Local Search (ILS) and Variable Neighborhood Descent (VND). Experimental results showed that the proposed ILS-VND algorithm is capable of producing better solutions compared with the benchmark algorithms proposed by [2]. For the 10 benchmark problems, 7 new better solutions were generated by ILS-VND. As future research we suggest to use other perturbation mechanisms, generate new test problems, and develop other meta-heuristics to solve the problem.

Acknowledgments. The authors thanks the financial support of CNPq, CAPES and FAPEMIG, Brazilian research agencies.

References

1. Benjamin, A.M., Beasley, J.E.: Metaheuristics for the waste collection vehicle routing problem with time windows, driver rest period and multiple disposal facilities. Comput. Oper. Res. **37**(12), 2270–2280 (2010)
2. Benjamin, A.M., Beasley, J.E.: Metaheuristics with disposal facility positioning for the waste collection VRP with time windows. Optim. Lett. **7**(7), 1433–1449 (2013)
3. Cordeau, J.-F., Gendreau, M., Laporte, G., Potvin, J.-Y., Semet, F.: A guide to vehicle routing heuristics. J. Oper. Res. Soc. **53**(5), 512–522 (2002)
4. Buhrkal, K., Larsen, A., Ropke, S.: The waste collection vehicle routing problem with time windows in a city logistics context. Procedia-Soc. Behav. Sci. **39**, 241–254 (2012)
5. Han, H., Ponce Cueto, E.: Waste collection vehicle routing problem: literature review. PROMET-Traffic Transp. **27**(4), 345–358 (2015)
6. Mladenović, N., Hansen, P.: Variable neighborhood search. Comput. Oper. Res. **24**(11), 1097–1100 (1997)
7. Kim, B.I., Kim, S., Sahoo, S.: Waste collection vehicle routing problem with time windows. Comput. Oper. Res. **33**(12), 3624–3642 (2006)
8. Lourenço, H.R., Martin, O.C., Stützle, T.: Iterated local search. In: Glover, F., Kochenberger, G.A. (eds.) Handbook of Metaheuristics, vol. 57, pp. 320–353. Springer, Heidelberg (2003)
9. Ombuki-Berman, B.M., Runka, A., Hanshar, F.T.: Waste collection vehicle routing problem with time windows using multiobjective genetic algorithms. Brock University 2007; Technical report CS-07-04, pp. 91–97 (2007)
10. Solomon, M.M.: Algorithms for the vehicle routing and scheduling problem with time window constraints. Oper. Res. **35**, 254–265 (1987)

An NLP-Based Ontology Population for Intentional Structure

Noura Labidi[1(✉)], Tarak Chaari[2], and Rafik Bouaziz[1]

[1] MIR@CL Laboratory, FSEGS, B.P. 1088, 3018 Sfax, Tunisia
noura.labidi@yahoo.fr, rafik.bouaziz@usf.tn
[2] ReDCAD Laboratory, University of Sfax, B.P. 1173, 3038 Sfax, Tunisia
tarak.chaari@redcad.org

Abstract. This paper describes an automatic approach that populates an ontology of intentions from textual client request in IT market. This approach is based on an ontology structure that models clients' intentions. It combines NLP (Natural Language Processing) techniques to populate the ontology by the client's intention recognized from an English written request. Our automatic approach ensures the segmentation, the analysis, the extraction of semantic components and finally the population of the ontological structure by the new instance. This last step is devoted to the building of an intentional instance in conformity with the ontology structure by applying a set of linguistic rules. An experiment on clients' requests from an online forum was considered to illustrate the efficiency of the proposed approach.

Keywords: Intentional structure · Ontology Population(OP) · Knowledge modeling · Intention recognition

1 Introduction

Nowadays, despite technological advances, clients still unaware of the technical vocabulary of suppliers. So, they find many offers with complex terms. On the other hand, providers have to well understand clients' requirements in order to be able to propose to them the best offer. Ontologies can help in bridging the gap between provider's offers and client's needs. With ontologies, we can match between their different terms by generating correspondences between the client and the provider terms. In this way, clients can express freely their request by using their own language and knowledge. To facilitate the understanding of clients' intention, a knowledge representation of client's intentions is needed. For that aim, we proposed an approach whose objective is to detect client's intention from its written text and to transform it to a formal representation (intentional structure) to be matched later with providers terms to find the suitable offer to its needs.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_89

Our work is focused more precisely on the concept of intentional structure which represents a structure of client intention ontology. This structure could be populated with the new instance extracted.

The proposed methodology uses linguistics knowledge and automatic natural language processing techniques. Its originality is in the fact that it treats one of the challenging tasks in the NLP field which is "understanding natural language texts". Our work focuses mostly on ontology population process by an intention detected from client free English request. The entire process of the system is made in an automatic and a semantic way.

The remainder of this paper is structured as follows. We begin first of all, by the related works. Section 3 describes in details our approach of OP. Experiments and results are reported in Sect. 4. Section 5 concludes the paper.

2 Related Works

2.1 Intention Recognition

Extracting intentions belongs to research issues pursued by multiple projects. The majority of works tackling the intention recognition aims to improve online services for users of web services. Some systems are based on methods based on NLP techniques, some others are based on learning techniques and statistical methods and others are based on classification techniques. Authors in [4] built an *Intention Insider* system architecture which is intended to be a cloud service. The tool used natural language processing (NLP) and text mining techniques [6] to extract intention from online forums to discover people's intentions in the social channel. The authors of this work were based on intention templates for extracting intentions of members of the forums without defining a specific intentional structure.

In [13] in order to extract user intention from terms in web service query, the authors proposed a goal-driven intention extraction approach. These works aim to satisfy user's intention by knowing which goal should be achieved within a multi-agent architecture named *BDIAgent* (Believe-Desire-Intention). This work is different from our issue because the input of the system is user-submitted keyword queries which do not exceed 3 or 4 words.

In order to improve the performance of an information research system, the authors in [10] use learning techniques and statistical methods to identify the intentions of the authors of scientific publications in the field of computer science. The authors built a recognition system called *RICAD* [8] based on a semi-automatic method to recognize the authors intentions from written scientific documents. The output of this method intends to enrich a knowledge base by new detected intentions. The final result of this approach is an ontology of intentions. The authors of this work consider the intention composed of goal, mean and reason but they do not explicit the intention' components in the ontology proposed.

Some related methods are also proposed in [9,14]. They also treat the user request by classifying a query submitted to a search engine to determine the

semantic intent of Web queries. In the same issue of intention classification but of social media posts, an approach based on intent classification from a paragraph written by a user of the forum is proposed in [5]. The authors focus on identifying intention from posts in online discussion forums; they propose a new transfer learning method (Co-class). Authors in that work are interested only in the intention "to buy", which restricts the extraction process. On the other hand, the labeling process is a manual process which is so hard, time consuming and expensive.

Our analysis of the state of the art led us to conclude that each work cited has both advantages and weaknesses: in some contributions such as in [4,11] many manual interventions are needed which is laborious, tedious and long. Therefore, an automation of intention recognition process is needed. *BDI Agent* is an automatic system for detecting and extracting intentions. However, it needs to know the user's profile to detect its intention. Such information may not be available if a client simply expresses his needs by writing a free text without leaving traces on its own characteristics. In addition, in the solutions proposed in [4,13], the authors are interested only in one component of the intention (the goal or the object) and they do not define a structure for the intention which enables a more understanding for the client' request. Contrariwise, the author in [11] defines a structure for intention and specifies its components. But, the proposed system (*RICAD*) is not able to provide the components of the intention extracted explicitly which requires an additional effort and time.

2.2 NLP Techniques for Ontology Population

In the context of ontology evolution, OP consists of associating new instances to concepts and relationships into an existing ontology. This task can be accomplished after the conceptual model of ontology is built. This issue has led to different approaches which are differentiated by the degree of automation and techniques used and the intended use of the ontology.

The project ArtEquAKT [1] is a system that extracts automatically knowledge about artists from the Web to generate personalized biographies. It uses a combination of syntactic and semantic analysis to populate a knowledge base by extracting named entity as instances of concepts. It requires a semantic knowledge of the domain to determine binary relationships between two instances.

Navigli and Velardi in [16] describe an automatic pattern-based method for automatic enrichment of a core ontology with the definitions of a domain glossary. It extracts instances from semi-structured corpora with the help of manually developed constraining rules as patterns. The system developed is domain specific.

SOBA [3] is an information extraction system that automatically populates a knowledge base by using information extracted from heterogeneous sources. It is devoted to extract named entities related to soccer matches and soccer-specific events.

Authors in [15] describe the use of language processing tools for the population of an ontology in the domain of Risk Management. The approach proposed is

semi-automatic and based on combined NLP techniques. The main idea consists in focusing on verbs in the instantiation process. More recently, in [7], authors propose a domain independent approach for the automatic population of ontologies. This process applies NLP and information extraction techniques to classify ontology instances. The proposed process is domain-independent.

Authors in [2] propose a lexicon-grammar based methodology for efficient information extraction and retrieval on unstructured medical records. They describe an NLP methodology for extracting RDF (Resource Description Framework) triples from unstructured medical records and populating an existing ontology built by a domain expert. The populated ontology will be used in e-health applications.

OP approaches can be compared based on many criteria. For example, we can differentiate works by the elements extracted. In fact, the majority of works can populate an ontology with instances of both concepts and relations. We can also mention those that are specialized in extracting one category of words such as verbs or named entities. Considering the degree of automation criterion, most of the cited works are automatic except the work in [15] which is semi-automatic. According to another criterion which is domain portability, some works are domain independent (such as [7,15]) while others are domain specific [2,3,16]. However, the project ArtEquAKT [1] is characterized by a limited domain portability. Those approaches are dedicated to the population of domain ontologies which is different from our context of work. Moreover, none of the cited works is devoted to the management of intentions. The treatment of this area is very specific. In fact, intention ontology population is a very delicate task that requires the extraction of unstructured and even a scarce knowledge. Such an ontology is not related to a particular domain but is created to ensure understanding between customer and supplier. Therefore, we have developed an appropriate automatic approach mixing several NLP techniques throughout its various stages detailed in the next section.

3 NLP-Based Ontology Population for Intentional Structure

This work is based on combined syntactic and semantic techniques from natural language processing for populating an intention ontology. It starts with building a structure for the intention defined as an ontology for client intention which describes the client intention model. We consider that an intention is composed of an action, a subject, goals and constraints. Indeed, an action concerns one subject. A subject can be under constraints. Each constraint is composed of a property, an operator and a threshold. The action performed is to achieve 0 or many goals. The proposed ontology structure is presented in Fig. 1. Our objective is to load the information extracted from the client's request as instance in the intention ontology. The outline of the proposed approach is shown in Fig. 2. The main phases of our approach are the following:

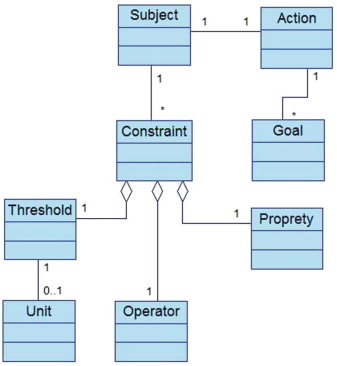


Fig. 1. Intention ontology structure [12]

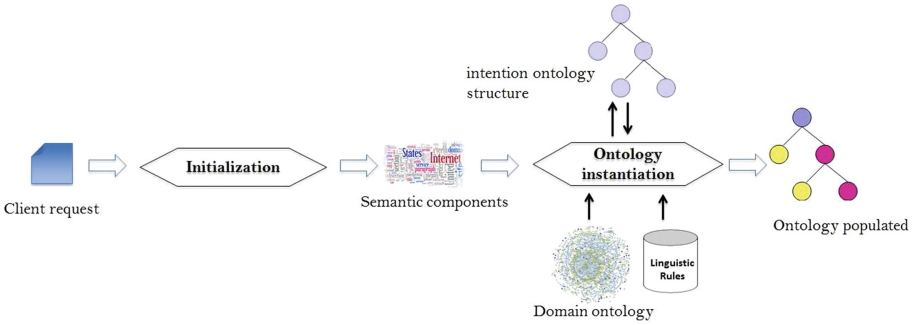


Fig. 2. Outline of the approach

3.1 Initialization Phase

The input of this phase is a client request represented as an informal text. This input is processed with the analyzer *TreeTagger chunker* in order to extract semantic components from the input text. Details about this step are described in our previous work [12].

3.2 Ontology Instantiation Phase

The second phase consists firstly in the classification of each component derived from the first phase in corresponding concepts (Action, Subject, Goal, Property, Operator, Threshold or Unit) and secondly in adding automatically relationships between the concepts' instances. Thereby, we populate the ontology with the new instance of intention detected.

Instances Classification in Concepts. In this second phase, we split each semantic component derived from the first phase into semantically coherent

tokens. Then, we classify each of them in its suitable concept existing in the intention ontology. For that purpose, we have conceived an exhaustive list of linguistic rules that cover all possible cases. Each rule concerns the grammatical category of the token given by the PoS (Part-of-Speech) tag obtained from the initialization phase and its linguistic environment, that is, the series of units that precede and follow the token. For example:

Rule: If the token t is a plural noun (NNS) and preceded by a cardinal number, then t is an instance of the concept UNIT.

More details about these rules are given in our previous work [12]. Therefore, this classification process is based on a matching mechanism between a token and the suitable linguistic rule according to its PoS and its linguistic environment. The task is accomplished by calling a specific function which applies the appropriate rule according to the grammatical characteristics of the token.

Creating Relationships Between Concepts' Instances. We intend, in this step, to complete the population of the ontology. As input, we have instances of concepts. In order to maintain the coherence of the ontology structure, it remains to add some default instances and the relationships between instances of concepts involved in the ontology.

The overall process of creation of associations' instances follows the Algorithm 1. The input of the algorithm is formed by instances composing the intention each assigned to its suitable concept. Basically, the algorithm starts by creating an instance of association between the instance of Action and the Subject instance (line 2). In the case when we don't find an Action instance, we add a default instance "to buy". Then, we create instances of associations between Action instance and Goal instances, as many as there are instances in Goals (line 3). After that, each Unit instance is linked with its Threshold instance. This step is achieved by UnitThresholdLink which retrieves the nearest instance from all the instances to the Unit instance and verifies if it belongs to the Threshold concept. In that case, it creates an association instance between the current unit and the Threshold instance; otherwise, it reiterates with the next nearest instance to the current Unit (lines 4–5).

On the other hand, a client expresses implicitly a number of constraints in his request. To each specific combination of three instances of the concepts Property, Operator and Threshold corresponds an instance of Constraint. So, we should link each created Constraint instance with each instance of its components. An operator is usually positioned, in the input text, between the property and the threshold. So, firstly we browse Operator instances, and for each one, the system creates a Constraint instance. It creates an association instance between this instance and the current Operator instance (lines 7–8). Secondly, the system identifies the corresponding Property and Threshold instances. The system locates the nearest next and the nearest previous instances (if they exist). Then, it verifies if each of those instances corresponds to a Property or a Threshold. In the positive case, it creates an association instance between the Constraint

Algorithm 1. Creating relationships between concepts' instances

Input : Instances _ a set of concepts' instances composing the current intention instance
Output: Intention instance

```
1  begin
2      Create Action-Subject instance link
3      Create Action-Goal instances links
4      for each instance u_i in UnitList do
5          UnitThresholdLink(u_i)
6      for each instance op_i in OperatorList do
7          Create an instance of Constraint cte_i
8          Create an instance link between cte_i and op_i
9          ConstraintThreshPropLink(op_i)
10     for each instance prop_j in PropertyList do
11         Create an instance of Constraint cte_i
12         Create an instance link between cte_i and prop_j
13         ConstraintThresholdLink(prop_j)
14         Create an instance link between cte_i and "equal"
15     for each instance thresh_k in ThresholdList do
16         Create an instance of Constraint cte_i
17         Create an instance link between cte_i and thresh_k
18         Create an instance link between cte_i and "equal"?
19         PropInst = Findhypernym(thresh_k)/* find a superordinate word for
           thresh_k */
20         Create an instance link between cte_i and PropInst
21     for each instance cte_i in ContraintList do
22         SubjectConstraintLink(cte_i)
```

instance and the suitable instance (line 9). At the end, each Constraint instance is linked to the Subject instance (lines 21–22).

As we deal with free texts written by clients, several cases of ambiguities may emerge in the population process. Those cases are specific but frequent. That's why our proposed algorithm deal with them (lines 10–20). For example, before inserting association's instances, we found some ambiguities in some concepts' instances and a lack of information on others. We present in the following paragraph some problems encountered and solutions proposed to clear up those ambiguities:

– **Case 1:** Creating an instance of the concept Constraint must be preceded by the creation of an instance of each of its components (Property, Operator and Threshold). However, in some cases we do not find instances in one or more of those components (e.g. lacks of information in the client request).
Solution: In the case where the Operator is missing (e.g. good screen resolution), the algorithm is extended by a browse of the remaining Property

instances (lines 10–14) (instances haven't relationships after the execution of the basic algorithm) to detect their corresponding Threshold instances. We adopt for that, the same strategy in the algorithm presented to search the nearest Threshold from the remaining instances. Then, we add a default value "equal" as instance in the concept Operator. In a final step, we browse the remaining instances in the Threshold concept that don't have corresponding Operator or Property instances (e.g. android phone), and then we add a default value as Operator and a significant value as an instance for Property (lines 15–20). The latter must be suitable to its corresponding Threshold instance (e.g. for the threshold "android", we add the instance "OS" as property). The Property instance is added by accessing to a specific domain ontology relative to the context of the client's request (OS is an hypernym for android).

- **Case 2:** An instance (especially Operator instance) can't be extracted explicitly.
  Solution: We transform the instance to an adequate one, like the expression "not lower than" which should be considered as an operator. So, we transform it into the symbol "$\geq$" in order to be an appropriate instance for the concept Operator.
- **Case 3:** The use of abbreviations and acronyms by clients in their request expressions to specify some units and brands.
  Solution: For abbreviations used to indicate units, we use a table of units composed of possible abbreviations and the corresponding unit for each one. For others which are used to denote a brand such as "Ip5" written by a client to indicate an "I phone 5", we refer to the domain ontology to solve such a problem.

4 Implementation Details and Results

To validate our approach, we have developed a Java prototype that implements all the presented steps of the approach including the linguistic rules. We used Protégé 3.4.8 for building the intention ontology. OWL (Web Ontology Language) is used to create the ontology which is populated using data from the input client requests. In order to evaluate our methodology, we used a test database extracted from the online forums (e.g. *Tom's Guide*[1]). This base presents a set of examples of clients' requests to express all kinds of questions to purchase hardware. Ten queries that contain different unstructured texts that express needs in different manners were randomly determined as shown in Table 1. To evaluate our experiment results, we adopted the measures of precision and recall. In our experiment, precision represents the number of correct created instances of concepts or relationships between concepts by our tool among the instances created; recall represents the number of correct created instances of concepts or relationships between concepts by our tool among all the essential instances to create an intention instance (detected by an expert). The population process

[1] http://www.tomshardware.fr/forum/.

Table 1. Recall and Precision for the test requests.

Label	Client request	Recall	Precision
R1	I'm looking forward to buy an android tablet for normal uses, with a good battery life and screen resolution not lower than 1280×800. Budget for that is 270 dollars	1	1
R2	I need a new mobile phone before Friday. What's the Mobile phone I can get for 120.00 or less	1	1
R3	I'm looking to buy a laptop. my uses would be for internet use. I'm also looking for one with the Windows 7 rather than the Vista. I would like one with the keyboard backlight. Price around 1000 or lower	1	0.57
R4	Hi, I am looking for an android phone with good speaker sound, big screen and a good camera. I am looking for phones in the average price range	1	1
R5	I have 300 and I want to know what good phone can I get for that kind of money	1	0.84
R6	I am looking for a tablet for 120 or less, and it has to be new. So I'd like to be able to watch my 720p videos with some nice clarity	0.4	0.8
R7	I want to buy a new cell. I want it with a good screen resolution, very good camera quality. Budget is around 25000	1	1
R8	Looking for a new phone and I have 350	1	1
R9	I have 80 dollars for a phone need a touch phone	0.55	1
R10	Looking for a simple mobile phone looking to spend less than 150	0.89	0.94
AVG		0.88	0.91

produces very good results. Overall, the results show that the system reaches a recall of 88% and a precision of 91%.

5 Conclusion

This paper presents an automatic NLP-based ontology population for intentional structure. The objective of this approach is to facilitate the understanding of clients needs. In this context, we extract the intention of client from a natural language text input as a client request. This intention is then inserted as an instance in the intention ontology. The OP approach presented is based on combined NLP techniques and linguistic rules. We propose algorithms that ensure the instanciation of the ontology which presents the model of intention proposed. We gave also some proposals for ambiguity resolution.

In our future work, we plan to use richer formal languages to represent intention. On another side, we believe that it is necessary to have a benchmark for textual intention recognition and this would be an interesting research axis.

References

1. Alani, H., Kim, S., Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H., Shadbolt, N.R.: Automatic ontology-based knowledge extraction from web documents. IEEE Intell. Syst. **18**(1), 14–21 (2003)
2. Amato, F., De Santo, A., Moscato, V., Picariello, A., Serpico, D., Sperlì, G.: A lexicon-grammar based methodology for ontology population for e-health applications. In: 2015 Ninth International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS), pp. 521–526. IEEE (2015)
3. Buitelaar, P., Cimiano, P., Racioppa, S., Siegel, M.: Ontology-based information extraction with soba. In: Proceedings of the International Conference on Language Resources and Evaluation, pp. 2321–2324. Citeseer (2006)
4. Castellanos, M., Hsu, M., Dayal, U., Ghosh, R., Dekhil, M., Ceja, C., Puchi, M., Ruiz, P.: Intention insider: discovering people's intentions in the social channel. In: Proceedings of the 15th International Conference on Extending Database Technology, pp. 614–617. ACM (2012)
5. Chen, Z., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Identifying intention posts in discussion forums. In: HLT-NAACL, pp. 1041–1050 (2013)
6. Ding, X., Liu, B., Zhang, L.: Entity discovery and assignment for opinion mining applications. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1125–1134. ACM (2009)
7. Faria, C., Serra, I., Girardi, R.: A domain-independent process for automatic ontology population from text. Sci. Comput. Program. **95**, 26–43 (2014)
8. Hassan, K., Ali, E., Chantal, S.d., Said, T.: Ontointention: an ontology for documents intentions. In: 2008 Second International Conference on Research Challenges in Information Science, RCIS 2008, pp. 301–306. IEEE (2008)
9. Hu, D.H., Shen, D., Sun, J.-T., Yang, Q., Chen, Z.: Context-aware online commercial intention detection. In: Zhou, Z.-H., Washio, T. (eds.) ACML 2009. LNCS (LNAI), vol. 5828, pp. 135–149. Springer, Heidelberg (2009). doi:10.1007/978-3-642-05224-8_12
10. Kanso, H., Elhore, A., Soule-Dupuy, C., Tazi, S.: Recognition and extraction of intentions based on ontology. In: 3rd International Conference on Information and Communication Technologies: From Theory to Applications, ICTTA 2008, pp. 1–5. IEEE (2008)
11. Kanso, H.: Vers la reconnaissance des intentions de communication: application au contenu de publications scientifiques. Ph.D. thesis, Toulouse (2009)
12. Labidi, N., Chaari, T., Bouaziz, R.: Towards an automatic intention recognition from client request. In: Nguyen, N.-T., Manolopoulos, Y., Iliadis, L., Trawiński, B. (eds.) ICCCI 2016. LNCS (LNAI), vol. 9875, pp. 163–172. Springer, Heidelberg (2016). doi:10.1007/978-3-319-45243-2_15
13. Lee, C.H.L.: Toward intention-aware services provision. In: TENCON 2007–2007 IEEE Region 10 Conference, pp. 1–4. IEEE (2007)
14. Li, X.: Understanding the semantic structure of noun phrase queries. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1337–1345. Association for Computational Linguistics (2010)

15. Makki, J., Alquier, A.M., Prince, V.: An NLP-based ontology population for a risk management generic structure. In: Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology, pp. 350–355. ACM (2008)
16. Navigli, R., Velardi, P.: Enriching a formal ontology with a thesaurus: an application in the cultural heritage domain. In: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge-OLP, pp. 1–9 (2006)

Metaheuristics Parameter Tuning Using Racing and Case-Based Reasoning in Scheduling Systems

Ivo Pereira(✉), Ana Madureira, and Bruno Cunha

GECAD - Knowledge Engineering and Decision Support Research Center, Institute of Engineering – Polytechnic of Porto (ISEP/IPP), Porto, Portugal
{iaspe, amd, bmaca}@isep.ipp.pt

Abstract. Real world optimization problems like Scheduling are generally complex, large scaled, and constrained in nature. Thereby, classical operational research methods are often inadequate to efficiently solve them. Metaheuristics (MH) are used to obtain near-optimal solutions in an efficient way, but have different numerical and/or categorical parameters which make the tuning process a very time-consuming and tedious task. Learning methods can be used to aid with the parameter tuning process. Racing techniques have been used to evaluate, in a refined and efficient way, a set of candidates and discard those that appear to be less promising during the evaluation process. Case-based Reasoning (CBR) aims to solve new problems by using information about solutions to previous similar problems. A novel Racing+CBR approach is proposed and brings together the better of the two techniques. A computational study for the resolution of the scheduling problem is presented, concluding about the effectiveness of the proposed approach.

Keywords: Artificial intelligence · Case-based reasoning · Learning · Metaheuristics · Parameter tuning · Racing · Scheduling

1 Introduction

As defined by Pinedo [1], Scheduling is *"a decision-making process that is used on a regular basis in many manufacturing and services industries"*. Scheduling problems deal with the allocation of resources (e.g. machines) to tasks (or jobs) over given time periods. The goal is to optimize one or more objective functions. In other words, Scheduling problems could be stated as: given a set of jobs (composed by a set of operations), a set of resources, and an optimization criterion for performance measure, the optimal plan is the best way to allocate the resources to the operations considering that precedence relations and the resource availabilities are respected and performance is optimized [2].

The Scheduling problem treated in this work was discussed in [3] and referred as Extended Job-Shop Scheduling Problem (EJSSP). It has some major extensions and differences compared to the classic Job-Shop Scheduling Problem (JSSP), in order to better represent real world problems. JSSP has a set of tasks processing in a set of machines. Each task has an ordered list of operations, each one characterized by the

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_90

respective processing time and machine where is processed. EJSSP additional constraints are: different release and due dates for each job; different priorities for each job; the possibility that not every machines are used for all jobs; a job can have more than one operation being processed in the same machine; two or more operations of the same job can be processed simultaneously; the possibility of existence of alternative machines, identical or not.

Metaheuristics (MH) have gained recognition over the past two decades by successfully solving many real-world problems, including scheduling problems [4]. Indeed, they allow the resolution of large dimension problems by achieving satisfactory solutions in reasonable execution times. These techniques have the objective of guiding and improving the search process to overcome local optimal solutions. They can obtain good quality solutions, very close to the optimal, in reasonable execution times [4].

Metaheuristics parameter tuning has a great influence in the effectiveness and efficiency of the search process. Parameter tuning is known to be a difficult, time consuming and tedious task. There are no "universal" values for the MH parameters. The process of defining the parameters is not obvious because they depend on the problem and the time that the user has to solve the problem [4].

One of the oldest and more important goals in the field of MH is to transfer part of the parameters tuning effort to the algorithm itself. The objective is to provide MH with intelligent mechanisms to be capable of self-adaptation to the problem or situation. Learning techniques can help with this aspect.

Racing is known to be a machine learning approach that has been used to evaluate, in a refined and efficient way, a set of candidates and release those that appear to be less promising during the evaluation process. Racing can be used to effectively and efficiently perform parameter tuning of MH.

Case-based Reasoning (CBR) is a problem-solving paradigm in many aspects different from other Artificial Intelligence approaches. Instead of relying only on general knowledge, CBR is able to use specific knowledge of concrete situations (cases) of a problem. Together with Racing, CBR usage seems very promising.

A hybrid Racing+CBR framework to perform MH parameter tuning is proposed in this paper, which brings together the better of the two learning techniques. A computational study for the resolution of the scheduling problem is carried out in order to validate its influence in the global system's performance. With this work we intend to make a contribution for the resolution of the MH parameter tuning problem, for solving real manufacturing scheduling problems.

The remaining sections of the paper are organized as follows: Sect. 2 describes the proposed Racing+CBR approach, as a novel contribution to the literature; in Sect. 3, the computational results are presented, stating the quality of the proposed hybrid approach; finally, conclusions and future work are presented in Sect. 4.

2 Racing+CBR Learning Approach

In this paper, a hybrid approach for parameter tuning is proposed, joining Racing and CBR together. As illustrated in Fig. 1, Racing is executed before the CBR, where Racing output (tuned parameters) acts as an input for the casebase. After the Racing

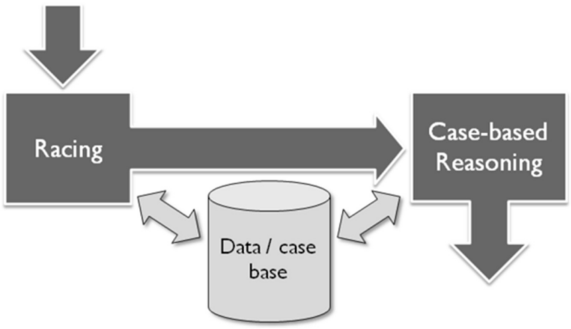


Fig. 1. Racing+CBR architecture [5]

module processing ends, CBR starts its own independent process, using the Racing output parameters to input the casebase. CBR evolves naturally through time and learns with experience. Both modules communicate through a data/casebase.

The idea of joining these two techniques relies on the fact that they can help each other while performing the learning process. Racing approaches perform an intelligent evaluation of learning candidates and gives suitable results, allowing e.g. categorization of candidates and generalization, it lacks on efficiency.

The most common Racing approaches need a lot number of instances to run, which make these learning techniques not very suitable for dynamic systems, as a Manufacturing Scheduling System is. Therefore, is not feasible to run over and over again a Racing algorithm to perform the recommendation of MH parameters each time one needs to optimize the system.

Case-based Reasoning comes into action as a very efficient technique that uses past knowledge to give recommendations. One of the most common problems using CBR is based on the fact that the casebase needs to be populated with good knowledge; otherwise CBR will need many cycles until achieve fairly good results. However, using an initial casebase populated by Racing, it is a very good starting point and the efficiency of the learning system increases.

The main objective of this paper is to show and understand how Racing and CBR can work together based on these premises. In the following subsections, each module is described in detail.

2.1 Racing Module

Racing techniques has been used to evaluate, in a refined and efficient way, a set of candidates and discard those that appear to be less promising during the evaluation process. Instead of training N times the structure of the model, a Racing algorithm can speed up the search for the best structure models, discarding the lower candidates once enough evidence against them are obtained.

The apparent advantage of Racing approaches on brute-force is based on the fact that it provides a better allocation of computational resources among the candidate

configurations. Instead of wasting time to accurately estimate the performance of the lower candidates, Racing approaches focus on the most promising. This aspect allows a more informed selection of the best candidate. In Racing approaches, a candidate is dropped from future assessments so that there is sufficient evidence that it is not the best. Throughout the evaluation process, the approach of Racing focuses increasingly on the most promising candidates. On the other hand, the brute-force approach tests all candidates in the same number of instances.

The algorithm of the Racing module (Table 1) undergoes a study of various combinations of MH parameters – candidates. Thus, as input a list of MH to validate is necessary. The information that follows is based on [5, 6].

Table 1. Racing algorithm

Step	Description
1	Get all the problem instances Get all the candidate parameters
2	Create race to evaluate candidates
3	For each instance still in the race For each candidate still in the race Perform a run to evaluate candidate Remove worst candidates
4	Return best candidate

The first step involves obtaining a list of candidate parameters of each MH to make a race between them. These races are held in a number of instances, which are also obtained in the first step. For each instance, each combination of parameters is tested in the system and the data are stored in the database. At each instance, the candidates who obtained the worst results are removed. At the end of the race, the best candidate is the one who was able to survive through the various instances.

The most important of this whole process is the remove candidates' algorithm, since it decides which candidates will be eliminated from the race. This algorithm takes as input parameters the current race and the list of candidates' parameters, returning the list of survivors' parameters. The first test to perform is check if the list of candidates has more than one combination; otherwise we are looking at the best case. Then we need to check which statistical test to be used. If the candidate list has more than two elements, Friedman's test is applied. Otherwise, Wilcoxon test is used. Both statistical tests are described in [7].

To apply Friedman's test, at least two blocks of results are required, i.e. all candidates have been carried out in at least two instances. This implies that all candidates survive at the end of the first instance. To apply the Friedman's test, one should first obtain the ordered rankings of the candidate parameters and then calculate the sum of each ranking for each candidate.

The number of survivors at each phase will be the best n, calculated using the following equation:

$$sobrev = round\left(\frac{1}{\log_2(inst+1)} \times cand\right) \quad (1)$$

where *cand* is the number of candidates and *inst* is the number of instances.

The number of survivors at each step relies on the number of existing candidates in that step and on the number of instances already executed. In this work, the number of survivors is calculated according to the inverse base-2 logarithm [5]. With this function, it is possible for even a small number of instances, to obtain a quick convergence to the best candidate. This point is an important contribution of this proposal, since the most common racing algorithms are not possible to apply in a small set of instances.

Birattari [6, 8] described that the Wilcoxon matched-pairs signed-ranks can be used when there are only two candidates running, since it has proved more robust and efficient. Following these guidelines, the Wilcoxon test is applied when there are only two candidates.

2.2 Case-Based Reasoning Module

Case-based Reasoning is an Artificial Intelligence methodology aiming to solve new problems by using information about solutions to previous similar problems [9]. On a direct analogy to human learning based on past experience, CBR uses the principle that similar problems may require similar solutions.

Reusing past cases is a powerful method for human and often applied to problem solving. Part of CBR grounds is its psychological plausibility. In CBR, a new problem is solved by searching for a similar past case, and reuse in the new situation. Another important aspect is that CBR used an incremental and sustained learning approach, since a new experience is retained each time a problem is solved, being readily available for future problems [9].

A CBR cycle consists in four main phases [9]:

1. **Retrieve** the most similar case or cases;
2. **Reuse** the retrieved information and knowledge;
3. **Revise** the proposed solution;
4. **Retain** the revised solution for future use.

The database or case repository is designated as casebase. In the first phase, a new case to be solved by CBR is used to retrieve an old case from the casebase. At the second phase, the objective is to retrieve the most similar previous case that was found in the casebase. In the Reusing phase, this retrieved case is used to suggest a solution to solve the new case. This means that, e.g., the solution used to solve this retrieved case can be used to solve the new case, since they are similar between them. In the Revising phase, the suggested solution is tested, e.g., by executing the system, and adapted if necessary. Finally, in the Retaining phase, the information is retained for future use, and the casebase is updated with the new learned case [9].

The objective of this CBR submodule is to recover the most similar case with the new problem, regardless the used MH. Thus, the returned case contains the MH and the respective used parameters. This approach turns out to be suited to the problem at hand

because it is important for the system to decide, at each time, what technique and parameters should be used.

As shown in Table 2, the CBR module consists in retrieving the most similar case or cases to the new problem, regardless the MH to use. Therefore, the case(s) containing the MH and its parameters are retrieved. It is important for the system to decide which MH to use and the respective parameters, because not all MH are suitable to every type of problems. As a consequence of the application of this approach, it will be possible to know e.g. which MH are more suitable to a particular class of problem.

Table 2. Case-based reasoning algorithm

Step	Description
1	Retrieving phase
2	Reusing phase Get retrieved case Get similarity between retrieved and new case
3	Revising phase Parameters tuning revision Running experience in the system
4	Retaining phase

At the start, every new problem leads to a new case for the system. In the first stage, the previous most similar cases are retrieved from the database. After recovering a list of most similar cases, the best case is reused, becoming in a suggested solution. In the Revise stage, the problem is executed using the given solution. To escape from local optimal solutions and stagnation, the use of some disturbance in the parameters of the proposed solution is proposed. In the end, the case are confirmed or not as a good solution, and if so, it is retained on the database as a new learned case, for future use.

In the Retrieving phase, previous cases are analyzed one by one, in order to select those that are similar enough with the new case, by calculating its similarity measure. The attributes to consider in the similarity measure are some characteristics from scheduling problem: number of jobs/tasks, number of machines/resources, problem type (some characteristics can be categorized to better identify a type of problem), number of precedence levels, etc. These attributes will depend exactly on the type of problem to solve and should be differently weighted, as shown in Eq. 2.

$$similarity = \sum_{i=0}^{n} w_i \times sim_i \tag{2}$$

where w_i is the weight of characteristic i and sim_i is the similarity of characteristic i

The different similarities can be calculated depending on the type of attribute. For example, the similarity of number of jobs and number of machines corresponds to the division of the lower value by the higher value, being in the interval [0;1]. The problem type similarity is a Boolean, representing if the attributes are the same or not. In the end of this phase, a list of most similar cases is retrieved.

In the Reusing phase, a case is selected from the list of most similar cases returned by the Retrieving phase, being the respective solution suggested as a candidate solution to solve the new case. So, the MH and respective parameters are returned to be used in the resolution of the new case.

In the Revising phase, the suggested solution is adapted, since the direct use of the solutions leads the system to stagnate and cannot evolve to the best results. So, to avoid local optimal solutions and system's stagnation, an algorithm that applies some diversity and perturbation to the suggested solution's parameters, using a global credit to assign different credit to the different parameters. The global credit is inversely proportional to the similarity of the reused case. This means that the less similar a case is the most perturbation is included in the suggested MH parameters.

Finally, in the Retaining phase, it becomes necessary to store the information in the casebase. First, values of conclusion and execution times are collected, resulting from the execution of the new case. The case is stored and, with this phase, the CBR cycle is concluded. In the next execution, this newly solved case will be available to be used in the resolution of a new case.

3 Computational Study

In this section, the computational study is presented that allows us to conclude about the effectiveness of the proposed Racing+CBR approach, in the resolution of the Scheduling problem. OR-Library[1] instances were used.

To solve the aforementioned Scheduling problem, a Multi-Agent System was proposed, consisting in Task and Resources agents, in which Resource Agents apply different MH, in order to produce effective scheduling plans. Used MH includes Tabu Search, Simulated Annealing, Genetic Algorithms, Ant Colony Optimization, Particle Swarm Optimization, and Artificial Bee Colony. The Multi-Agent System is detailed described in [5, 10] and previous work can be found in [3, 11, 12].

In order to take advantage of the potential of Racing and CBR modules, in this work a hybrid approach (Racing+CBR) is proposed, which employs Racing to initialize the casebase and uses CBR to solve new cases based on past experience. The proposed approach aims to select a technique and perform parameter tuning, depending on the instance to solve.

All obtained results correspond to makespan minimization (completion time - Cmax). Each instance was performed 5 times, and the average of the values obtained was calculated. In order to normalize the values, the ratio between the optimum value and the average value of Cmax is used, to estimate the deviation of the value obtained from the value of the optimal solution referenced in the literature.

In this hybrid approach, 150 initial cases were included, with values corresponding to the parameters in instances throughout the study Racing (6 results in 25 MH instances). Using the initial case base filled with the results obtained from Racing, it was carried out after the development of the study of the use CBR, running the system

[1] http://people.brunel.ac.uk/~mastjjb/jeb/info.html.

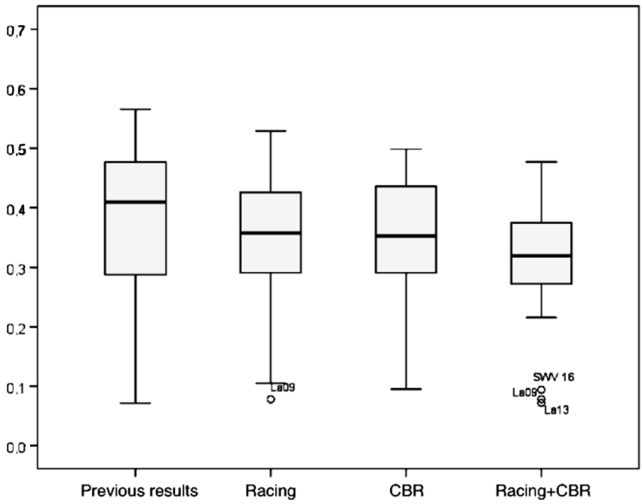


Fig. 2. Average values ratio comparison between initial (without learning), Racing, CBR and Racing+CBR results

in test instances. Table 3 shows the analysis of the ratio of the mean Cmax with all learning all approaches, including the Racing+CBR. From Fig. 2 chart analysis it is possible conclude on the advantage of the hybrid approach analyzing median and dispersion of data.

Table 3. Analysis of average values ration from initial (without learning), Racing, CBR and Racing+CBR results

	Min. value	Max. value	Avg.	Std. dev.
Initial	0.07	0.57	0.3792	0.13217
Racing	0.08	0.53	0.3474	0.11834
CBR	0.09	0.50	0.3388	0.10736
Racing+CBR	0.07	0.48	0.3142	0.10999

Comparing the results of the hybrid approach with the results of Racing module, we can see a slight improvement of the results, especially in the dispersion, which indicates that the inclusion of CBR module can maintain greater consistency of results. As shown in Table 3, there is an average and a standard deviation lower than in the Racing module, and moreover a minimum value and a maximum value below which is important since this is a minimization problem.

Comparing the hybrid approach with the results of CBR module, the improvement is not as significant, which expects to join the two modules can be advantageous. In Table 3, improvements are noted in minimum and maximum value, and the average, but the standard deviation is slightly higher, although the difference is minor.

The results of the Racing+CBR approach were compared with the initial results. And here, clear improvements are shown, which indicates a statistically significant advantage in the use of learning algorithms in MH parameter tuning for optimization problems. In Table 3 one can note significant improvements, especially in the middle of the results and the standard deviation. The minimum value was equal but the maximum value is significantly lower. Finally, comparing the first two approaches with the hybrid approach, you can check on the advantage of use of the latter.

At this point it becomes important to analyze the statistical significance of these results. Comparing the previous results and the results of the hybrid approach (Table 4) the values $t(29) = 5.475$, $p < 0.05$ allows to conclude, with a degree of confidence of 95%, that there are statistically significant differences between the initially obtained results (without training) and the results obtained by the Racing+CBR approach. Observing this, we can conclude about the advantage of using the Racing+CBR approach. In the other hand, comparing the results of the Racing+CBR approach with the results of Racing, and observing the values $t(29) = 5.068$, $p < 0.05$, one can also say with a confidence level of 95%, there are statistically significant differences between the two approaches, with advantages for the hybrid approach. Finally, comparing the results of CBR with the results of the hybrid approach, you can also complete, with a confidence level of 95%, there are statistically significant differences between the two approaches, as $t(29) = 2.581$; $p < 0.05$.

Table 4. Student's *t* test for paired samples

	Avg.	Std. dev.	*t*	DoF	p-value
Initial vs. Racing+CBR	0.06496	0.06499	5.475	29	0.000
Racing vs. Racing+CBR	0.03314	0.03582	5.068	29	0.000
CBR vs. Racing+CBR	0.02454	0.05208	2.581	29	0.015

4 Conclusions and Future Work

The objective of this paper was to present a novel approach for the selection and tuning of MH, in the resolution of Scheduling problems. The proposed hybrid approach is based on Racing and CBR. While Racing is used to evaluate a set of candidates and release those that appear to be less promising during the evaluation process, in a refined and efficient way, CBR which uses previous similar cases to solve new cases, providing a learning from experience. Considering the importance of tuning in the process of designing and implementing MH, a computational study was performed to evaluate the effectiveness of the proposed Racing+CBR approach.

The presented computational study aimed to evaluate the performance of a Multi-Agent Scheduling System, with the incorporation of different learning modules. The results of the hybrid Racing+CBR approach were compared with the results obtained previously, without the incorporation of learning mechanisms. All results were validated by analyzing the statistical significance and a significant statistically advantage in the use of Racing+CBR learning module was confirmed. All the presented approaches have improved the results obtained the system in relation to the previous

results, however the hybrid Racing+CBR approach achieved better average results. From the results, it was possible to conclude about the existence of statistical evidence on the inclusion of learning on the process of MH tuning.

In the future, we intend to make a more extensive computational study and also compare our novel contribution with other learning techniques, as such Reinforcement Learning, for instance.

Acknowledgments. This work is supported by FEDER Funds through the "Programa Operacional Factores de Competitividade - COMPETE" program and by National Funds through FCT "Fundação para a Ciência e a Tecnologia" under the project: FCOMP-01-0124-FEDER-PEst-OE/EEI/UI0760/2014.

References

1. Pinedo, M.: Scheduling: Theory, Algorithms, and Systems. Springer, Heidelberg (2016). ISBN 978-3-319-26580-3
2. Baker, K.R., Trietsch, D.: Principles of Sequencing and Scheduling. Wiley, New York (2009)
3. Madureira, A.: Meta-heuristics application to scheduling in dynamic environments of discrete manufacturing, Ph.D. Dissertation, University of Minho (2003). (in portuguese)
4. Talbi, E.-G.: Metaheuristics - From Design to Implementation. Wiley, New York (2009)
5. Pereira, I.: Intelligent system for scheduling assisted by learning, Ph.D. thesis in Electrical and Computer Engineering, Department of Electrical and Computer Engineering, University of Trás-os-Montes and Alto Douro (2014). (in portuguese)
6. Birattari, M., Balaprakash, P., Dorigo, M.: The ACO/F-RACE algorithm for combinatorial optimization under uncertainty. In: Doerner, K.F., Gendreau, M., Greistorfer, P., Gutjahr, W., Hartl, R.F., Reimann, M. (eds.) Metaheuristics. Operations Research/Computer Science Interfaces Series, vol. 39, pp. 189–203. Springer, Heidelberg (2007)
7. Box, G., Hunter, J., Hunter, W.: Statistics for Experimenters: Design, Innovation, and Discovery. Wiley, New York (2005)
8. Birattari, M., Stutzle, T., Paquete, L., Varrentrapp, K.: A racing algorithm for configuring metaheuristics. In: Langdon, W.B., et al. (eds.) Genetic and Evolutionary Computation Conference, pp. 11–18. Morgan Kaufmann Publishers Inc., San Francisco (2002)
9. Kolodner, J.: Case-Based Reasoning. Morgan Kaufmann, San Francisco (2014)
10. Madureira, A., Santos, J., Pereira, I.: A hybrid intelligent system for distributed dynamic scheduling. In: Chiong, R., Dhakal, S. (eds.) Natural Intelligence for Scheduling, Planning and Packing Problems, Studies in Computational Intelligence. SCI, vol. 250, pp. 295–324. Springer, Heidelberg (2009)
11. Madureira, A., Santos, J., Pereira, I.: MASDSheGATS – scheduling system for dynamic manufacturing environments. In: MultiAgent Systems. In-Tech (2009)
12. Pereira, I., Madureira, A.: Self-Optimization module for Scheduling using Case-based Reasoning. Appl. Soft Comput. **13**(3), 1419–1432 (2013). Elsevier

A Branch-and-Price Algorithm for the Double Vehicle Routing Problem with Multiple Stacks and Heterogeneous Demand

Jonatas B.C. Chagas(✉) and André G. Santos(✉)

Departamento de Informática, Universidade Federal de Viçosa,
Viçosa, Minas Gerais, Brazil
jonatas.chagas@ufv.br, andre@dpi.ufv.br

Abstract. This paper introduces the Double Vehicle Routing Problem with Multiple Stacks and Heterogeneous Demand (DVRPMSHD). This problem is a more realistic generalization of the Double Vehicle Routing Problem with Multiple Stacks when customers have heterogeneous demands and the demand of each customer cannot be divided among two or more vehicles. The objective of this problem is to find optimal routes for a fleet of vehicles to visit all pickup and delivery locations while ensuring the feasibility of the loading plan. An integer linear programming and a branch-and-price algorithm are proposed. The computational results show that our branch-and-price algorithm is able to find the optimal solution for most of the instances reported and find better solutions than those found by the integer linear programming formulation for some instances.

1 Introduction

Given the increasing competition among companies that seek to achieve consolidation and stability, scientific/technological advances and strategic business planning are crucial so that they become more competitive and able to conquer the consumer market. In this way, the companies seek to steadily improve the quality of their logistics processes.

The transportation is one of the most important logistics processes, being fundamental in providing services to customers [1]. The transport planning has been gaining importance in the current scenario and arousing attention for researchs that seek to adapt logistic processes to current needs, since the costs associated with transport has a high factor in logistic costs in most companies [2].

Due to economic importance, one of the most studied problems in combinatorial optimization is the Vehicle Routing Problem (VRP). Since it was first proposed in 1959 by Dantzig and Ramser [3], several studies have been done with the aim to propose new methods for solving the variations of this problem. According to Desrochers et al. [4] the variants of VRP's differ in aspects related

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_91

to the type of vehicle fleet, location and customer demand, existence of temporal and spatial constraints, type of loading, objective of the problem and several other characteristics related to a specific scenario.

Among the VRP's variants, the one which considers load constraints, known as Routing Problems with Loading Constraints, has become very attractive because these problems are beyond the theoretical field, being viewed in the real-world. Iori and Martello [5] made an study about the problems that combine routing and loading constraints, reporting its characteristics, importance and applications.

In 2009, Petersen and Madsen [6] formulated the Double Travelling Salesman Problem with Multiple Stacks (DTSPMS) motivated by a real world application arising in transportation of pallets between a pickup region and a delivery region. The DTSPMS arises in a context where the pickup and delivery regions are widely separated. In this problem, a single vehicle must load all products in the pickup region before unload them in the delivery region. All products have the same size and shape, and the loading compartment of the vehicle (container) is divided into stacks of fixed height. The product are collected in the pickup region and stored in stacks. The products cannot be rearranged in container, i.e., once that a particular product is stored in some stack it should remain unmoved up to the unloading. In the delivery region the products must be unloaded and delivered respecting a LIFO policy, i.e., at any time, the customers served must be the ones with items on top of a stack, according to the loading in the pickup region. The cost of transport between the two regions is fixed and is not considered as part of the optimization problem. The objective is to find two tours, one for each region, of minimum total cost while ensuring the feasibility of the loading plan.

Since Petersen and Madsen [6] proposed a mathematical formulation and a heuristic approach based on the simulated annealing metaheuristic to solve the DTSPMS, a great interest on the problem aroused in the academic community, emerging several approaches to solve it. Felipe et al. [7] proposed a neighborhood structure and a variable neighborhool serch algorithm. Carrabs et al. [8] proposed a branch-and-bound algorithm for a special case of the problem, when the vehicle has exactly two stacks. Alba Martínez et al. [9] proposed an exact algorithm based on a branch-and-cut framework.

In 2015, Iori and Riera-Ledesma [10] formulated a generalization for the DTSPMS motivated by the fact that not always a single vehicle is able to meet the demand of products, thus more than one is needed. Similarly as in the DTSPMS, the goal is to collect the products in the pickup region and deliver them in their respective places in the delivery region, respecting the LIFO policy on the containers. However, the difference is that now there is a fleet of vehicles available to performe the needs. The vehicle fleet is fixed, i.e., the number of vehicles used is already defined. Also, the fleet of vehicles may be heterogeneous, i.e., the vehicle loading compartment may be different.

Regarding the proposed approaches to solve the DVRPMS, Iori and Riera-Ledesma [10] proposed three exact algorithms to solve the DVRPMS: a branch-and-cut, a branch-and-price and a branch-and-cut-and-price. Silveira et al. [11]

proposed three heuristic algorithms based on simulated annealing, iterated local search and variable neighboord descent.

In this paper we consider a more realistic generalization of the DVRPMS, which considers that customers have heterogeneous demand and all demand of each customer should be served by a single vehicle. We call this generalization the Double Vehicle Routing Problem with Multiple Stacks and Heterogeneous Demand (DVRPMSHD).

Figure 1 represents a viable solution example for an instance of the DVRPMSHD involving 12 requests and 3 vehicles with containers of size (2×4), (2×3) and (2×3). The demand value of each vertex is indicated by the number beside of each one. For each vehicle are associated a tour in the pickup region, a tour in the delivery region and a loading plan of the container. The tours of each vehicle respect the LIFO policy on the container. Notice that the demand of a customer may be split in different stacks, as long as they belong to the same vehicle.

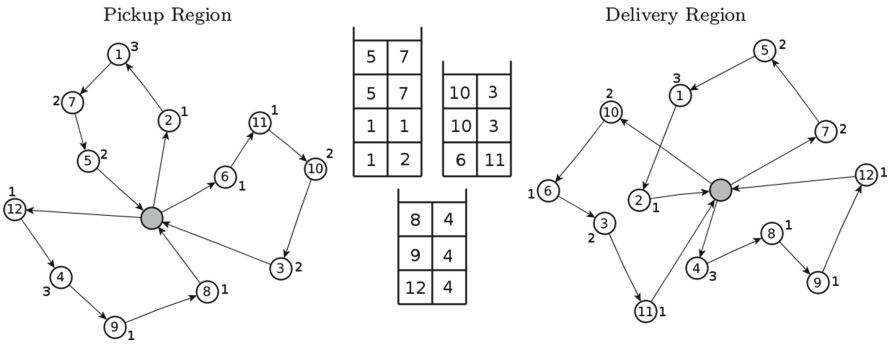


Fig. 1. An example of a feasible solution for the DVRPMSHD

The remaining of this paper is organized as follows. In Sect. 2 the DVRPMSHD is formally defined. In Sect. 3 we present an Integer Linear Programming (ILP) formulation for the problem. Section 4 presents the proposed branch-and-price algorithm. Finally, the computational experiments are reported in Sect. 5 and in Sect. 6 we present our conclusions.

2 Problem Definition

The DVRPMSHD can be formally described as follows. Let $I = \{1, 2, ..., n\}$ be the set of n customer requests, and let also $V_c^P = \{1^P, 2^P, ..., n^P\}$ be the set of vertices that represent the locations of the pickup customers and $V_c^D = \{1^D, 2^D, ..., n^D\}$ representing the locations of the delivery customers. Each request $i \in I$ is associated with a pickup vertex i^P, a delivery vertex i^D and a

value d_i representing the request demand, i.e., the number of products that has to be loaded at the vertex i^P, transported and unloaded at the vertex i^D.

The DVRPMSHD is defined on two complete directed graphs, $G^P = (V^P,\ E^P)$ and $G^D = (V^D,\ E^D)$, where $V^P = V_c^P \cup \{0^P\}$ and $V^D = V_c^D \cup \{0^D\}$ represents, respectively, the vertices in the pickup and delivery regions, being 0^P and 0^D the depots on each region. The set of edges in pickup and delivery regions are defined by $E^P = \{(i^P, j^P, c_{ij}^P)\ \forall i^P \in V^P, \forall j^P \in V^P \mid j^P \neq i^P\}$ and $E^D = \{(i^D, j^D, c_{ij}^D)\ \forall i^D \in V^D, \forall j^D \in V^D \mid j^D \neq i^D\}$, where c_{ij}^P and c_{ij}^D are, respectively, the routing costs associated to edges (i^P, j^P) and (i^D, j^D).

The vehicles of set K depart always from the pickup and delivery depots. The loading compartment of each vehicle $k \in K$ is divided into R_k stacks of height L_k. The vehicles must collect all products from customers located in the pickup region, storing them in their containers and then deliver them in the delivery region. The routes of each vehicle must satisfy the LIFO policy in all stacks of all vehicles, i.e., if a pickup vertex i^P is visited before the pickup vertex j^P, and some product of the request i and some product of the request j are loadead into the same stack, then the delivery vertex j^D must be visited before the delivery vertex i^D.

In order to simplify the notation, we use $T = \{P,\ D\}$ to define the properties applied in both graphs. We also adopted the notation described by Iori and Riera-Ledesma in [10] using i to denote both i^P and i^D, and (i, j) to denote both (i^P, j^P) and (i^D, j^D).

3 ILP Formulation

Our formulation for the DVRPMSHD is based on the ILP formulations for the DTSPMS defined by Petersen and Madsen [6] and for the DVRPMS proposed by Iori and Riera-Ledesma [10]. The decision variables used are:

- $x_{i,j}^{kT}$: binary variable that gets 1 if vehicle k crosses the arc (i, j) and 0 otherwise. T indicates that the properties of the variable are valid for both regions, P or D.
- y_{ij}^{T}: binary variable that gets 1 if vertex i is visited before vertex j in region T and 0 otherwise.
- w_i^{kT}: binary variable that gets 1 if vehicle k visits the customer in i of region T and 0 otherwise.
- z_{ih}^{rk}: binary variable that gets 1 if the h-th item of customer i is placed in stack r of the vehicle k and 0 otherwise.

Using these variables we can obtain the following ILP formulation:

$$\min \sum_{k \in K} \sum_{T \in \{P,D\}} \sum_{(i,j) \in E^T} c_{ij}^T x_{ij}^{kT} \tag{1}$$

$$\sum_{j \in V_c^T} x_{0j}^{kT} = 1 \qquad \forall k \in K, \forall\, T \in \{P, D\} \qquad (2)$$

$$\sum_{i \in V_c^T} x_{i0}^{kT} = 1 \qquad \forall k \in K, \forall\, T \in \{P, D\} \qquad (3)$$

$$\sum_{i \in V^T \setminus \{j\}} x_{ij}^{kT} = w_j^{kT} \qquad \forall k \in K, \forall\, T \in \{P, D\}, \forall j \in V_c^T \qquad (4)$$

$$\sum_{j \in V^T \setminus \{i\}} x_{ij}^{kT} = \sum_{j \in V^T \setminus \{i\}} x_{ji}^{kT} \qquad \forall k \in K, \forall\, T \in \{P, D\}, \forall i \in V_c^T \qquad (5)$$

$$\sum_{k \in K} w_i^{kT} = 1 \qquad \forall\, T \in \{P, D\}, \forall i \in V_c^T \qquad (6)$$

$$\sum_{i \in I} \sum_{h=1}^{d_i} z_{ih}^{rk} = L_k \qquad \forall k \in K, \forall r = 1..R_k \qquad (7)$$

$$\sum_{h=1}^{d_i} \sum_{r=1}^{R_k} z_{ih}^{rk} = d_i w_i^{kP} \qquad \forall k \in K, \forall i \in I \qquad (8)$$

$$w_i^{kP} = w_i^{kD} \qquad \forall k \in K, \forall i \in I \qquad (9)$$

$$y_{ij}^T + y_{ji}^T = 1 \qquad \forall\, T \in \{P, D\}, \forall i \in V_c^T, \forall j \in V_c^T \setminus \{i\} \qquad (10)$$

$$y_{il}^T + y_{lj}^T \leq y_{ij}^T + 1 \qquad \forall\, T \in \{P, D\}, \forall l \in V_c^T, \forall i \in V_c^T, \forall j \in V_c^T \setminus \{i\} \qquad (11)$$

$$x_{ij}^{kT} \leq y_{ij}^T \qquad \forall k \in K, \forall\, T \in \{P, D\}, \forall i \in V^T, \forall j \in V^T \setminus \{i\} \qquad (12)$$

$$y_{ij}^P + z_{ih}^{rk} + z_{jh'}^{rk} \leq 3 - y_{ij}^D \qquad \begin{array}{r} \forall k \in K, \forall i \in I, \forall j \in I \setminus \{i\}, \\ \forall h = 1..d_i, \forall h' = 1..d_j, \forall r = 1..R_k \end{array} \qquad (13)$$

$$\sum_{j \in I} j x_{0j}^{kP} \leq \sum_{j \in I} j x_{0j}^{k'P} \qquad \forall k \in K, \forall k' \in K \mid k < k', R_k = R_{k'}, L_k = L_{k'} \qquad (14)$$

$$x_{ij}^{kT} \in \{0, 1\} \qquad \forall k \in K, \forall\, T \in \{P, D\}, \forall (i, j) \in E^T \qquad (15)$$

$$y_{ij}^T \in \{0, 1\} \qquad \forall\, T \in \{P, D\}, \forall i \in V^T, \forall j \in V^T \setminus \{i\} \qquad (16)$$

$$z_{ih}^{rk} \in \{0, 1\} \qquad \forall k \in K, \forall i \in I, \forall h = 1..d_i, \forall r = 1..R_k \qquad (17)$$

$$w_i^{kT} \in \{0, 1\} \qquad \forall k \in K, \forall\, T \in \{P, D\}, \forall i \in V_c^T \qquad (18)$$

The objective function (1) minimizes the total routing cost of the vehicles. The constraints (2) and (3) force each vehicle to leave and arrive in the depot. The constraint (4) ensures that a edge (i, j) is crossed by a vehicle k only if the request j is served by vehicle k. The constraint (5) ensures that the same vehicle must arrive and leave of one vertex. The constraint (6) guarantees that each customer is attended by just one vehicle. The constraint (7) ensures that the capacity of the stacks is not extrapolated. The constraint (8) forces that all demand of a customer must be carried by a single vehicle. The constraint (9) ensures that the same vehicle must attend the same customer in the pickup region and in the delivery region. The constraints (10) and (11) respectively

impose an order of visit of customers on each region and ensures a transitivity on this order. The constraint (12) ensures that if the edge (i, j) is used by some vehicle, then i is visited before j. The constraint (13) expresses the LIFO constraints: if products of customers i and j are loaded in the same stack r of vehicle k, with i visited before j in the pickup region, then i cannot be visited before j in the delivery region. The constraint (14) breaks the symmetry arising from the formulation imposing a lexicographical order on the routes of vehicles with the same container configuration. And, finally, the constraints (15) to (18) define the scope and domain of the decision variables.

This formulation completely describes the problem and may used by an ILP solver to solve a given instance. An alternative formulation, more suitable to be solved by a column generation technique, is given in the next section.

4 Branch-and-Price Algorithm

In this section we describe our branch-and-price algorithm, detailing the set partitioning formulation, the components of the column generation (pricing) and the criteria used in the branching process.

4.1 Set Partitioning Formulation

Let $\mathscr{S}_k$ be the set of all feasible loadings for vehicle $k \in K$. In order to be feasible, a loading must serve a subset of requests, include all demand of the requests served, do not exceed the vehicle's capacity, and the products be positioned in a way that there is a sequence of customer's visit that respects the LIFO policy. For a given $s \in \mathscr{S}_k$, let c_k^s be the minimal total routing cost to load and unload the items considering the loading pattern s, and let p_{ki}^s describe the requests served, being 1 if the request i is attended by the loading pattern s and 0 otherwise.

In order to formulate the DVRPMSHD using a set partitioning formulation we define the decision binary variable λ_k^s that assumes 1 if the requests of the loading pattern s are attended by the vehicle k, and 0 otherwise. Then, DVRPMSHD can be formulated by:

$$\min \sum_{k \in K} \sum_{s \in \mathscr{S}_k} c_k^s \lambda_k^s \tag{19}$$

$$\sum_{k \in K} \sum_{s \in \mathscr{S}_k} p_{ki}^s \lambda_k^s = 1 \qquad \forall i \in I \tag{20}$$

$$\sum_{s \in \mathscr{S}_k} \lambda_k^s = 1 \qquad \forall k \in K \tag{21}$$

$$\lambda_k^s \in \{0, 1\} \qquad \forall k \in K, \forall s \in \mathscr{S}_k \tag{22}$$

By the constraints (21), exactly one loading pattern is chosen for each vehicle. The chosen patterns must serve all requests (constraint (20)) and have minimal total cost (objective function (19)).

Although simple to describe, this formulation has an implicit complexity, which is the necessity to generate (and work with) the set of all feasible loading patterns for each vehicle configuration. The size of these sets may be quite huge. Fortunately the sets may be generated on-the-fly using a column generation technique, which is described in the following. However, this technique requires the integrality constraint (22) to be relaxed, forcing the process to be embedded in a branch-and-bound approach, leading to a branch-and-price algorithm, which will be detailed later.

4.2 Column Generation

As pointed before, the model (19)–(22) may be difficult to solve due to the huge number of variables. In fact, this number grows exponentially with the number of customers and stacks on the vehicles. In this formulation, a variable (from now on called column) represents a feasible pattern (routes and loading plan). In a column generation approach, the model is decomposed into a master problem, that is a linear version of the original problem containing only a restricted number of columns, and in a subproblem, that iteratively generates promising columns. In order to iteratively improve the current solution, the master problem sends the dual prices of its constraints to the subproblem, which in turn uses these dual prices to evaluate and generate patterns whose reduced cost improves the current master solution by the simplex algorithm. The process continues until the subproblem is not able to generate improve columns. If the subproblem is solved to optimality, the last solution of the master problem will be (linear) optimal. The master problem and the subproblem of our approach are detailed in the following. We suggest the book [12] for the readers interested in the theory behind the column generation method.

Master Problem. The master problem is a linear version of the set partitioning based formulation, i.e., objective function (19), constraints (20), (21) and the following (23) instead of (22). Moreover, instead of using the full set $\mathscr{S}_k$ of feasible columns for each vehicle, the model uses a subset $\widehat{\mathscr{S}_k} \subseteq \mathscr{S}_k$ of the columns, which is filled iteratively by the subproblem. Every time the master problem is solved, besides a linear solution, the simplex method provides dual prices associated to its constraints. We denote by π_i the dual price for request i and by α_k the dual price for vehicle k, respectively from constraints (20) and (21). These values are passed to the subproblem in order to obtain new promising columns.

$$0 \leq \lambda_k^s \leq 1 \qquad \forall k \in K, \forall s \in \mathscr{S}_k \tag{23}$$

Subproblem. Guided by the dual prices provided by the master problem, the subproblem searches a column of minimal reduced cost, or at least a column with negative reduced cost. A column, in our context, is a loading pattern, including the routing cost. The reduced cost is the routing cost discounted the

dual prices, which varies according to the requests attended and the vehicle used. The objective is then expressed by (24). Instead of only one subproblem, we work with $|K|$ subproblems, one for each vehicle. This is possible because the only dependence between the different vehicles is that each request must be attended by one and only one vehicle, a constraint already covered in the master problem. The ILP formulation for the subproblem is then very similar to the model (1)–(18) with the following differences: the index k is dropped in all variables as it deals with one vehicle a time; constraints (6), (9) and (14) are removed; and constraint (10) is substituted by (25), because not all requests are to be attended by a single vehicle.

$$\min \sum_{T \in \{P,D\}} \sum_{(i,j) \in E^T} c_{ij}^T x_{ij}^T - \sum_{i \in I} \pi_i w_i - \alpha_k \tag{24}$$

$$y_{ij}^T + y_{ji}^T = w_i \qquad \forall\, T \in \{P,D\}, \forall i \in V_c^T, \forall j \in V_c^T \setminus \{i\} \tag{25}$$

4.3 Branching

If the column generation process runs until no column with negative reduced cost exists, the final solution of the master problem is optimal. It may not be integral, however, as the model solved is a linear relaxation of the set partitioning model. In this case, we start a branch-and-bound (B&B) enumeration to achieve the integer optimal solution. The difference to a standard B&B is that columns, i.e., new variables, may be generated on each node of the enumeration tree. The whole process is then called branch-and-price [12].

In every node of the B&B tree, if a non-integral solution is found, a request is chosen to be fixed (and later forbidden) in a certain vehicle. Similarly as was demonstrated by Savelsbergh in [13] for the generalized assignment problem, if $\lambda_k^{s'}$ is fractional for any $s' \in \mathscr{S}_k$, then some request is only partially served by the vehicle k, i.e., there is a request i covered by s' such that $\sum_{s' \in \mathscr{S}_k} p_{ki}^{s'} \lambda_k^{s'} < 1$. Among such requests not integraly served by a vehicle we chose the one that has close value to zero. In one side of the tree, the vehicle is forbid to serve this request, by fixing $w_i = 0$ in the subproblem of vehicle k. In the other side, it is forced to be served by this vehicle, by fixing $w_i = 1$ in the subproblem of vehicle k and $w_i = 0$ in other subproblems. Additional constraints on values of the variables are also added on the master problem accordingly. Integer solutions found during the process updates the best solution, which are kept as an upper bound to the final solution.

Due to time constraints, the process may not run until the end. We would have a subset of columns that not necessarily contains the ones of the optimal solution. In this case, we simply run the master problem again, this time imposing integrality on all variables. This may lead to a feasible solution, although not optimal.

5 Computational Experiments

Our implementations were coded in C/C++ using the Concert Technology Library of CPLEX 12.5 academic version, with CPLEX default settings, except for the runtime that was limited to 1 h. All experiments were performed on an Intel Core i7-4790K CPU @ 4.00 GHz x 8 desktop computer with 32 GB RAM, Ubuntu 14.04 LTS 64 bits.

5.1 Instances Description

In order to validate our approaches, the computational experiments were performed using 90 instances built from the DTSPMS instances proposed by Petersen and Madsen [6], where the items are randomly arranged in two regions (pickup and delivery) of dimension 100×100, being the depot located at (50, 50). Euclidean distances were rounded to the nearest integer and the sets of items were named R00 to R04. In some of these instances, not all items are used. Therefore, the items were chosen in ascending order starting from the first item in the list.

Based on Iori and Riera-Ledesma [10], we defined 18 types of fleet configurations, whose characteristics are show in Table 1. Each type is descrived by the

Table 1. Configuration of the containers for each type of fleet

$Type$	$\|K\|$	$(R \times L)'s$	$\sum_{k \in K} R_k L_k$
(a)	2	(2×3) (2×3)	12
(b)	2	(2×2) (2×4)	12
(c)	3	(2×2) (2×2) (2×2)	12
(d)	2	(2×3) (3×4)	18
(e)	3	(2×3) (2×3) (2×3)	18
(f)	4	(2×2) (2×2) (2×2) (2×3)	18
(g)	2	(3×4) (3×4)	24
(h)	3	(2×4) (2×4) (2×4)	24
(i)	4	(2×3) (2×3) (2×3) (2×3)	24
(j)	2	(3×5) (3×5)	30
(k)	3	(2×5) (2×5) (2×5)	30
(l)	4	(2×3) (2×4) (2×4) (2×4)	30
(m)	2	(3×6) (3×6)	36
(n)	3	(3×4) (3×4) (3×4)	36
(o)	4	(2×4) (2×4) (2×4) (3×4)	36
(p)	2	(3×8) (3×8)	48
(q)	3	(3×5) (3×5) (3×6)	48
(r)	4	(3×4) (3×4) (3×4) (3×4)	48

Table 2. Uniform distribution of the demand values for each number of customers

Type	# customers	Demands																							∑ demands
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
(i)	6	3	1	3	1	2	2																			12
(ii)	9	3	1	3	2	2	2	1	1	3																18
(iii)	12	3	1	3	2	3	2	1	1	3	2	1	2													24
(iv)	15	2	1	3	2	3	2	1	1	3	2	1	2	1	3	3										30
(v)	18	2	1	3	2	3	2	3	1	3	2	1	1	1	3	3	2	1	2							36
(vi)	24	2	1	3	3	1	2	3	1	3	2	1	2	1	3	3	2	1	2	2	2	1	3	1	3	48

number of vehicles $|K|$, the configuration of the containers of the vehicle fleet and the maximum capacity of the freight.

In order to define the demand values for each customer, we used an integer-valued uniform distribution between 1 and 3. Table 2 shows the values generated considering 6, 9, 12, 15, 18 and 24 customers.

5.2 Experimental Results

Here we report the results obtained by our approaches for the DVRPMSHD. Notice that since there is no literature regarding this problem, we are only going to compare the results of our approaches.

In Table 3 the four first columns describe the instances, being that each instance is identified by a number (column ID), a name indicating the pickup and delivery regions (column R), a type of vehicle feet (column F) (see Table 1) and a demand distribution (column D) (see Table 2). The results obtained by the ILP formulation are described in the following columns. The column UB shows the best solution value obtained at the end of the computation, LB_0 informs the value of the lower bound computed relaxing the ILP formulation to an LP formulation, $\%g_0$ shows the relative gap between UB and LB_0 computed as $(UB - LB_0)/UB$, %g shows the relative gap between UB and LB computed as $(UB - LB)/UB$, where LB is the lower bound value at the end of the computation. The column Opt indicates by an asterisk the instances solved to optimality and the column t(s) informs the total processing time. The following columns describe the results obtained for the branch-and-price algorithm. The column LB_0 informs the value of the lower bound computed in the root node, the columns #N and #C show, respectively, the number of nodes expanded and the number of columns added by the algorithm. The other columns have the same meaning of the columns for the ILP formulation.

Notice that the best results are boldfaced, and each empty cell means that the information has not been obtained with the processing time limited to 1 h.

For all instances which the branch-and-price algorithm was not able to find the optimal solution, the following procedure was performed. All variables λ_k^s included in the master problem (columns) were defined to be binary ($\lambda_k^s \in \{0,1\} \quad \forall k \in K, \forall s \in \widehat{\mathscr{S}_k}$), and the master problem is solved again. The goal of

this process is to find an integer solution using the columns already inserted in the master problem.

The results show that, within a 1 h time timit, our ILP formulation and the branch-and-price algorithm were able to solve 52 and 66 instances, respectively, of the 90 instances. It is also noticed that for 80% of the instances the branch-and-price algorithm found equal or better solutions compared to the solution obtained by ILP formulation.

Note that, for all instances, when the root node was fully processed by both methods, the initial gap value (column $\%g_0$) of the branch-and-price algorithm is zero or close to zero, whilst the gaps for the ILP formulation are very high, indicating that the model based on set partitioning formulation is stronger, as it has a better linear relaxation value. Even though, for the smaller instances the initial gap of the ILP formulation could be reduced to zero (column $\%g$), proving optimality.

We can also observe that for several instances the branch-and-price algorithm consumes a lot of processing time, not being able to process completely the root node (then, the linear optimal LB_0 is not found). Even when this occurs, the columns already generated may be enough to compose a good integer solution. For example, for instances with ID 64, 76, 78 and 79, the UB value are better than the ones found using the ILP formulation.

Regarding the computational time, both methods were limited to one hour. Analysing the average computation time of all 90 instances, the branch-and-price algorithm was more efficient, being 1.5 times faster than the ILP formulation.

Table 3. Comparison between ILP formulation and branch-and-price algorithm

Instance				ILP formulation						Branch-and-Price						
ID	R	F	D	UB	LB_0	$\%g_0$	$\%g$	Opt	t(s)	UB	LB_0	$\%g_0$	#N	#C	Opt	t(s)
1	R00	(a)	(i)	**555**	478.0	13.87	0.00	*	0.1	**555**	555.0	0.00	1	14	*	0.1
2	R01			**636**	419.0	34.12	0.00	*	0.2	**636**	636.0	0.00	1	14	*	0.2
3	R02			**614**	513.0	16.45	0.00	*	0.1	**614**	614.0	0.00	1	10	*	0.2
4	R03			**507**	446.0	12.03	0.00	*	0.2	**507**	507.0	0.00	1	16	*	0.2
5	R04			**573**	443.0	22.69	0.00	*	0.2	**573**	573.0	0.00	1	16	*	0.3
6	R00	(b)		**572**	556.0	2.80	0.00	*	0.2	**572**	572.0	0.00	1	15	*	0.5
7	R01			**589**	412.1	30.03	0.00	*	0.1	**589**	589.0	0.00	1	17	*	1.2
8	R02			**590**	590.0	0.00	0.00	*	0.1	**590**	590.0	0.00	1	15	*	0.9
9	R03			**480**	399.0	16.88	0.00	*	0.1	**480**	480.0	0.00	1	16	*	0.6
10	R04			**559**	452.1	19.12	0.00	*	0.2	**559**	559.0	0.00	1	14	*	0.8
11	R00	(c)		**699**	699.0	0.00	0.00	*	0.2	**699**	699.0	0.00	1	21	*	0.2
12	R01			**783**	527.0	32.69	0.00	*	0.3	**783**	783.0	0.00	1	12	*	0.1
13	R02			**746**	548.0	26.54	0.00	*	0.2	**746**	746.0	0.00	1	12	*	0.1
14	R03			**616**	400.6	34.97	0.00	*	0.3	**616**	616.0	0.00	1	18	*	0.1
15	R04			**678**	438.6	35.31	0.00	*	0.3	**678**	678.0	0.00	1	18	*	0.2
16	R00	(d)	(ii)	**725**	581.0	19.86	0.00	*	1.2	**725**	725.0	0.00	1	30	*	4.1
17	R01			**741**	534.0	27.94	0.00	*	1.5	**741**	741.0	0.00	1	28	*	5.8
18	R02			**776**	514.9	33.65	0.00	*	4.9	**776**	776.0	0.00	1	25	*	5.7

(continued)

Table 3. *(continued)*

Instance				ILP formulation						Branch-and-Price						
ID	R	F	D	UB	LB_0	$\%g_0$	%g	Opt	t(s)	UB	LB_0	$\%g_0$	#N	#C	Opt	t(s)
19	R03			**693**	545.2	21.33	0.00	*	1.5	**693**	693.0	0.00	1	26	*	4.5
20	R04			**679**	545.3	19.69	0.00	*	1.2	**679**	679.0	0.00	1	32	*	7.3
21	R00	(e)		**887**	580.8	34.52	0.00	*	3.5	**887**	887.0	0.00	1	51	*	1.8
22	R01			**932**	538.3	42.24	0.00	*	3.7	**932**	932.0	0.00	1	60	*	2.6
23	R02			**995**	522.0	47.54	0.00	*	3.3	**995**	987.5	0.75	11	58	*	8.4
24	R03			**870**	546.0	37.24	0.00	*	22.1	**870**	854.0	1.84	11	73	*	10.7
25	R04			**863**	541.0	37.31	0.00	*	20.4	**863**	863.0	0.00	1	72	*	3.3
26	R00	(f)		**1013**	591.9	41.57	0.00	*	3.5	**1013**	999.0	1.38	17	52	*	9.4
27	R01			**1054**	534.2	49.32	0.00	*	4.5	**1054**	1054.0	0.00	1	47	*	1.5
28	R02			**1077**	524.0	51.35	0.00	*	4.2	**1077**	1077.0	0.00	1	39	*	1.2
29	R03			**985**	546.0	44.57	0.00	*	33.9	**985**	985.0	0.00	1	47	*	1.6
30	R04			**977**	541.0	44.63	0.00	*	16.9	**977**	977.0	0.00	1	44	*	2.0
31	R00	(g)	(iii)	**844**	621.3	26.39	0.00	*	87.0	**844**	844.0	0.00	1	86	*	33.3
32	R01			**852**	604.0	29.11	0.00	*	42.0	**852**	852.0	0.00	1	52	*	13.6
33	R02			**813**	550.6	32.28	0.00	*	151.0	**813**	813.0	0.00	1	84	*	78.6
34	R03			**836**	624.0	25.36	0.00	*	133.2	**836**	836.0	0.00	1	80	*	54.2
35	R04			**763**	560.5	26.54	0.00	*	58.8	**763**	758.0	0.66	7	138	*	85.9
36	R00	(h)		**959**	623.0	35.04	0.00	*	80.5	**959**	959.0	0.00	1	77	*	7.2
37	R01			**1032**	604.0	41.47	0.00	*	67.2	**1032**	1032.0	0.00	1	81	*	8.3
38	R02			**944**	550.5	41.68	0.00	*	178.3	**944**	944.0	0.00	1	93	*	13.6
39	R03			**988**	621.3	37.12	0.00	*	231.5	**988**	988.0	0.00	1	90	*	17.1
40	R04			**842**	562.0	33.25	0.00	*	59.6	**842**	842.0	0.00	1	105	*	17.2
41	R00	(i)		**1162**	623.0	46.39	0.00	*	575.6	**1162**	1162.0	0.00	1	100	*	7.5
42	R01			**1176**	604.0	48.64	0.00	*	224.7	**1176**	1176.0	0.00	1	68	*	5.2
43	R02			**1129**	549.8	51.30	0.00	*	384.2	**1129**	1103.5	2.26	27	176	*	75.0
44	R03			**1095**	620.0	43.38	0.00	*	510.8	**1095**	1079.5	1.42	83	209	*	159.7
45	R04			**992**	563.1	43.24	0.00	*	366.8	**992**	992.0	0.00	1	96	*	6.9
46	R00	(j)	(iv)	**909**	632.0	30.47	0.00	*	2740.5	**909**	908.5	0.06	7	180	*	1873.4
47	R01			**935**	587.7	37.14	0.00	*	703.0	**935**	935.0	0.00	1	74	*	338.5
48	R02			**822**	566.7	31.06	0.00	*	864.8	**822**	822.0	0.00	1	125	*	1701.3
49	R03			**938**	668.9	28.69	0.00	*	608.3	**938**	938.0	0.00	1	104	*	751.9
50	R04			**855**	681.1	20.34	0.00	*	376.4	**855**	855.0	0.00	1	122	*	277.2
51	R00	(k)		**1094**	632.0	42.23	2.96		1h	**1094**	1076.0	1.65	49	344	*	1017.4
52	R01			**1088**	589.2	45.85	0.00	*	3302.0	**1088**	1088.0	0.00	1	147	*	79.0
53	R02			1010	563.5	44.21	8.86		1h	**1006**	1006.0	0.00	1	165	*	163.8
54	R03			**1115**	663.4	40.50	5.45		1h	**1115**	1103.0	1.08	41	333	*	752.2
55	R04			**1008**	668.6	33.67	0.00	*	1863.1	**1008**	1008.0	0.00	1	147	*	92.5
56	R00	(l)		1162	632.0	45.61	16.78		1h	**1151**	1151.0	0.00	1	102	*	44.2
57	R01			**1243**	587.1	52.77	15.04		1h	**1243**	1243.0	0.00	1	116	*	45.3
58	R02			1129	567.6	49.73	20.35		1h	**1126**	1126.0	0.00	1	121	*	63.3
59	R03			1227	672.7	45.18	19.29		1h	**1208**	1203.3	0.39	3	165	*	118.5
60	R04			**1098**	668.6	39.11	10.68		1h	**1098**	1098.0	0.00	1	131	*	57.9
61	R00	(m)	(v)	**1005**	665.0	33.83	19.33		1h	1118	—	-	1	98		1h
62	R01			**985**	690.8	29.87	20.68		1h	-	—	-	1	116		1h
63	R02			**884**	645.5	26.98	12.33		1h	1026	—	-	1	94		1h
64	R03			1081	687.8	36.37	22.32		1h	**1062**	—	-	1	106		1h
65	R04			**930**	680.0	26.88	13.92		1h	1108	—	-	1	152		1h
66	R00	(n)		**1196**	665.0	44.40	32.77		1h	-	—	-	1	143		1h

(continued)

Table 3. *(continued)*

Instance				ILP formulation						Branch-and-Price						
ID	R	F	D	UB	LB_0	$\%g_0$	%g	Opt	t(s)	UB	LB_0	$\%g_0$	#N	#C	Opt	t(s)
67	R01			1148	676.5	41.07	28.97		1h	**1113**	1113.0	0.00	1	222	*	930.3
68	R02			**1103**	624.0	43.43	28.10		1h	1154	–	-	1	213		1h
69	R03			1234	687.5	44.29	34.38		1h	**1162**	1162.0	0.00	1	216	*	2779.9
70	R04			1151	680.0	40.92	29.94		1h	**1101**	1065.5	-	6	262		1h
71	R00	(o)		1265	665.0	47.43	35.70		1h	**1255**	1255.0	0.00	1	149	*	441.2
72	R01			1431	674.1	52.89	40.85		1h	**1254**	1254.0	0.00	1	195	*	595.6
73	R02			1266	619.9	51.03	37.47		1h	**1182**	1182.0	0.00	1	206	*	3058.0
74	R03			1385	687.3	50.38	40.48		1h	**1337**	1303.0	2.54	58	418		1h
75	R04			1271	680.0	46.50	34.40		1h	**1191**	1191.0	0.00	1	187	*	432.8
76	R00	(p)	(vi)	2645	724.2	72.62	69.80		1h	**2072**	–	-	1	58		1h
77	R01			**3047**	769.8	74.74	73.52		1h	-	–	-	1	58		1h
78	R02			2619	753.6	71.23	68.56		1h	**1642**	–	-	1	56		1h
79	R03			2783	813.8	70.76	69.28		1h	**1793**	–	-	1	60		1h
80	R04			**1716**	717.6	58.18	54.50		1h	-	–	-	1	50		1h
81	R00	(q)		**1624**	726.6	55.26	53.45		1h	-	–	-	1	69		1h
82	R01			**1854**	755.3	59.26	57.43		1h	-	–	-	1	75		1h
83	R02			**1725**	755.6	56.20	54.00		1h	-	–	-	1	78		1h
84	R03			**1646**	816.7	50.38	48.72		1h	-	–	-	1	75		1h
85	R04			**1348**	715.1	46.95	43.18		1h	-	–	-	1	76		1h
86	R00	(r)		**1802**	729.6	59.51	58.84		1h	-	–	-	1	252		1h
87	R01			**1954**	755.7	61.33	59.95		1h	-	–	-	1	236		1h
88	R02			**1709**	757.5	55.68	54.94		1h	-	–	-	1	212		1h
89	R03			**1842**	810.6	55.99	53.93		1h	-	–	-	1	244		1h
90	R04			**1706**	717.5	57.94	56.25		1h	-	–	-	1	240		1h
Sum/Average				1109.8	612.4	39.11	15.64	52	1672.6	961.3	910.2	0.21	4.4	104.9	66	1140.8

6 Conclusions and Further Investigations

In this paper we introduced and proposed two exact methods to solve a generalization of the Double Vehicle Routing Problem with Multiple Stacks (DVRPMS). This generalization was called Double Vehicle Routing Problem with Multiple Stacks and Heterogeneous Demand (DVRPMSHD) in order to model a more realistic case of the DVRPMS, as customers usually have heterogeneous demand.

We have tested our approaches for the DVRPMSHD with 90 instances created based on instances proposed in the literature, and our computational experiments shows that the branch-and-price algorithm is able to solve most of them.

As future work for the DVRPMSHD we intend to develop heuristic algorithms in order to find good solutions in short computational time.

Acknowledgments. The authors thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and GAPSO Company for financial support of this project.

References

1. Grabara, J., Kolcun, M., Kot, S.: The role of information systems in transport logistics. Int. J. Educ. Res. **2**, 1–6 (2014)
2. Aggelakakis, A., Bernardino, J., Boile, M., Christidis, P., Condeco, A., Krail, M., Schippl, J.: The future of the transport industry. Technical report, Institute for Prospective and Technological Studies, Joint Research Centre (2015)
3. Dantzig, G.B., Ramser, J.H.: The truck dispatching problem. Manag. Sci. **6**, 80–91 (1959)
4. Desrochers, M., Lenstra, J.K., Savelsbergh, M.W.: A classification scheme for vehicle routing and scheduling problems. Eur. J. Oper. Res. **46**, 322–332 (1990)
5. Iori, M., Martello, S.: Routing problems with loading constraints. TOP **18**, 4–27 (2010)
6. Petersen, H.L., Madsen, O.B.: The double travelling salesman problem with multiple stacks - formulation and heuristic solution approaches. Eur. J. Oper. Res. **198**, 139–147 (2009)
7. Felipe, Á., Ortuño, M.T., Tirado, G.: The double traveling salesman problem with multiple stacks: a variable neighborhood search approach. Comput. Oper. Res. **36**, 2983–2993 (2009)
8. Carrabs, F., Cerulli, R., Speranza, M.G.: A branch-and-bound algorithm for the double TSP with two stacks. Networks **61**, 58–75 (2013)
9. Alba Martínez, M.A., Cordeau, J.F., Dell'Amico, M.: A branch-and-cut algorithm for the double traveling salesman problem with multiple stacks. INFORMS J. Comput. **25**, 41–55 (2013)
10. Iori, M., Riera-Ledesma, J.: Exact algorithms for the double vehicle routing problem with multiple stacks. Comput. Oper. Res. **63**, 83–101 (2015)
11. Silveira, U.E.F., Benedito, M.P.L., Santos, A.G.: Heuristic approaches to double vehicle routing problem with multiple stacks. In: 15th IEEE International Conference on Intelligent Systems Design and Applications, pp. 231–236. IEEE Press, Marrakesh (2015)
12. Desaulniers, G., Desrosiers, J., Solomon, M.M.: Column Generation. Springer Science & Business Media, Berlin (2006)
13. Savelsbergh, M.: A branch-and-price algorithm for the generalized assignment problem. Oper. Res. **45**, 831–841 (1997)

BigDimETL: ETL for Multidimensional Big Data

Hana Mallek[1,2(✉)], Faiza Ghozzi[1,2], Olivier Teste[1,2], and Faiez Gargouri[1,2]

[1] MIR@CL Laboratory, University of Sfax, BP 242, 3021 Sfax, Tunisia
[2] Université de Toulouse, IRIT 5505, 118 Route de Narbonne, 31062 Toulouse, France
mallekhana@gmail.com, jedidi.faiza@gmail.com, Olivier.Teste@irit.fr, faiez.gargouri@gmail.com

Abstract. With the broad range of data available on the World Wide Web and the increasing use of social media such as Facebook, Twitter, YouTube, etc. a "Big Data" notion has emerged. This latter has become an important aspect in nowadays business since it is full of important knowledge that is crucial for effective decision making. However, this kind of data brings with it new problems and challenges for the Decision Support System (DSS) that must be addressed. In this paper, we propose a new approach called BigDimETL (Big Dimensional ETL) that deals with ETL (Extract-Transform-Load) development process. Our approach focuses on integrating Big Data taking into account the MultiDimensional Structure (MDS) through a MapReduce paradigm.

Keywords: ETL · Data warehouse · Bigdata · Twitter · MapReduce · Parallel processing · Multidimensional structure

1 Introduction

The DSS [11] delivers useful data for supporting business decisions. It maintains permanently generated data in order to be stored and collected in the Data warehouse (DW). This integration process referred as ETL (Extract-Transform-Load), which is responsible for integrating data at regular intervals such as daily, weekly or monthly. Nowadays, the DSS is facing a huge volume of data due to the exponential growth of new and heterogeneous data sources such as web, social networks and ubiquitous applications (Smart-phones, GPS, etc.). According to IBM[1], 2.5 trillion gigabytes of data are generated every day. In fact, this huge amount of data introduces the concept of "Big Data", which is often defined by the 3Vs characteristics, i.e. Volume (quantity of data), velocity (the speed of generating, capturing and sharing of data), and variety (range of data types and sources). This data evolution creates big challenges in the decisional domain in

[1] http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_92

order to handle, store, transfer and analyze it, at the right time. Big data requires several sets of new integration technologies such as Hadoop [22] MapReduce [6], Hive [18], etc. These technologies present a new opportunity to process and analyze data for DSS. Therefore, supporting new requirements and technologies of Big Data is a challenging issue for the conventional ETL tools [2]. In this paper, we propose a new approach called BigDimETL for ETL process in Big Data context. It aims to adapt the Extraction and the Transformation phases of ETL processes with MapReduce (MR) technology, in order to minimize time consuming through the parallel data processing. Although, MR technology treats a huge amount of data, without taking into account the MultiDimensional Structure (MDS) of the DW. Thus, our solution is based on dividing the input data vertically according to the MDS metadata. The MDS is considered as a high level DW/ETL specific constructs [13] and it is dedicated to Online Analytical Processing (OLAP) and Business Intelligence applications. Consequently, decision makers are not concerned to learn a new language to analyze big data any more. As mentioned previously, our approach focuses on extraction and transformation phases. Accordingly, in the extraction phase, a division method is used in order to minimize the overload into the transformation and loading phases. While in the transformation phase, we choose to focus on the "select operation" since it is one of the most frequently used query. It was picked from different primordial operations like projection, conversion, etc. This query will be used in treating and filtering data steps. To validate our approach, we carried out a series of theoretical experiment which validate the feasibility of our proposed approach. The rest of this paper is structured as follows: A state of the art of studies dealing with ETL process, for big data and data warehouse, is presented in Sect. 2. In Sect. 3, we present our platform as well as several key concepts. In Sects. 4 and 5, algorithms of our proposed extraction and transformation approaches are presented. Finally, in Sect. 6, we describe our contribution and different analysis of the proposed algorithms complexity for testing the feasibility our approach.

2 Related Work

On the DSS field, various studies consider ETL modeling process as an efficient solution to acheive data warehousing projects. The objective of modeling is to reduce the costs of DW implementation, and produce useful information in MDS through OLAP cubes [16]. Some researchers have focused on business data sources to model ETL processes. These processes are modeled as new specific tasks which are implemented as a framework called ARCTOS [21] and EMD [9]. However, working with new conceptual constructions proves to be significant, but the non-standardization is still a limit since it is a vital asset in terms of modeling. Therefore, the standardization appeared with [19] which presents ETL processes as activities using Activity Diagram. In the same context, [8] uses BPMN language to design ETL process. With the works of [7] and [4] a semantic design of ETL processes is presented, based on RDF (Resource Description Framework) graph in order to generate semantic DW. On the other hand, several studies focused on the development of ETL functionalities model which are

considered as important sub-processes of ETL. For example, [17] designs CDC functionality (Changing Data Capture) as colored network (CPN) describing and presenting the behavior of ETL processes, and [15] presents SCD (slowly Changing Dimension) and SKP (Surrogate key Pipeline) through a coordination Reo language to improve the synchronization between the entities that interact with each other. Today, Information System researchers are facing a big explosion of web data, especially with social media (Facebook, Twitter, Youtube,). Hence, Traditional Database Systems (TDS) cannot support this big volume of data. Indeed, computing industry starts looking at new options, namely parallel processing to provide a more economical solution. In this case, other lines of works deal with data integration using new technologies introduced to handle Big Data context, such as Hadoop, MapReduce [6], Hive, etc. Consequently, it is very crucial to use these technologies to manage a big amount of data. In fact, previous works have shown the power and the gain in time for processing and storing large volumes of data using Hadoop framework. Thus, [14] proposed a web service called TAREEG that collect geographic data using MR paradigm to represent real spatial data. However, few studies have focused on the exploitation of DSS with the MR paradigm. Moreover, the amount of data pushes the experts to adapt ETL processes in order to support Big Data requirements. It raises several problems related to Big Data 3Vs characteristics, which must be considered while modeling ETL process. In fact, authors in [2,13] use MR paradigm to add the parallelism concept to these processes in a physical level of integration. Similarly we find [14] which uses MR and Hive system to define a Framework called CloudETL that supports the representation of DW Star schema. The MDS is, moreover, treated in the transformation phase. [2] proposed a Framework called P-ETL which presents a parallel ETL through the MR paradigm. This work employs the parallelism strategy while integrating data, without considering the MDS, which is considered as the paramount stage for further analysis operations. [1] uses Hive and Hadoop to support DW in a cloud computing environment and to build OLAP cubes. In fact, the MDS is designed to solve complex queries in real time [11]. According to [3], the performance of OLAP queries can be advanced if the data warehouse is perfectly selected and integrated. Besides, several works try to adopt querying OLAP in Big Data context [5]. In fact, these previous studies are very interesting in the data warehousing context. Although, these works share some similarities with ours, the pretreatment phase, including the extraction and the partitioning, aren't covered. As a conclusion, our goal is to reduce the cost of DW implementation and produce relevant information for decision support. Among the possible solutions, the MR paradigm proves a powerful solution for parallel processing of massive data. Hence, we propose to handle the extraction and the transformation phase according to a MDS from the first step using vertical partitioning.

3 BigDimETL Architecture Overview

Figure 1 presents our proposed BigDimETL architecture which adapts the three main steps of typical ETL process to MapReduce paradigm. Thus, our main concepts are defined in the following:

- Multidimensional structure composed of a list of dimensions of analyses (D), where, $D = \{D_1, D_2, ...\}$ and fact (F) where, $F = \{F_1, F_2, ...\}$ to present the subject of analysis. Each dimension D_i is composed of several attributes $Att = \{att_1, att_2...\}$. Each fact composed of numerical measures $M = \{m_1, m_2, ..\}$. A fact is linked to associated dimension.
- MapReduce is a framework developed by Google described as a programming model used for a parallel processing of massive data sets. The input data must be divided into several partitions according to the default size (64MB). Each split is browsed separately by several jobs to perform transformation tasks. Each job has to be specified by the means of two functions: The first function Map takes key/value pairs from splits as input. The resulted key/value pairs play the role of input value for the second function Reduce.
- ETL process is described as three principal phases (Extraction, Transformation, Loading), considered as the dynamic and responsible part dealing with the workflow synchronization and the transmission of data to be loaded into the DW [20]. It is a big challenge, in this typical process to deal with a big amount of data, and generate multidimensional data into the DW. According to [20], ETL processes present several critical operations needed for the transformation phase in order to model ETL. In our case, we will classify these operations in two main classes according to the type of operation: Elementary Operations (EO) or Complementary Operations(CO). While EO requires as input one operand (Select, Project, Conversion operations), the CO requires two or more operands as input(Union, Join, SUM operations). To remap our operations according to map and reduce functions while conserving key/value structure, we find that EOs can be running as map or reduce functions and COs need both functions of MapReduce. In this paper, we focus on the "select operation" for the BigDimETL that will be executed as MapOnly function.

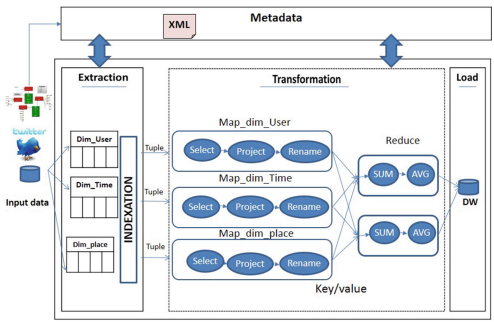


Fig. 1. BigDimETL approach

BigDimETL has three important phases described as follows: The extraction phase takes the case of capturing the data and then do the correspondence between the MDS metadata and the input data. In this phase, we need to split our input data into sub logical parts in order to work in a distributed and parallel aspect. We also need to ensure the vertical partitioning of column oriented data according to Facts and Dimensions described in the XML metadata. Hence, each separate part represents a group of columns related to a specific dimensions and facts. This phase faces two main challenges: (i) the large volume of the dataset and (ii) the time consumed when data is extracted. Details of the data extraction phase are presented in Sect. 4. Then, the data are transferred to the transformation phase, which covers several operations of data processing such as Selection, projection, conversion, etc. Thus, the transformation phase corresponds to Map and Reduce functions which cover all types of the transformation operations of ETL processes. This phase will present EO and CO. Finally, the loading phase feeds the processed data into the final database following MDS where the result of Reduce or Map phase is loaded.

4 Using Dimension Partitioning in Extraction Phase with Map Reduce

As mentioned in the introduction, Big Data is characterized by 3Vs (Volume, Velocity, Variety). "Twitter" shares the same Big Data characteristics, it is a popular social network used by millions of users to exchange news, informations, opinions in different domains such as political, commercial and educational domains. Therefore, the processing of these data is considered as a big challenge and it is hard to extract useful information for experts and decision makers. Thus, to help ETL processes to work efficiently, the metadata will be used as a dictionary guide for the input data content described by all the data elements. The MDS is described in a metadata XML file, including the global structure of the DW which should be given in the beginning. Figure 2 shows a part from our XML file which describes the star schema of Tweets [12]. This file will be saved as a file system in HDFS (Hadoop Data File System), in order to

```
<Tweet>
  <Dimensions>
    <Dimension name="Time" refD="ID-T">
      <Aggregation_level>
        <level> second </level>
        <level> Minute </level>
        <level> Hour </level>
        <level> Day </level>
        <level> Month </level>
        <level> Year </level>
      </Aggregation_level>
    </Dimension>
    <Dimension name="Place" refD="ID-PL">
      <aggregation_level>
        <level> Place_type </level>
      </aggregation_level>
      <aggregation_level>
        <level> Country-code </level>
        <level_attribute>country</level_attribute>
      </aggregation_level>
```

Fig. 2. Multidimensional structure of Twitter

be accessible during ETL steps. Therefore, we define at first, the structure of facts, which is composed of identifier "id-fact", foreign key of dimensions, and a list of measures. At the second sight, we find a description of the dimensions which is composed of a list of attributes identified by "id-dim". So, we have as input the metadata of our Data Warehouse and the Twitter data presented in a JSON format as a tree of objects.

4.1 Algorithm for Extracting Phase with BigDimETL

In our work, we partitionate the input data vertically in order to store them in distributed Database. The work of [10] considers vertical partitioning as partial loading data in order to ameliorate row data processing. Vertical partitioning is very significant, which leads us to use this principle in our case, especially, to create the MDS. Thus, we propose to convert the tree structure of JSON (object oriented) into columns structure (Columns oriented) (CSV). In this latter structure, the first line is conserved to present keys of elements and its values are reserved for the rest of file, in which each line is a record (tuple) that presents a tweet. The main advantage is that CSV structure is oriented columns which can represent different dimensions and facts clearly. Also, the structuring into CSV format is a compliant with MR paradigm [2]. In our case, the process to convert JSON structure to CSV file is based on a Metadata structure of our DW. The Algorithm 1 shows the details of the whole processes including converting, partitioning and indexing data using MR paradigm. The operations in the first step are reserved to parse our XML metadata, then the second step is reserved to parse JSON file and to compare the correspondence between XML elements and JSON keys. The map function's Key value, in this case, receives the JSON key attached with its nested element. Then, the value of the map function will receive its value. Next, with the use of the reduce function, we group all values which have the same key and we make it as CSV column. After the partitioning phase, we need to immediately index each partition in order to reduce the partition search area by indexing directly the desired partition. Browsing sequentially big amount of data is a fastidious operation, especially for query processing. The aim of indexing phase is to identify each partition contents to facilitate parsing the information. In our context, the indexing phase is based on Dimensions identifiers "id_dim".

Example: We consider to integrate tweets into the ED. Each tweet is subdivided into several dimensions ("User", "Place", "Time" and "Source"). Taking as an example "User" dimension which is described as an elementary JSON object at JSON file and can be composed of sub JSON elements. With MR paradigm the key is the elementary JSON objects, for example: key= {"Name", "Location", "Lang"} and its value is {"Thaddeus", null, "en"}. In CSV file these object are described as NameDimension_attribute structure, "User_Name", "User_Location", "User_Lang". In this case, we obtain a grouping of attributes list in the same partition called "Dim_User". In this case, the key is the NameDimension_attribute and the value is a list of key values:

<key, value> = <user_name,{"Thaddeus", "Henry Allup", "philippe All"}>.

Algorithm 1. Extract and partitioning data

Input: MDD:Metadata document, FJ:JSON File, FCSV : CSV file
Output: Key list , Value list

```
Parsing Metadata
Extract dimensions
Extract facts
Parsing JSON Input data
Compare extracted dimensions (datt) and facts (fatt) according to Keys in JSON file
foreach Dimension and Fact do
    if datt = metadata dim or fatt= metadata fact then
        Key=dimension_name
        first line of CSV = key
        Value= list of value of the same key
        Put values in the rest of CSV file
Create CSV column,
Index Facts and Dimensions : give index for each group of columns (fact, dimension)
```

5 Selection Operation Algorithm in BigdimETL

The transformation phase is responsible for handling the extracted data. Typically, this phase is composed of series of operations (Select, Project, Conversion ...) applied on the extracted data. The huge volume of data requires a new methodology to be treated in the nick of time. So, the integration of MR principal in the classical operations can minimize a lot of time consuming. In our approach, BigDimETL integrates MR paradigm into DSS and offers the capabilities to manage ETL Operations. Hence, the transformation phase is treated as Map and Reduce functions. In this paper, we focus on SELECT operation since it is the most frequently used query. It is responsible of restricting specific data to be loaded into the DW. Moreover, this operation is the basic query to retrieve data, and enables the selection of one or many columns from oriented columns structure. The result of SELECT statement is a set of records from one or several sources. In our context, it is very interesting to parallelize the treatment of this operation. The basic syntax of SELECT statement in relational algebra is as follows: $\sigma(Q)_P$, Where Q is a selection condition or a predicat. This predicat is applied independently to each individual tuple t in a partition P.

Example: We need to extract all tuples from Dim_Place where the attribute "Place" is not NULL.
$t = \sigma(Place <> NULL)_{Dim_Place}$.
In our case, we need to translate the "select operation" statement according to MR structure(key, value). Thus, each Map operation retrieves specific dimension "Dim_Place" which is identified in the indexed file. Then, we apply the predicate verification $Q = Place <> NULL$. The Algorithm 2, proposed in this work,

affects key and value of received data: Hence, we have as input key each tuple t from the input data. The input value is the same tuple t as the key, where the predicate P is improved else the value will take 0 as value. Thus, we will take as input the partitioned CSV according to the dimensions and facts. As a result we find (key, value) in the (t,t) structure if the predicate is satisfied else we obtain (t,0). In this case, Selections really do not need the full power of MR. We can conserve the output data without the shuffle and the combination phases which are considered as time consuming phases.

Algorithm 2. Select operation

Input: $key = t$, $P = Predicat$, $value = t$, index= List of indexed partition
Output: key , Value

```
1 Select(P)
2 begin
3     Get partition index (direct access)
4     foreach tuple t in Dimension D do
5         if Predicat=true then
6             Value= tuple
7         else
8             Value = 0
9 return(Key, value)
```

6 Theoretical Experiments

As mentioned above, the twitter data were used as input for our approach. The default JSON format of these data leads to several problems such as: (i) Huge time consuming during the parsing process because of the existence of nested elements (object tree structure). (ii) In the case of nested elements, it is difficult to identify the next one, so we need to determine the start of the next JSON element. We propose to convert the JSON into CSV format and to execute the following select operation: $\sigma(Place <> NULL)_{Dim_{Place}}$ on these two formats with MapReduce algorithm in two different cases. The first case uses Horizontal Partitioning (HP) data and in the second case, we use Vertical Partitioning (VP) data according to MDS. In this section, we aim to analyze the feasibility of the proposed approach. Therefore, Table 1 presents the complexity of the select operation execution under HP and VP. Using the HP, we notice that the complexity of the select operation is $O(n^2)$ for the JSON and CSV formats. This complexity is due to the parsing of all elements and sub elements with nested loops which extract "Place element" from the JSON format. While for the CSV format, the complexity is the result of parsing all tuples and all columns in order to verify the predicate P. Consequently, the collect of the treated data on the reduce function for each next query job (corresponding to a join) presents a time consuming task. On the other side, we notice that the complexity of the select

operation using the VP is equal to $O(n)+O(n)$. The oriented columns structure (CSV) of tweets is composed by several partition (dimension or fact content). Hence, it is easy to select the reference of Place partition using the index file. As a result, the access will be straight to the suitable partition. As a conclusion, for the theoretical point of view, the cost of the same queries is estimated by the number of loops, and research time.

Table 1. Algorithm complexity of select operation

Input	Partitioning	Select operation complexity
JSON	Horizontally	$O(n^2)$
CSV	Horizontally	$O(n^2)$
	Vertically	$O(n)+O(n)$

7 Conclusion

In this paper, we have introduced the first and the second part of our approach BigDimETL. Our approach leads to integrate Big Data with conserving the MDS of DW. The integration of the parallelism aspect of ETL processes through MapReduce paradigm is very useful in our case. We concentrate on the definition with the extraction and Transformation phase with MapReduce paradigm. Hence, to distribute our input data we have used vertical partitioning according to dimensions of multidimensional structure described in the metadata. As a future work, we aim to implement the ETL phases, and support the various functionalities of ETL (CDC, SCD).

Acknowledgments. This work is dedicated to the soul of my supervisor, Dr. Lotfi Bouzguenda, who left us in juin 2016. We are very grateful for his help, his advice and his prestigious remarks. May his soul rest in peace.

References

1. Arres, B., Kabachi, N., Boussaid, O.: Building OLAP cubes on a cloud computing environment with MapReduce. In: ACS International Conference on Computer Systems and Applications, AICCSA, pp. 1–5 (2013)
2. Bala, M., Boussaïd, O., Alimazighi, Z.: P-ETL: parallel-ETL based on the MapReduce paradigm. In: 11th IEEE/ACS International Conference on Computer Systems and Applications, AICCSA, pp. 42–49 (2014)
3. Bellatreche, L., Schneider, M., Mohania, M., Bhargava, B.: PartJoin: an efficient storage and query execution for data warehouses. In: Kambayashi, Y., Winiwarter, W., Arikawa, M. (eds.) DaWaK 2002. LNCS, vol. 2454, pp. 296–306. Springer, Heidelberg (2002). doi:10.1007/3-540-46145-0_29
4. Berro, A., Megdiche, I., Teste, O.: Graph-based ETL processes for warehousing statistical open data. In: Proceedings of the 17th International Conference on Enterprise Information Systems, pp. 271–278 (2015)

5. Chung, W.C., Lin, H.P., Chen, S.-H., et al.: JackHare: a framework for SQL to NoSQL translation using MapReduce. Autom. Softw. Eng. **21**(4), 489–508 (2014)
6. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
7. Deb Nath, R.P., Hose, K., et al.: Towards a programmable semantic extract-transform-load framework for semantic data warehouses. In: Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP, pp. 15–24 (2015)
8. Akkaoui, Z., Mazón, J.-N., Vaisman, A., Zimányi, E.: BPMN-based conceptual modeling of ETL processes. In: Cuzzocrea, A., Dayal, U. (eds.) DaWaK 2012. LNCS, vol. 7448, pp. 1–14. Springer, Heidelberg (2012). doi:10.1007/978-3-642-32584-7_1
9. El-Sappagh, S.H.A., Hendawi, A.M.A., El Bastawissy, A.H.: Original article: a proposed model for data warehouse ETL processes. J. King Saud Univ. Comput. Inf. Sci. **23**(2), 91–104 (2011)
10. Jaspreet Kaur, K.K.: A new improved vertical partitioning scheme for non relational databases using greedy method. Int. J. Adv. Res. Comput. Commun. Eng. **2** (2013)
11. Kimball, R., Ross, M.: The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd edn. Wiley, Hoboken (2002)
12. Kraiem, M.B., Feki, J., Khrouf, K., et al.: Modeling and OLAPing social media: the case of Twitter. Soc. Netw. Anal. Min. **5**(1), 47:1–47:15 (2015)
13. Liu, X., Thomsen, C., Pedersen, T.B.: ETLMR: a highly scalable dimensional ETL framework based on MapReduce. Trans. Large-Scale Data Knowl. Cent. Syst. **8**, 1–31 (2013)
14. Liu, X., Thomsen, C., Pedersen, T.B.: CloudETL: scalable dimensional ETL for hive. In: 18th International Database Engineering & Applications Symposium, IDEAS, pp. 195–206 (2014)
15. Oliveira, B., Belo, O.: Using REO on ETL conceptual modelling: a first approach. In: Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP, DOLAP 2013, pp. 55–60 (2013)
16. Orlando, S., Orsini, R., Raffaetà, A., Roncato, A., Silvestri, C.: Trajectory data warehouses: design and implementation issues. JCSE **1**(2), 211–232 (2007)
17. Silva, D., Fernandes, J.M., Belo, O.: Assisting data warehousing populating processes design through modelling using coloured petri nets. In: 2013 - Proceedings of the 3rd International Conference on Simulation and Modeling Methodologies, pp. 35–42 (2013)
18. Thusoo, A., Sarma, J.S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., Murthy, R.: Hive: a warehousing solution over a map-reduce framework. Proc. VLDB Endow. **2**(2), 1626–1629 (2009)
19. Trujillo, J., Luján-Mora, S.: A UML based approach for modeling ETL processes in data warehouses. In: Song, I.-Y., Liddle, S.W., Ling, T.-W., Scheuermann, P. (eds.) ER 2003. LNCS, vol. 2813, pp. 307–320. Springer, Heidelberg (2003). doi:10.1007/978-3-540-39648-2_25
20. Vassiliadis, P., Simitsis, A., Skiadopoulos, S.: Conceptual modeling for ETL processes. In: Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP, DOLAP 2002, pp. 14–21. ACM, New York (2002)
21. Vassiliadis, P., Vagena, Z., et al.: ARKTOS: towards the modeling, design, control and execution of ETL processes. Inf. Syst. **26**(8), 537–561 (2001)
22. White, T.: Hadoop: The Definitive Guide. O'Reilly Media, Inc., Sebastopol (2012)

Knowledge Integration in Collaborative Environments Using Supervised Ontological Alignment

Leandro Pupo Natale(✉) and Nizam Omar

Universidade Presbiteriana Mackenzie,
Rua da Consolação, 978, São Paulo, Brazil
leobrasil@gmail.com

Abstract. The increasing demand for new technologies allow greater interactivity and participation of people in online environments. This interaction occurs, mainly, by social relationships in different levels, and the multimedia resource's search and sharing. These interactions produce a large amount of data and information. These have provided the rise of social networks, which operate with information and resource sharing tools. The information and resources produced in these collaborative environments are heterogeneous in essence, and stored in different kinds of data repositories. In order to provide a more easily and efficient way to found and share those data more efficiently, those repositories must be integrated in a single uniform format, automatically, and negotiated safe. Based on these aspects, this paper proposes ontological alignment method that allow integrate semantically different data and information repositories through an ontological layer, used to data conceptualization. This integration allows a collaborative system's users to exchange resources in a optimally way, using its own concepts as a filter to execute any tasks in the collaborative environment.

Keywords: Knowledge management · Ontological alignment · Collaborative systems

1 Introduction

The new Information and Communication Technologies (ICT) have made the relationship between people and companies more dynamic. This relationship is also gaining speed and complexity due to the amount of data and exchanged information.

The data collected by the Information Systems (IS), associated with various types of apps, allow the contents used in these transactions to offer greater knowledge about the different areas involved.

However this data is, at its core, heterogeneous and, as seen before, distributed into different information repositories. Analyzing, integrating and managing all this data so that they can transmit some information, in order to enable the extraction of knowledge of these bases, is one of the most complex computational challenges [1, 3, 7].

The task of integration of this data for more accurate analysis, systemic and social integration, has many inherent uncertainties. These uncertainties arise due to the representation format of the data and techniques used for information extraction [8, 9].

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_93

In order to minimize these uncertainties, this paper presents a model of representation of knowledge of these bases and heterogeneous data integration in an automatically, negotiated and safe approach. The WKI - Web Knowledge Integration [10] - model was developed with the goal of establishing alignment among different ontologies.

Thus, the use of efficient knowledge representation methods brings better results in integration activities. Ontologies are guided to represent domains of knowledge computationally, enabling computational agents to automatically interact with each other [11].

The ontologies alignment was originally defined [6] as a method for establishing an automatic translation between two ontologies. A direct mapping one-to-one may often not entirely express the conceptualization [6]. Thus the WKI promotes improved quality pairings, through the use of an interlanguage created from the interaction of the creators of ontologies.

The WKI is composed of the taxonomy model of ontologies to be aligned, algorithms used for extraction and creation of ontologies, ontological alignment algorithms and a set of operators that promote the interaction of the creators of each ontology. This algorithm is implemented in software architecture WProfile [14].

The WProfile architecture [14] was designed to manage a social network that allows the search and sharing multimedia features. Each multimedia feature in this system is represented by a local ontology. The interaction of users seeking and sharing resources is carried out in a transparent manner with respect to the local knowledge. This is possible through the ontological alignment model created by WKI.

The details of WKI, how it was implemented in WProfile architecture, and test cases used to validate the implementation, focusing on tasks of ontological alignments, will be presented in this paper.

2 Organization of Knowledge and Searches

The organization of knowledge for further search and sharing is an inherent task to humans. In the early days, knowledge was organized and compiled by hieroglyphics, papyrus scrolls and books. The storage of these multimedia documents began to be implemented in special buildings, called libraries.

The need for representation of information in order to allow searching and sharing has become a computational task. The complexity is not only in how to extract information from stored documents, but also in how to rank it according to their relevance, and how to share it in a collaborative environment.

The organization and classification of knowledge is a particular judgment of each user. This task depends directly on the context and the control of information linked to the relevance attributed by users. In addition, the relevance may change over time as new information becomes available or depending on the devices that will be used as tools.

The standard vocabulary of a thesaurus is a strong candidate to be used in indexing and searching of documents. Besides, it enables the creation of hierarchies and standard classifications of used concepts, reducing errors in search tasks, and also enabling a semantic identification.

Ontologies are intended to formally store the knowledge of a specific domain so that they can be reused by apps and people. However, a domain can have different settings, each represented in a different ontology. This heterogeneous scenario is one of the main features of a collaborative environment. Thus, integration techniques and ontological combination are needed to make ontologies reuse and sharing more efficient [15].

The ontological knowledge bases have large amounts of data and relationships, forming dense graphs with a high degree of vertices. For the processing of this type of base to be effective, the ontological foundation must be built following methodologies that allow its reuse, re-engineering, and efficient processing to facilitate integration with other systems [15].

The evolution of the Web has changed the way users perform queries; robust machines are currently being used to perform searches of large amounts of data. These machines deal with large amounts of information and enable the search process through the use of large indexes.

In this sense, different categories of queries were created, trying to automatically predict user intent [2]. Broder (2002) defined taxonomy for web searches, creating three classes: Navigational, Informational, Transactional.

A major problem of these techniques is related to correctly identifying the purpose of the search, which can be ambiguous. For example, if a user makes a search using the name of his or her favorite singer, he or she may be looking for biographical information, music, concerts, or even other singers who have a rhythmically similar music production. This problem of search ambiguity occurs because the sought context often is unknown [2].

According to Noy and Musen (1999) two types of ontological integration are defined: alignment and merge. The alignment consists in establishing simple bonds or mappings between two ontologies (not necessarily from the same domain), preserving the original ontologies. In this technique, rules should be used to interconnect two concepts, creating an association or interlanguage between the two ontologies. Ontological integration using merge techniques creates a new ontology from the original pair. The merge technique presupposes that the two ontologies are from the same domain; the concepts present in the initial ontologies should be present in the new, more general domain ontology created.

Ontological integration by alignment has a great advantage over the merge process because it allows to join two ontologies from different domains. The process may be conducted in three ways: automatically, supervisingly or in a hybrid form.

3 Ontological Integration

3.1 Solution Proposals

The paths followed by different authors who have sought the ontological alignment as a way to integrate ontologies of different domains followed mainly the use of merge process structures and structural alignment.

The PROMPT, algorithm developed to perform the alignment processes and Ontology merging [6] analyzes the relationships structure of the concepts that were paired in the neighborhood. This algorithm starts by selecting a set of classes based on the similarity of names. From this point, there is an interaction with the user that directs the alignment or the merger of the ontologies terms. In each interaction, the neighborhoods of each relationship are evaluated in order to optimize the process.

This method uses one of the principles of Knowledge Management techniques, the expert supervision during the process of building an ontology. This provides better results at the end of the process because the expert can direct the construction of knowledge.

The method ONIONS (Ontological Integration of Naive Sources) [13] uses the supervised form of ontologies alignment. This approach is guided by the stratification of generic, intermediate and domains ontologies. The alignments are created directly among the metadata from the different ontologies involved.

Unlike previous methods, FCA-Merge [12] considers, besides the ontologies to be matched, a set of domain documents from each of the ontologies. Thus, concepts are extracted from documents that belong to both ontologies.

Thus, if the C1 concept of ontology O1 has instances in a document containing instances of C2 concept of ontology O2, C1 and C2 are candidates to be considered the same concept.

Hwang (2012) describes the OntoMProcess as a method capable of integrating two ontologies from the same domain in a collaborative environment. In this method the integration of ontology is created from a "seed" concept and it should be conceptualized collaboratively. However, this process has some problems related to the treatment of redundancy of terms and definitions of the concepts for each user.

The OntoMProcess does not predict that each user should have a formally established local ontology, making the average time to build ontologies around 1.5 min, which is relatively long for a web environment. Thus, some semantic inconsistencies and redundancies in definitions are generated, and they must be solved by a system administrator [4].

Following different approaches to integration of knowledge and documents, the e-Share provides the manual integration among different user groups. The creation of sharing groups does not use any management tool of social networks; this way, it is not possible to perform searches for shared resources in this environment. This tool has no means to create a data taxonomy as a process optimization way.

The Mendeley, specialized in academic resources, allows the creation of research groups with free sharing of resources. This service has a ranking engine of articles as a way to optimize the sharing. There are also services that allow to index the articles and, thus, to facilitate local and global search. Similar services are provided by Google Drive and Research Gate.

In addition to these important features of integration of documents and databases, sharing is achieved through the use of e-mails. This infrastructure, which is widely scalable and presents high performance, provides greater interaction among individuals, especially in a corporative environment, organizing discussions hierarchically, and resulting in topics and sub-topics in the databases.

Despite the possibility of this type of interaction, users are not (directly) connected to each other through personal or professional interests because there is only one large email address book.

Analyzing the characteristics of a collaborative environment, local data repositories are created from each user's features. For these resources can be stored efficiently in order to allow the creation of a correct and consistent local ontology, specialized tools must be used for extraction and recovery of information. These tools should still be able to work with different types of features (such as text, audio, video, etc.) in order to provide users with better experiences and information storage possibilities.

3.2 The Web Knowledge Integration Method (WKI)

Grüninger (2005) proposes *FLOWS (First-order Logic Ontology for Web Services*) as modular ontologies, created from semantic inferences based on details and descriptions previously stored. It is structured as webServices, with autonomous, and preferably atomic, modules, self-contained, able to represent a conceptualisation and may be interconnected [11, 16] and reusable [17].

It comprises a set of autonomous ontological modules that are interconnected by connective intermodulares that allow access and import the knowledge contained in the connected modules.

The development of this method extends this concept by proposing the integration of humans, in collaborative environments through interconnected ontologies.

Thus, the WKI method is an ontological alignment method, based on local ontology of users from collaborative environments, social networks, and similar [10]. The WKI consists of an automated and a supervised phase. The main approach purpose of this method is to create an interlanguage between two ontologies from different domains, using the supervision of their experts.

This method adapts two important processes for ontological alignment in collaborative environments. The first one relates to the automatic processing, similar to that used by ONIONS, followed by an asynchronous supervised process.

Thus, the detailed activities necessary to accomplish this ontological alignment [10] follow below.

- Activity 1. Construction of Superior Level ontologies. First, ontologies should be built to different domains, enabling conceptualize multimedia features. In this step a semantic ontology used for the semantic attributes and properties conceptualization of each multimedia resource to be classified is also built.
- Activity 2. Construction of the local ontologies. At this stage, local ontologies via an information extraction process and metadata from user's multimedia resources are built. For this, the Superior Level ontologies from Activity 1 are used and each attribute or property can also be conceptualized.
- Activity 3. Taxonomic Analysis. Each user must classify and define the attributes and properties of his or her resources. This phase is of prime importance, because each user (ontological specialist) describes the particular meaning of each term and resource. This process allows the user to classify semantically terms such as:

  - Polysemic: A term can acquire different meanings depending on the context. For example, Java can mean an archipelago in Indonesia (in a geographical context), a programming language, or a type of drink made with coffee beans. A different concept for each Java meaning must be created.
  - Synonyms: Two or more different spelling terms have the same meaning.
  - Antonyms. Different terms with opposite meanings.
- Activity 4. Local Ontological Expansion. This step uses the API JENA inference engine to establish new relationships, expanding semantically local relationships. In this stage, a local semantic alignment occurs.
- Activity 5. Semantic Refinement. When a user performs a search for resources in a collaborative environment, he or she must provide a term that identifies it, if it is not already defined in its local ontology, it must refined by running Activity 2. In this activity a new class or instance can be created in local ontology.
- Activity 6. Ontological Selection. Process in which a search is performed in the various environmental ontologies in order to find equivalence with the search term. This process is called selection, as a selection of the ontologies that have some conceptualization of the searched term or its definitions is performed. In case of none of the ontologies is selected, the user must execute Activity 2 again in order to further refine the term or concept he or she is seeking for.

In this interaction, WKI uses the same concept used by the FCA-Merge, but in an extended form. If a C1 concept, or one of its derivatives concepts belonging to an O1 ontology (in this case, the local ontology) also has a similar concept C2 in a O2 ontology (any of the other local ontology in the environment), they are candidates to have some kind of relationship (not necessarily to be the same concept).

In this case, the pairing will occur only after the analysis of the creators of each of the local ontologies on Multi-Local Ontological Negotiation stage.

- Activity 7. Multi-Local Ontological Negotiation. In this activity each term is formalized in a negotiated manner, as users of the two ontologies analyze the concepts in order to accept or reject the new concept. For both users (the one who conducted the search, and the one with the desired feature) ontologies sections are presented in order to facilitate the understanding of the new concept.
- Activity 8. Interlanguage Creation. This is the activity in which the ontological integration interlanguage is created. Automatically, based on the concepts negotiated on Activity 7, an ontology for new concepts union is created through the Subclass-Of, Superclass-Of, Part-Of, Associated-by, Inverse-Of and Same-As relations.
- Activity 9. Ontological Enrichment. This last activity aims to carry out an inference operation in this new ontology integration based on the definitions and concepts of both adjacent ontologies. Unlike Activity 4, the semantic analysis performed at this point is limited only to transitive operations, avoiding false relations, because the adjacent ontologies newly connected, are from different domains.

To minimize incorrect ontologies combinations, local ontologies are integrated into pairs of entities. The negotiated pairing allows the creation of a mapping among the shared resources, creating an interlanguage of integration and minimizing the uncertainty of these relationships.

The WKI allows the repositories of multimedia features integration in an ontological level. Each repository must have a local ontology for its representation. This enables the creation of interoperability among local ontologies. The pairing of all ontologies provides the appearance of a general ontology of superior domain that describes all the concepts of all integrated data repositories.

4 Integrating Ontologies: Modeling and Architecture

This section presents the use of the WKI as support to integration and resource sharing in WProfile collaborative environment. This is a collaborative environment that allows users to perform searches and sharing of multimedia resources according to the semantic content of these resources, unlike the tools normally used.

This environment allows the user to include multimedia features, all these represented from a domain ontology. This ontological layer allows a better conceptualisation of stored resources.

From this ontological layer, the environment allows the creation of a semantic layer for each one of the content. This layer allows each user to computationally define particular concepts, related to each resource.

In this way, each user can define, in particular, the meaning of each term or expression related to the resource. All storage devices must have a RDF document describing information and providing a basic semantic layer. This layer uses an RDF document due to compatibility with OWL used to create the Ontologies of each user.

The ontological layer was created with ontologies from different domains, all based on the taxonomy created by Shema.org Consortium. This taxonomy, initially created to promote and improve the structure of information through markup elements in a web environment, has the attribute's definition for various types of entities and their relationships.

The semantic layer is defined as a part of each ontology domain, previously unified. Initially, there is only support to the simplest definitions as definitions "it is a", "it is not a" "it is a part of", "it consists of", "it is a sub-class of." All these definitions are specializations of existing terms in the OWL language.

The alignment of ontologies, conducted by WKI, starts from each search performed by users at WProfile environment. When a user searches for a term in the environment by following the activities defined by the WKI, the definition and conceptualization of the searched term is checked in the local ontology. If it is a term that has not been defined in the local ontology, the user is directed to perform this setting.

Then the environment starts the search of these terms and similar terms in all environmental ontologies. Finding similar terms, a notification is created for the user, owner of the local ontology, stating that his or her resources were found in a search.

Figure 1 shows the interlanguages of integration among the individuals of the environment. Each interlanguage is created from the concepts negotiation for every two individuals who sought information on the environment. Local data repositories have an ontological layer, forming an ontological database. This layer stores metadata of stored resources and the conceptualization of each term.

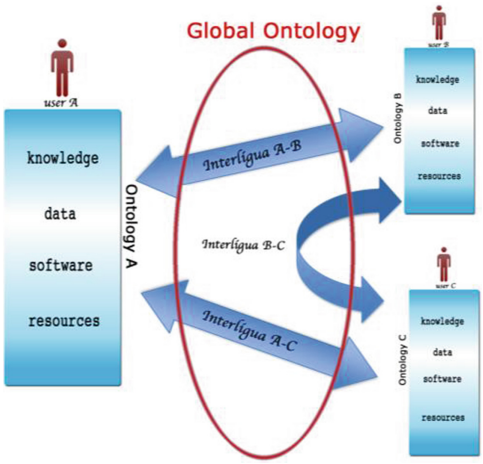


Fig. 1. WProfile architecture for semantic integration and resource sharing [14]

From the functional point of view these connections represent an interlanguage among each user's metadata. Structurally, the interlanguages form an ontological database, with conceptualized terms through collaborative interaction of the users in the environment, as shown in Fig. 2.

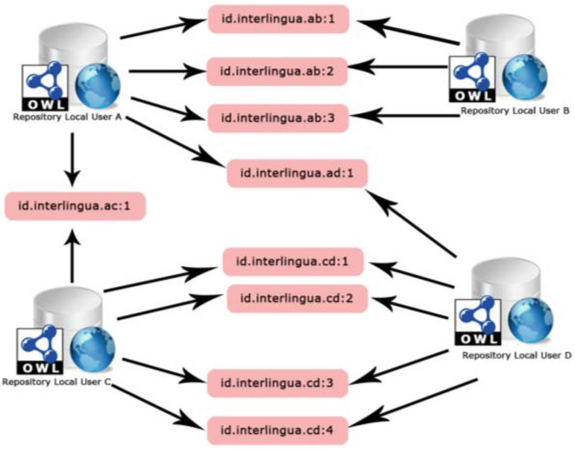


Fig. 2. Interlanguages representation [14]

Applications built based on this architecture have the ability to provide its users a greater interaction and efficient sharing of resources. Each user creates their profile in the collaborative environment, and has the ability to add various multimedia features with their descriptions. The interaction among individuals, from this point, can be managed and treated more efficiently and automatically by the environment [5].

Semantic integration among the data repositories of each user enables users' searches and resources in the environment to be carried out simply, and initially based on local ontology of each user. The environment begins to use the local ontology to create a semantic graph that will help it find other users in the environment. Complementarily, environmental ontological agents can use the semantic graph to find similar concepts in other stored local ontologies.

5 Conclusion

Collaborative environments managing systems such as social networks, knowledge management, collaborative learning, handle a large amount of data and heterogeneous resources. The integration of users and their resources must occur in a transparent manner and as automatic as possible, in order to provide better usability of the tools.

The creation of interlanguage, created from the concepts negotiation, enables the integration of data repositories and heterogeneous resources in an ontological perspective. Negotiation is a collaborative process.

In this sense, the ontological integration based on this kind of cooperation, creates a process that allows the emergence of a superior level ontology, created from a common sense. This ontology is refined and it evolves while the local ontologies evolve as well and new agreements are established, thereby refining the global ontology.

Stressing this model and checking for new languages that allow different types of integration is a plan for future work. From this point it will be possible to measure the quality that each type of iteration provides the integration of data repositories and resources.

References

1. Ashish, N., Bitton, D., Carey, M., Draper, D., Pollock, J., Rosenthal, A., Ssikka, V.: Enterprise information integration: success, challenges and controversies. In: SIGACM-SIGMOD, Baltimore, Maryland, EUA (2005)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca. Bookman Editora (2013)
3. Bernstein, P.A., Haas, L.M.: Information integration in the enterprese: a guide to the tools and core technologies for merging information from disparete sources. Commun. ACM, **51**, 72–79. September 2008
4. Hwang, D., Lee, K.I., Jung, J.J.: ONTOCS: a web-based system for collaborative ontology contruction. Comput. Inf. (2009)
5. Natale, L.P.: Utilização de Banco de Dados Ontológicos e Análise de Redes Sociais de Cidadãos em Sistemas de Governo Eletrônico. São Paulo, Brasil (2007)
6. Noy, N., Musen, M.: SMART: automated support for ontology merging and alignment. In: 12th Workshop on Knowledge Acquisition, Modeling and Management, Banff, Canada, October 1999

7. Seligman, L., et al.: Openll: an open source information integration toolkit. In: SIGMOD, Indianápolis, Indiana, EUA, June 2010
8. Dong, X., Halevy, A., Yu, C.: Data integration with uncertainty. Proc. VLDB **18**, 469–500 (2008)
9. Berners Lee, T., James, H., Ora, L.: The Semantic Web. Scientific American, maio (2001). Disponível em. http://www.scientificamerian.com/2001/0501issue/0501berners-lee.html. Acesso em: nov. 2004
10. Natale, L.P., Omar, N.: Compartilhamento de Informações entre Pares Utilizando Alinhamento Ontológico Negociado. In: Conferência IADIS Ibero-Americana WWW/Internet Lisboa (2016)
11. SWSF Committee. Semantic Web Services Framework (SWSF) overview, Disponível em: http://www.w3.org/Submission/SWSF. Acesso em Set 2010
12. Stumme, G., Maedche, A.: FCA-Merge: bottom-up merging of ontologies. In: IJCAI, vol. 1 (2001)
13. Steve, G., Gangemi, A., Pisanelli, D.M.: Integrating medical terminologies with ONIONS methodology. In: Information Modelling and Knowledge Bases IX, vol. 9, p. 1 (1998)
14. Natale, L.P., Omar, N.: Utilização de Banco de Dados Ontológicos em Integração de Bases de dados HEterogêneas. In: Proceedings of ICECE (2013)
15. Gomez-Perez, A., Fernández-López, M., Corcho, O.: Ontological Engineering: with Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web. Springer Science & Business Media, London (2006)
16. Hwang, D.S., Lee, I.K., Jung, J.J.: OntoCS: a web-based system for collaborative ontology construction. Comput. Inf. **28**(6), 781–793 (2012)
17. Shrivastava, J., Grüninger, M.: Developing an ontology for joints in furniture design. In: Cuel, R., Young, R. (eds.) FOMI 2015. LNBIP, vol. 225, pp. 88–99. Springer, Heidelberg (2015). doi:10.1007/978-3-319-21545-7_8

A General VNS Heuristic for a Three-Stage Assembly Flow Shop Scheduling Problem

Saulo C. Campos(✉), José Elas C. Arroyo, and Ricardo G. Tavares

Department of Computer Science, Universidade Federal de Viçosa - UFV, Viçosa, MG, Brazil
saulo.campos@ufv.br, jarroyo@dpi.ufv.br, ricardogoncalvetavares@gmail.com

Abstract. This paper studies a flow shop scheduling problem with three stages. At the first stage, different parts of a job are independently manufactured on m unrelated parallel machines. At the second stage, the manufactured parts are collected and transferred to the assembly stage where the parts are assembled into a final product. In this problem we considered sequence dependent setup times at each stage, and the objective to be minimize is the total tardiness. Motivated by the computational complexity and the practical relevance of the problem, we propose a heuristic based on General Variable Neighborhood Search with random VND (RVND) local search procedure to obtain near-optimal solutions. In the proposed RVND, a neighborhood is selected according to a selection probability that is self-updated with the number of improvements obtained by the neighborhood. As shaking procedure, we use a destruction-construction strategy that is used in the Iterated Greedy meta-heuristic. The effectiveness of the presented heuristic is evaluated by comparing its results to a basic GVNS, a Simulated Annealing proposed in the literature, and the CPLEX solver used to solve the mixed integer programming model of the problem. The obtained results have been statistically analysed.

Keywords: Assembly flow shop scheduling · Variable Neighborhood Search · Total tardiness · Sequence dependent setup times

1 Introduction

In this study we address a three-stage assembly flow shop scheduling (3sAFS) problem which is defined as following. There are n jobs to be scheduled, where each job consists of m parts. There are m unrelated parallel machines at the first stage (production or fabrication stage), and each machine just produces a specific part of a given job. When all m parts of the job are completed at the first stage, parts are collected and transferred to the assembly machine through the second stage (transportation stage). Finally, at the third stage (assembly

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_94

stage), all m parts are assembled by a single machine to obtain the final product. On the machines of the three stages, we consider sequence-dependent setup times (SDSTs). Setup time is the time required to prepare the necessary resource (machines) to perform a task (operation or job). Setup operations include obtaining tools, positioning work in process material, return tooling, cleanup, setting the required jigs and fixtures, adjusting tools, inspecting materials, etc. The most complex situation arises when the setup times are sequence-dependent, that is, the amount of setup time depends on the machine, the job that the machine was processing, and the job that comes next [1]. The objective is to find a schedule such that the total tardiness is minimized.

In this problem we only consider permutation schedules. When there is only one machine at the first stage ($m = 1$), then the problem is reduced to the three-machine permutation flow shop scheduling problem. The problem under study is NP-hard, since the single machine scheduling problem with total tardiness minimization is NP-hard [2].

The 3sAFS problem has been studied recently due to its applicability in realistic environments of a production system. To minimize the weighted sum of the mean flowtime and maximum tardiness in a 3sAFS problem with SDSTs only at the first stage, a tabu search (TS) and a Simulated Annealing (SA) metaheuristic were developed by Hatami et al. [3]. Andrés and Hatami [4] considered SDSTs at the first and third stages. They proposed a mixed integer programming (MIP) model for the total completion time minimization. Maleki et al. [5] considered SDSTs at the first stage and blocking times between each stage in such a way that the weighted sum of mean completion time and makespan is minimized. These authors propose a meta-heuristic based on SA. To the best of our knowledge, the 3sAFS problem with total tardiness performance measure and with SDSTs at the three stages has not been addressed so far. Allahverdi and Aydilek [6] considered a two-stages assembly flow shop scheduling without SDSTs. To minimize the total tardiness, they proposed heuristics based on genetic algorithm (GA) and SA. Allahverdi and Aydilek [6] showed that the SA combined with an insertion algorithm, named PSA, performs better than the GA.

In this paper, to tackle the 3sAFS problem under study, we develop a GVNS based local search heuristic [7]. In our heuristic we propose the combination of different new ideas, such as a random VND with selection probability, and a shake procedure based on random-destruction and greedy-construction as used in the Iterated Greedy meta-heuristic [8].

2 Problem Statement and MIP Formulation

The 3sAFS problem is formally defined as follows. There are n jobs to be scheduled such that each job has exactly $m + 2$ operations which should be processed by different machines. The first m operations are processed in parallel by m different machines at the first stage (production stage). The next two operations are processed at the second and third stages (transportation and assembly stages), respectively. That is, there is only one machine in the last two stages.

The transportation operation at the second stage may start only after all m operations at the first stage are completed. Similarly, the assembly operation at the third stage may start only after the transportation operation is completed. In this problem, the jobs are available for scheduling at time zero. Each machine is available continuously and can process only one job at a time. Preemption is not allowed, i.e. when an operation of a job is being processed, it continues until that operation is completed. In this problems permutation schedules are only considered, i.e., the sequence of jobs at each stage is the same.

For each job j, the following parameters are known: the processing time on machine k at the first stage (P_{kj}), due date (d_j), transportation time at the second stage (TT_j), and assembly time at the third stage (AT_j). Furthermore, between the processing of two consecutive jobs i and j, on a machine k at the first stage, there is a sequence-dependent setup time $S1_{kij}$. The setup times on the transportation and assembly machines are defined by $S2_{ij}$ and $S3_{ij}$, respectively, where i and j are two consecutive jobs. The goal of the problem is to find a schedule such that the total tardiness is minimized. For a given schedule S, the total tardiness is computed as follow: $f(S) = \sum_{j=1}^{n} max\{0, d_j - C3_j\}$, where $C3_j$ is the completion time of job j at stage three.

Based on the position-based model of Andrés and Hatami [4], we developed a MIP formulation for the 3sAFS problem. The goal of the model is to determine an optimal schedule (sequence of jobs which produces the minimum total tardiness). A sequence is a permutation of n jobs $\{j_1, j_2, ..., j_n\}$, where j_h is the h-th job of the sequence, i.e., the job processed in position h.

The following indices are used in the model:

$i, j = 0, 1, ..., n$: job index.

$k = 1, ..., m$: machine index at stage one.

$h = 1, ..., n$: position index in a sequence.

Note that we introduce a dummy job 0 which precedes the first job in the sequence. The following decision variables are used within the model:

X_{jh}: equal to 1 if job j occupies the h-th position in the sequence, otherwise zero.

$C1_{hk}$: completion time of the h-th job on machine k at stage one.

$C2_h$: completion time of the h-th job at stage two.

$C3_h$: completion time of the h-th job at stage three.

The mathematical model is defined according to Eqs. (1–14). The objective function (1) is to minimize the total tardiness. Constraint set (2) ensures that every position of a sequence has only one job assigned, while Constraint set (3) ensures that each job is assigned to exactly one position. Constraint sets (4) and (5) determine the completion time of every job on the machines at stage one by using the processing and sequence-dependent setup times of this stage. Constraint sets (6) and (7) determine the completion time of every job at stage two. Constraint set (8) guarantee that the processing of each job at stage two may start only after all m operations of the job at stage one are completed. These constraint set establishes the relationship between the completion times of jobs at stages one and two. Constraint sets (9) and (10) determine the

completion time of each job at stage three. Constraint set (11) establishes the relationship between the completion times of jobs at stages two and three, that is, the processing of a job at stage three may start only after the processing of this job at stage two is completed. The tardiness for each job is determined by constraint set (12). Constraints (13) represent non-negativity conditions of the decision variables $C1_{hk}$, $C2_h$, $C3_h$ and T_h, while constraints (14) ensure that X_{jh} are binary variables.

$$Min \sum_{j=1}^{n} T_j \tag{1}$$

$$s.t. \sum_{h=1}^{n} X_{jh} = 1, \quad \forall j. \tag{2}$$

$$\sum_{j=1}^{n} X_{jh} = 1, \quad \forall h. \tag{3}$$

$$C1_{1k} \geq \sum_{j=1}^{n} (S1_{k0j} + P_{kj}) X_{j1}, \quad \forall k. \tag{4}$$

$$C1_{hk} \geq C1_{(h-1)k} + \sum_{j=1}^{n} P_{kj} X_{jh} + \sum_{i=1}^{n} \sum_{j=1}^{n} S1_{kij} X_{i(h-1)} X_{jh}, \quad \forall h \geq 2, \forall k. \tag{5}$$

$$C2_1 \geq C1_{1k} + \sum_{j=1}^{n} (S2_{0j} + TT_j) X_{j1}, \quad \forall k. \tag{6}$$

$$C2_h \geq C2_{(h-1)} + \sum_{j=1}^{n} TT_j X_{jh} + \sum_{i=1}^{n} \sum_{j=1}^{n} S2_{ij} X_{i,h-1} X_{jh}, \forall h \geq 2. \tag{7}$$

$$C2_h \geq C1_{hk} + \sum_{j=1}^{n} TT_j X_{jh} + \sum_{i=1}^{n} \sum_{j=1}^{n} S2_{ij} X_{i(h-1)} X_{jh}, \forall h \geq 2, \forall k. \tag{8}$$

$$C3_1 \geq C2_1 + \sum_{j=1}^{n} (S3_{0j} + AT_j) X_{j1}. \tag{9}$$

$$C3_h \geq C3_{(h-1)} + \sum_{j=1}^{n} AT_j X_{jh} + \sum_{i=1}^{n} \sum_{j=1}^{n} S3_{i,j} X_{i,h-1} X_{j,h}, \forall h \geq 2. \tag{10}$$

$$C3_h \geq C2_h + \sum_{j=1}^{n} AT_j X_{jh} + \sum_{i=1}^{n} \sum_{j=1}^{n} S3_{ij} X_{i(h-1)} X_{jh}, \quad \forall h \geq 2. \tag{11}$$

$$T_h \geq C3_h - \sum_{j=1}^{n} d_j X_{jh}, \quad \forall h \geq 1. \tag{12}$$

$$C1_{hk} \geq 0, \quad C2_h \geq 0, \quad C3_h \geq 0, \quad T_h \geq 0, \quad \forall h, k, j \geq 1. \tag{13}$$

$$X_{jh} \in \{0, 1\}, \quad \forall j, h \geq 1. \tag{14}$$

We note that the above MIP model is non-linear, because there is a non-linear term $X_{i(h-1)}X_{jh}$ in constraints (5), (7), (8), (10) and (11). In order to obtain a linear model, we define an auxiliary binary variable Y_{ijh} to replace the non-linear term ($Y_{ijh} = X_{i(h-1)}X_{jh}$). The use of this variable requires the addition of the following constraints:

$$Y_{ijh} - X_{i(h-1)} \leq 0, \quad \forall h \geq 2, \forall i, j \geq 1. \tag{15}$$

$$Y_{ijh} - X_{ih} \leq 0, \quad \forall h, i, j \geq 1. \tag{16}$$

$$X_{i(h-1)} + X_{ih} - Y_{ijh} \leq 1, \quad \forall i, j \geq 1, \forall h \geq 2. \tag{17}$$

3 The GVNS-RVND Heuristic

General VNS (GVNS) is a VNS heuristic with variable neighbourhood descent (VND) as local search [7]. In the VND method, the change of neighbourhoods is performed in a deterministic way. In this work we use a Random VND (RVND) method to determine neighbour solutions. A pseudo-code description of our GNVS-RVND heuristic is given in Algorithm 1. The algorithm starts with an *Initialization* phase (line 1) in which the initial best solution S_{best} is generated by a constructive heuristic. In line 2 the shake variable, d, is initialized. The iterations of the algorithm are computed from line 3 to 14 until a *stopping* is satisfied. During each iteration, a *Shake* procedure is applied on S_{best} (line 4). This procedure consists in destroy randomly the solution S_{best} and reconstruct it obtaining a new solution S. The destruction depends on the shake variable d. Then, the RVND local search is applied on S (line 5), resulting in a local minimum S'. If the best solution is improved, it is updated and the variable d is set to its lowest value (line 7). Otherwise, the variable d is increased by one (line 9). When d reaches a maximum value (d_{max}), it is set to its lowest value (lines 10 and 11). The best obtained solution is returned by the Algorithm (line 15). In the next subsections, we describe the three procedures of our algorithm: *Initialization*, *Shake* and *RVND*.

3.1 Initialization Procedure

A solution S for the 3sAFS problem is represented by a permutation of the n jobs. An initial solution is determined as follows. First, the jobs are initially arranged in increasing order of the due dates. Two possible sequences containing the first two sorted jobs are evaluated. The best among the two is kept. Then, the third sorted job is inserted in all possible positions of the best sequence. The job is finally placed in the position resulting in the best total tardiness value. The process continues with the fourth job and completes when all jobs have been inserted, obtaining a complete sequence.

3.2 Shake Procedure

We use a Shake procedure based on Destruction and Construction of the current solution S. This procedure is used in the Iterated Greedy meta-heuristic [8].

Algorithm 1. GVNS-RVND (*stopping*, d_{max})

```
1: S_best ← Initialization( );
2: d ← 2;
3: while stopping not satisfied do
4:    S ← Shake(S_best, d); //Shaking
5:    S' ← RVND(S);
6:    if f(S') < f(S_best) then
7:       S_best ← S'; d ← 2;
8:    else
9:       d ← d + 1;
10:      if d > d_max then
11:         d ← 2;
12:      end if
13:   end if
14: end while
15: return S_best;
```

In the Destruction phase, a random block of d consecutive jobs are removed from the solution S, obtaining a partial solution Sp. The Construction phase inserts, one by one, all removed jobs from the block back into the partial solution. A removed job is inserted into all positions of Sp, obtaining ns new partial solutions. For example, if the first job is inserted into all positions, $ns = n-d+1$ partial solutions are generated. The best among the generated solutions is kept and it replaces Sp.

3.3 RVND Local Search

The RVND local search proposed in this paper is similar to one proposed by [9]. In our RVND, the neighborhoods are chosen randomly with a certain probability. The better neighborhoods have a greater chance of selection than worse ones. The steps of the proposed RVND are showed in Algorithm 2. Firstly, a neighbourhood list L containing nv predefined neighbourhoods is initialized (line 1). For each neighborhood k, an improvement counter c_k is defined (line 2). In the main loop of the algorithm (lines 3–12), the selection probabilities of the neighborhoods are defined (line 4). Note that, all the neighborhoods initially have the same selection probability. A neighbourhood $N_i \in L$ is chosen at random (with probability p_i) and then the best neighbour solution S' is determined (lines 5 and 6). If the S' improves the current solution S, the counter c_i of N_i is increased and the list L is populated with all the neighbourhoods (line 8), i.e., in the next iteration all the neighbourhoods will participate of the selection. Otherwise, N_i is removed from L. As the iterations progress, good neighborhoods will have the highest selection probabilities. The RVND local search finishes when L is empty.

In our RVND local search, three neighbourhoods are explored: N_1, N_2 and N_3. The moves that define N_1, N_2 and N_3 are job reallocation, two jobs exchange and two blocks (with two consecutive jobs) exchange, respectively.

Algorithm 2. RVND(S)

```
Initialize the Neighborhood List: L ← {N_1, ..., N_nv};
c_k ← 1, ∀k = 1, ..., nv; //Initialize the improvement counters
while L ≠ ∅ do
    p_k ← c_k / Σ_{N_i ∈ L}(c_i), ∀N_k ∈ L; //probability of selection
    Select a neighborhood N_i of L with probability p_i;
    S' ← Find the best neighbor solution in N_i(S);
    if f(S') < f(S) then
        c_i ← c_i + 1; L ← {N_1, ..., N_nv}; S ← S';
    else
        L ← L \ {N_i};
    end if
end while
return S;
```

4 Computational Experiments

We compare the proposed GVNS-RVND against the basic GVNS [7] and a SA heuristic (named PSA) [6]. GVNS uses a VND local search, in which the change of neighbourhoods is performed in a deterministic way. The *Shake* procedure and the neighbourhood structures are the same used in GVNS-RVND. The PSA heuristic was proposed in [6] to solve the 2sAFS problem. We adapt this heuristic for the 3sAFS problem. In GVNS and PSA, the initial solutions are generated as in GVNS-RVND. All of the heuristics have been coded in C++ and run on a PC with an Intel Core i5-3570 CPU (3.4 GHz), with 16 GB of the RAM.

To evaluate the performance of our heuristic, two sets of random instances were generated. Small-size instances: n and m are assume in {8, 10, 12, 14} and {2, 4, 6}, respectively. Large-size instances: n and m are assume in {30, 50, 80, 100, 200} and {5, 10, 20}, respectively. The processing times, setup times and due dates are uniformly distributed in the ranges $U[0, 99]$, $U[0, 49]$ and $U[3p, 3p + (n + 1)p]$, respectively. p is the average value of the processing times. For each configuration $n \times m$, 10 instances were generated. Therefore, a total of 120 small-size instances and 150 large-size instances were tested. To solve each instance, each heuristic was run 10 times with the same stopping criterion which is based on an amount of CPU time: $n \times m \times 100$ ms. For the parameter d_{max} used in GVNS and GVNS-RVND, three the values were tested: $3n/4, n/10, n/2$ and $n/4$. The best value found for d_{max} was $n/2$. The results of this test are not shown in detail due to reasons of space.

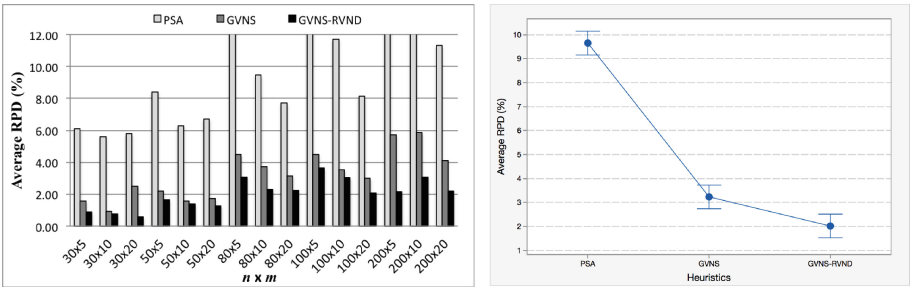
In order to obtain optimal solutions, the IBM-ILOG CPLEX 12.5 solver is applied to solve the mathematical MIP model with a threshold CPU time of two hours for each small instance. That is, if after the established time no optimal solution is obtained, the best current solution (upper bound) is returned by CPLEX. Table 1 presents the average RPDs (over 10 instances) of the heuristics PSA, GVNS and GVNS-RVND for each combination of n and m. Bold values represent best results. For a total of 120 small instances, 108 optimal solutions

were generated by CPLEX, from which 102, 82 and 69 optimal solutions were obtained by GVNS-RVND, GVNS and PSA, respectively. CPLEX was capable to obtain the optimal solutions for all the instances of up to 12 jobs. For these instances, the heuristics GVNS-RVND, GVNS and PSA present overall average relative distances (GAPs) of 0.2%, 0.47% and 1.14%, respectively. On twelve instances with $n = 14$, all heuristics perform better than CPLEX. For these instances, the heuristics GVNS-RVND, GVNS and PSA present average improvements of 4.42%, 4.25% and 2.24%, respectively. It is important to mention that, to solve an instance of size $n = 12$ and $m = 6$, CPLEX spend, on average, 926 s. For instances with $n = 14$ and $m = 6$, CPLEX consumed, on average, 4,990 s. The heuristics need 84 ms to solve an instance with $n = 14$ and $m = 6$.

Table 1. Results for small-size instances: avg. RPDs and CPLEX CPU times(s).

Instances $n \times m$	PSA	GVNS	GVNS-RVND	CPLEX (CPU times)
8×2	0.24	0.41	**0.31**	1.72
8×4	0.50	0.36	**0.17**	2.05
8×6	2.06	**0.00**	**0.00**	2.08
10×2	1.17	0.19	**0.00**	13.41
10×4	0.73	0.06	**0.07**	13.82
10×6	1.16	1.58	**0.74**	15.95
12×2	0.92	0.58	**0.31**	350.89
12×4	2.69	**0.00**	**0.00**	651.96
12×6	0.76	1.03	**0.19**	926.08
Mean	1.14	0.47	**0.20**	219.77
14×2	−0.50	−0.97	**−1.17**	6989.44
14×4	−0.39	−2.85	**−2.99**	5140.94
14×6	−0.89	**−1.07**	**−1.07**	4990.57
Mean	−0.60	−1.63	**−1.74**	5706.74

Next, we present the comparison of the heuristics on large-size instances. The performance of each heuristic is measured by the relative percentage deviation (RPD) measure determined by RPD $= \frac{f_{mean} - f_{best}}{f_{best}} \times 100$, where f_{mean} is the average total tardiness obtained by a given heuristic among ten runs, and f_{best} is the minimum total tardiness obtained among all the heuristics. The obtained RPDs are shown in Table 2. The average RPDs are grouped by the number of jobs and the number of machines ($n \times m$). For each group, 10 instances are considered. The heuristics GVNS-RVND, GVNS and PSA have overall average RPDs of 2.02%, 3.23% and 9.65%, respectively. Figure 1a summarizes the average RPDs of the heuristics. Each group of three bars represents the average RPD of the three heuristics over 10 instances, for each combination n and m ($n \times m$).



(a) Average RPDs (b) Mean plot and confidence intervals

Fig. 1. Comparison of the heuristics on large-size instances.

A short vertical bar implies that the solutions are closer to the reference solutions. The Figure shows that the performance of GVNS-RVND is better than that of GVNS, which in turn is far better than the PSA. Notoriously, the heuristic PSA present the worst performance.

Since the differences between the heuristics GVNS-RVND and GVNS are quite small, we carry out a non-parametric Kruskal-Wallis statistical test using the RPD measure as response variable. The results of the test indicate that there is statistically significant difference between the obtained results at a 95% confidence level. Figure 1b shows, for the three heuristics, the means plot and Tukey

Table 2. Average RPDs: results for large-size instances.

Instance n×m	PSA	GVNS	GVNS-RVND
30×5	6.09	1.55	**0.86**
30×10	5.60	0.91	**0.76**
30×20	5.79	2.50	**0.55**
50×5	8.42	2.19	**1.63**
50×10	6.29	1.55	**1.39**
50×20	6.70	1.73	**1.28**
80×5	13.65	4.47	**3.08**
80×10	9.48	3.72	**2.31**
80×20	7.71	3.15	**2.24**
100×5	13.29	4.49	**3.63**
100×10	11.70	3.54	**3.04**
100×20	8.12	2.97	**2.06**
200×5	15.43	5.72	**2.13**
200×10	15.20	5.86	**3.07**
200×20	11.32	4.11	**2.17**
Mean	9.65	3.23	**2.02**

honestly significant difference intervals with 95% confidence level from the statistical test. Non overlapping confidence intervals indicate that observed differences are statistically significant. From the perspective of performance, GVNS-RVND is statistically the best heuristic, followed by the heuristic GVNS.

5 Conclusions

In this study, a general VNS heuristic with random VND local search is proposed for solving the 3sAFS problem with SDSTs and total tardiness criterion. To the best of our knowledge, this problem has not been addressed so far. The main contributions of the proposed GVNS-RVND are as follows: (1) In the RVND local search, the neighborhoods are chosen with a certain probability (every time a chosen neighborhood improves the current solution, then its probability of choice is increased, thus, the best neighborhoods are more likely to be chosen). (2) We use a shake procedure based on random-destruction and greedy-construction. We have performed a comparative study between the proposed GVNS-RVND heuristic and the basic GVNS heuristic by using small and large-size instances. All the tested heuristics were confronted with the same amount of CPU time. Computational experiments reveal that the performance of GVNS-RVND is significantly better than GVNS and PSA.

Acknowledgments. The authors thanks the financial support of CNPq, CAPES and FAPEMIG, Brazilian research agencies.

References

1. Ruiz, R., Stützle, T.: An iterated greedy heuristic for the sequence dependent setup times flowshop problem with makespan and weighted tardiness objectives. Eur. J. Oper. Res. **187**(3), 1143–1159 (2008)
2. Du, J., Leung, J.Y.T.: Minimizing total tardiness on one machine is NP-hard. Math. Oper. Res. **15**(3), 483–495 (1990)
3. Hatami, S., Ebrahimnejad, S., Tavakkoli-Moghaddam, R., Maboudian, Y.: Two meta-heuristics for three-stage assembly flowshop scheduling with sequence-dependent setup times. Int. J. Adv. Manuf. Technol. **50**(9–12), 1153–1164 (2010)
4. Andrés, C., Hatami, S.: The three stage assembly permutation flow shop scheduling problem. In: V International Conference on Industrial Engineering and Industrial Management, pp. 867–875 (2011)
5. Maleki-Darounkolaei, A., Modiri, M., Tavakkoli-Moghaddam, R., Seyyedi, I.: A three-stage assembly flow shop scheduling problem with blocking and sequence-dependent set up times. J. Indust. Eng. Int. **8**(1), 1–7 (2012)
6. Allahverdi, A., Aydilek, H.: The two stage assembly flow shop scheduling problem to minimize total tardiness. J. Intell. Manuf. **26**, 1–13 (2013)
7. Hansen, P., Mladenović, N., Pérez, J.A.M.: Variable neighbourhood search: methods and applications. 4OR **6**(4), 319–360 (2008)
8. Ruiz, R., Stützle, T.: A simple and effective iterated greedy algorithm for the permutation flowshop scheduling problem. Eur. J. Oper. Res. **177**(3), 2033–2049 (2007)
9. Subramanian, A., Drummond, L.M.D.A., Bentes, C., Ochi, L.S., Farias, R.: A parallel heuristic for the vehicle routing problem with simultaneous pickup and delivery. Comput. Oper. Res. **37**(11), 1899–1911 (2010)

Towards NoSQL Graph Data Warehouse for Big Social Data Analysis

Hajer Akid[1,2(✉)] and Mounir Ben Ayed[1,2]

[1] Research Groups in Intelligent Machines, University of Sfax, National School of Engineers (ENIS), BP 1173, Sfax 3038, Tunisia
akid.hajer@gmail.com, mounir.benayed@ieee.org
[2] Faculty of Science of Sfax, BP 1171, Sfax 3000, Tunisia

Abstract. Big Data generated from social networking sites is the crude oil of this century. Data warehousing and analysing social actions and interactions can help corporations to capture opinions, suggest friends, recommend products and services and make intelligent decisions that improve customer loyalty. However, traditional data warehouses built on relational databases are unable to handle this massive amount of data. As an alternative, NoSQL (Not only Structured Query Language) databases are gaining popularity when building Big Data Warehouses. The current state of the art of proposed NoSQL data warehouses is captured and discussed in this paper. The paper will also focus on the opportunities and challenges of using NoSQL graph databases for storing and querying Big Social Data and how graph theory can help to mine information from these data warehouses.

Keywords: Big data · Social graph · Decision support systems · NoSQL data warehouses · Graph-NoSQL data warehouse · Graph theory · Big Social Data Analysis

1 Introduction

Social network is a term used to describe web-based services that allow individuals to construct a public or semi-public profile within a bounded system, articulate a list of other users with whom they share a connection, and view and traverse their list of connections and those made by others within the system [10]. During the last decade, the number of shared activities and exchanged opinions about products and services on social networks has grown exponentially. Therefore, many researchers and organizations are looking for making sense of this river of massive data to make better strategic decisions. Since data warehouses are accepted as the heart of decision support systems, an emerging topic is building Big Data Warehouses able to handle such amount of data. Traditionally, data warehouses are mainly built using R-OLAP (Relational-OLAP) approach based

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_95

on relational databases. Despite their maturity, these databases are facing many challenges [20,22] such as scalability and concurrency issues, high cost of the join operation and inability to analyse unstructured data. These problems make it difficult to store and manage big data effectively and enterprises are therefore turning to NoSQL database technology. Even if NoSQL data models and solutions have been well introduced, compared and discussed [3,11,12,14], no survey on proposed approaches to build NoSQL data warehouses is available. Hence, this paper intends to highlight the most important features of proposed logical models and make comparison between them to show advantages and drawbacks of each model. Further, and more importantly, Big Data has been mainly tied to Hadoop, MongoDB, and Cassandra. However, we believe that NoSQL graph databases are going to have a great impact on the technology landscape due to the re-emergence of interest in storing and analysing graph data. An important aspect of social networks is the connectivity of its nodes. Hence, we will introduce later a new V to describe Big Social Data. Storing and analysing these connections can give more informations about people's interests than their names and occupations. Unlike other database management systems, which require us to infer connections between entities using contrived properties such as foreign keys, or out-of-band processing like map-reduce, relationships are first-class citizens of the graph data model [24]. Although NoSQL graph databases have been recently introduced, the interest towards graphs is not new. Indeed, graph Theory, and more generally a graph, has been around for nearly 300 years and it is considered as one of the best and clearest way to represent and detect the structure defined by the way a relation connects different individuals of a social group [19]. In addition, many authors [2,15,27] proved that graph theory is a powerful and very useful in social networks analysis. Therefore, building NoSQL graph data warehouse and taking advantage of graph theory would be an efficient way to store and mine Big Social Data.

The reminder of this paper is organized as follows. Section 2 explains how the common V's of Big Data are not enough to characterize Big Social Data. In Sect. 3, we discuss related works. In Sect. 4, we argue the efficiency of building NoSQL graph data warehouse for Big Social Data Analysis. Our general approach is proposed in Sect. 5. Finally, Sect. 6 concludes the paper and presents future works.

2 Big Social Data

The term Big Data is an evolving term and it has been defined in several ways in the literature. According to Gartner, Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making and process automation. IDC, International Data Corporation, defines Big Data technologies as a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery and/or analysis. As for the Mckinsey

Global Institute (MGI), Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. However, since the appearance of the web 2.0 construct, the web has become highly connected. Thus, we believe that these definitions and the common 3 V's of Big Data are not enough to describe Big Data generated from social networking sites. We propose to add another V which is "Very-Connected". Very Connected data is data whose informations are related and where value is hidden in connections between things.

3 Related Works

Big Data analysis has become an emerging area. It aims to extract hidden values, provide suggestions and help decision makers. Many methods, architectures and tools for big data analysis must be established to meet the challenges of this phenomenon. Thus, new data warehouse and database technologies are introduced to address this problem. In particular, NoSQL are emerging as a new generation of databases. In this section, we discuss some related works in building Big Data Warehouses using NoSQL databases in order to introduce our approach. We will also focus on how graphs are the most suitable when dealing with connected data.

3.1 NoSQL Data Models

Recent years have seen a meteoric adoption of NoSQL, a family of data storage that emerged as an alternative to relational databases. There are four types of NoSQL data stores which vary by their data model.

Key-Value Stores. These systems are often considered as the most simple type of NoSQL stores. They store objects in the form of key-value pair. The key is the index used to fetch the data. All the key-value stores support insert, delete, and lookup operations and provide scalability through key distribution over nodes [3]. Yet, the queries supported by such systems are very simple since they can only cover the key. The users have to make efforts and extract the desired value form the block of returned data.

Over the past few years, many key-value databases have appeared such as: DynamoDB, LevelDB, GenieDB, RocksDB, Berkeley DB, Oracle NOSQL Database, Azure Table Storage, Riak, Redis and Voldemort.

Column Stores. Column family stores were motivated by the success of Googles BigTable. Unlike relational databases, a column-oriented database stores data by column rather than by row. Indeed, relational model has a static schema that requires the definition of the same columns for the several rows. However, the use of a column oriented database allows to define different columns for each line which avoids columns having NULL values. Column family databases are made up of column, super column, column family.

When you have volume and variety of big data, you should use a column-oriented databases such as: HBase, Cassandra, Hypertable, Amazon SimpleDB and BigTable.

Document Stores. In document stores, value associated with a unique key is a document which have a different schema to support more complex data forms. The structure of the documents and their parts is represented in JSON (JavaScript Object Notation), BSON (Binary JSON) or XML (eXtensible Markup Language) formats.

Document databases are most useful when you have to produce a large amount of data and they need to be dynamically assembled from elements that change frequently. MongoDB, SimpleDB, CouchDB and Couchbase Server are some examples of NoSQL databases using this model.

Graph Stores. Is a type of NoSQL database that uses graph theory to the storage of information about the relationships between entries such as the relationships between people in social networks or between items and attributes in recommendation engines. A graph database is essentially a collection of nodes and edges connecting each other through relations. Each node represents an entity such as a person, customer, businesses, accounts, or any other item and each edge represents a connection or relationship between two nodes. In a graph database, a node is defined by an unique identifier, a set of outgoing or incoming edges and a set of properties expressed as key/value pairs. Each edge is defined by an unique identifier, a starting and ending node and a set of properties.

Past surveys on NoSQL have done a great job of describing their applications, advantages and challenges. In particular, many authors [16,21,25] gave an overview on the multiple commercial and open-source implementations of NoSQL databases for helping users to "Use the right tool for the job". They have shown the need to choose the suitable NoSQL data model after evaluating data. In addition, some authors [17,27], focused on the comparison between relational and NoSQL databases based on their features and their tools. Many criteria such as data model, scalability, Big Data handling and ability of use for data warehousing were considered.

3.2 Graph Databases

Graph databases are the best way to model, store, and manage connected data. They are almost used in three major groups of applications which are chemical and biological data, social networks, and the web [26]. Far from being a recent data handling development, graphs ans graph theory were actually around for nearly 300 years and can be attributed to Leonhard Euler. Yet, it is only in the past few years that graph theory started to be used on information management. In that time, graph databases have helped solve important problems in the areas of social networking, master data management, geospatial, recommendations, and more [24].

Many surveys have discussed the opportunities behind graph databases. For instance, in [1] the author encouraged the use of graph data models in real life applications where interconnectivity is a key feature and concentrated on data structures, query languages, and integrity constraints. Also, some others made a comparison between graph databases and relational databases [26] based on objective and subjective measures. The objective measures include processing speed, disk space requirements and scalability. Subjective measures include maturity, ease of programming and security. Authors found that graph database did better than the relational at the structural type queries and full text character searches. In other work the author presented three major groups of applications for graph data which are chemical and biological data, social networks, and the web. Moreover, in an other paper [23], authors found that the use of relational databases in highly connected data applications such as web, computer networks, geographical structure is inefficient. Thereby he proposed Graph Database Management System to convert a relational database to graph database. This idea can be revisited taking advantage of the current graph NoSQL tools ans solutions.

As both of NoSQL surveys have evaluated its performance away from the field of decision support systems (DSS), this paper will focus more on the approaches proposed to implement Big Data warehouses using NoSQL.

3.3 NoSQL Columnar Warehouses

To build columnar big data warehouses, an indirect approach was proposed [18] to convert in two phases the multidimensional model of a data warehouse to HBase which is column-oriented database management system similar to Big Table. The first phase is to transform the relational schema of the data warehouse into HBase schema based on the data model of HBase. In the second phase, relationships between two schemas are expressed as a set of nested schema mappings. Another indirect approach [13] helped to implement three-dimensional data model, which uses the version dimension of HBase to store the values of a data item over time. Results showed that the performance was improved by the use of the third dimension of HBase. Despite their ability to build columnar NoSQL warehouses, these approaches are limited to the transformations of a logical model (Relational model) to another logical representation (NoSQL data model) and do not respect the constraints described by the conceptual level. This problem gave birth to another generation of approaches. First, a benchmark called CNSSB (Columnar NoSQL Star Schema Benchmark) was proposed for the columnar NoSQL data warehouse [9]. Yet, authors didn't focus on the mapping process. In a further work [8], they proposed three approaches called NLA, DLA, and DLA-CF to convert the multidimensional model to NoSQL data model. The NLA allows the mapping with a normalized approach. The DLA proposes transforming the data conceptual model into a model based on a large structure (table) called BigFactTable. The DLA-CF proposes transforming the data conceptual model into columnar NoSQL logical model. Results showed that DLA and DLA-CF are more efficient than NLA when queries requires to

join dimensions. Other authors [5] implemented two decision support systems based on column-oriented NoSQL and document-oriented NoSQL using respectively HBase and MongoDB. Results have shown the performance of the columnar NoSQL data warehouse when loading data generated by the benchmark TPC-DS.

3.4 NoSQL Document Warehouses

Document-oriented NoSQL attracted other researches. In [7], authors evaluated the performance, scalability and fault-tolerance of using MongoDB which is a document-oriented NoSQL database with Hadoop for scientific data analytic. Results shown important a great interest of using MongoDB for storage and Hadoop for analysis. However authors didn't give any detail on the transformation process. This latter has been the aim of a recent work [5]. A direct approach was proposed to build document NoSQL warehouse using MongoDB. Comparing this latter to column NoSQL warehouse has shown the performance of the document NoSQL warehouse when querying the database. Also, other authors [4] three approaches called MLD0, MLD1 and MLD2. The MLD0 approach transform every star schema to a collection of documents. This latter stores fact's measures and dimension's attributes as key-value pairs. The first approach doesn't use the concept of embedded documents. The MLD1 approach transforms every star schema to a collection of documents where measures are stored in an embedded document and the attributes that belong to the same dimension are also stored as embedded documents. The MLD2 approach store the star schema in a collection of documents using references instead the concept of embedded document to refer the documents of the corresponding dimensions. Results showed that MLD0 and MLD1 are more efficient than MLD2.

4 NoSQL Graph Data Warehouse: Opportunities and Challenges

Nowadays, social networks services have explosively grown. Social networks are often modeled as large graphs which contains a huge number of interconnected nodes. Building Graph NoSQL social warehouses requires to store the multidimensional data into graph-oriented NoSQL database. The transformation process have to express the concepts of multidimensional model in terms of nodes and edges. Building such a data warehouse can help to navigate on the collected data for supporting decision makers and make recommendations. However, while On-line analytical processing (OLAP) is a powerful primitive for structured data analysis, it faces major challenges in manipulating interconnecting data. Thus, authors [6] proposed in recent work a new data warehousing model, namely Social Graph Cube to support OLAP technologies on multidimensional social networks. Other challenges are the absence of a standard query language and the variety of data models. Research and industry will certainly influence the future direction of NoSQL graph database.

In summarize, both of direct and indirect approaches enabled the built of big data warehouses. Yet, only the direct approaches have considered the conceptual multidimensional model during the transformation process. The main drawback of the direct approaches hierarchies didn't be conserved after the transformation process. Here we can mention that the concepts of embedded documents and super columns can be of great interest to store hierarchies. Also, we noticed that only column data models and document data models were chosen as target models.

5 NoSQL Graph Data Warehouse: A Novel Approach

We propose in this section a new architecture for the construction of a graph data warehouse for the storage and analysis of Big Social Data. As this category of database management systems is adapted to the complexity of the problem, not volume, we find it necessary to use first a document-oriented NoSQL DBMS for the storage of the information shared on the social web. Than we will store in graph NoSQL database the nodes and links that we want to analyse. Although the approaches proposed in the literature only allow to migrate from a relational database management system to another NoSQL database management system, we will propose a new Intra-NoSQL approaches to migrate from NoSQL to NoSQL. This is a recursive approach. The loading of data will be automatic using the ETL process. Finally, we will focuse into the analysis and extraction of relevant informations using the graph algorithms known by their maturity (in-depth courses, courses in widths, etc.).

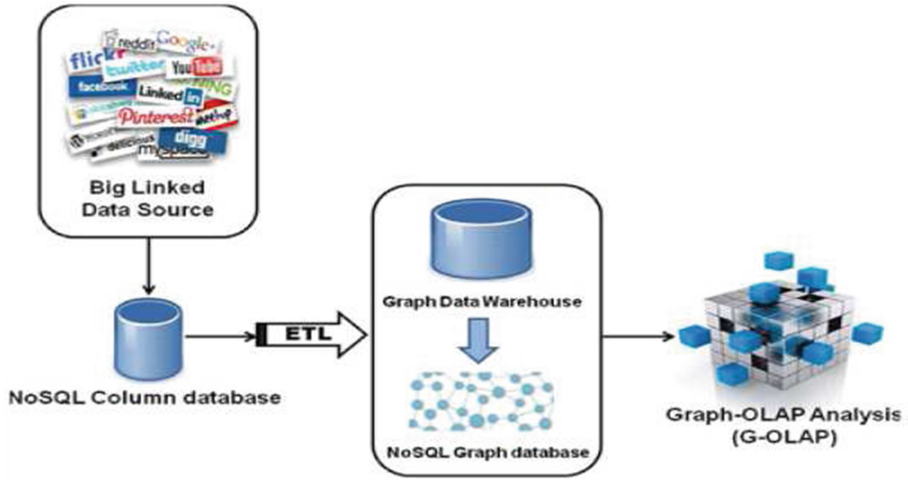


Fig. 1. Building NoSQL graph data warehouse

6 Conclusion

This paper gave an overall summary of the current state of big data warehouses. We focused almost on the process of transformation that convert multidimensional model to NoSQL data model. Moreover, we discussed the opportunities and challenges behind the implementation of graph NoSQL warehouses in the context of social networking. Indeed, the use of Graph NoSQL databases to data warehouse linked data such as social data is of great interest. Therefore we proposed a novel approach to build NoSQL Graph Data Warehouse to storage and mine Big Social Data.

References

1. Angles, R., Gutierrez, C.: Survey of graph database models. ACM Comput. Surv. (CSUR) **40**(1), 1 (2008)
2. Cartwright, D., Harary, F.: A graph theoretic approach to the investigation of system-environment relationships. J. Math. Soc. **5**(1), 87–111 (1977)
3. Cattell, R.: Scalable SQL and NOSQL data stores. ACM SIGMOD Record **39**(4), 12–27 (2011)
4. Chevalier, M., El Malki, M., Kopliku, A., Teste, O., Tournier, R.: Implantation Not only SQL des bases de données multidimensionnelles. In: Colloque VSST (2015)
5. Chevalier, M., El Malki, M., Kopliku, A., Teste, O., Tournier, R.: Implementing multidimensional data warehouses into NOSQL. In: 17th International Conference on Enterprise Information Systems (ICEIS15), Spain (2015)
6. De Virgilio, R., Maccioni, A., Torlone, R.: R2G: a tool for migrating relations to graphs. In: EDBT, pp. 640–643 (2014)
7. Dede, E., Govindaraju, M., Gunter, D., Canon, R.S., Ramakrishnan, L.: Performance evaluation of a mongoDB and hadoop platform for scientific data analysis. In: Proceedings of the 4th ACM Workshop on Scientific Cloud Computing, pp. 13–20. ACM (2013)
8. Dehdouh, K., Bentayeb, F., Boussaid, O., Kabachi, N.: Using the column oriented NOSQL model for implementing big data warehouses. In: Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), p. 469. The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2015)
9. Dehdouh, K., Boussaid, O., Bentayeb, F.: Columnar NoSQL star schema benchmark. In: Ait Ameur, Y., Bellatreche, L., Papadopoulos, G.A. (eds.) MEDI 2014. LNCS, vol. 8748, pp. 281–288. Springer, Heidelberg (2014). doi:10.1007/978-3-319-11587-0_26
10. Ellison, N.B., et al.: Social network sites: definition, history, and scholarship. J. Comput.-Mediat. Commun. **13**(1), 210–230 (2007)
11. Gajendran, S.K.: A Survey on NOSQL Databases. University of Illinois, Champaign (2012)
12. Grolinger, K., Higashino, W.A., Tiwari, A., Capretz, M.A.: Data management in cloud environments: NOSQL and NEWSQL data stores. J. Cloud Comput.: Adv. Syst. Appl. **2**(1), 1 (2013)

13. Han, D., Stroulia, E.: A three-dimensional data model in HBase for large time-series dataset analysis. In: 2012 IEEE 6th International Workshop on the Maintenance and Evolution of Service-Oriented and Cloud-Based Systems (MESOCA), pp. 47–56. IEEE (2012)
14. Han, J., Haihong, E., Le, G., Du, J.: Survey on NOSQL database. In: 2011 6th international conference on Pervasive Computing and Applications (ICPCA), pp. 363–366. IEEE (2011)
15. Harary, F., Norman, R.Z.: Graph Theory as a Mathematical Model in Social Science. University of Michigan, Ann Arbor (1953)
16. Hecht, R., Jablonski, S.: NOSQL evaluation. In: International Conference on Cloud and Service Computing, pp. 336–41. IEEE (2011)
17. Jatana, N., Puri, S., Ahuja, M., Kathuria, I., Gosain, D.: A survey and comparison of relational and non-relational database. Int. J. Eng. Res. Technol. **1**(6), 1–5 (2012)
18. Li, C.: Transforming relational database into HBase: a case study. In: 2010 IEEE International Conference on Software Engineering and Service Sciences, pp. 683–687. IEEE (2010)
19. Martino, F., Spoto, A.: Social network analysis: a brief theoretical review and further perspectives in the study of information technology. PsychNology J. **4**(1), 53–86 (2006)
20. Mohamed, M.A., Altrafi, O.G., Ismail, M.O.: Relational vs. NOSQL databases: a survey. Int. J. Comput. Inf. Technol. **3**(03), 598–601 (2014)
21. Moniruzzaman, A., Hossain, S.A.: NOSQL database: new era of databases for big data analytics-classification, characteristics and comparison. arXiv preprint arXiv:1307.0191 (2013)
22. Nayak, A., Poriya, A., Poojary, D.: Type of NOSQL databases and its comparison with relational databases. Int. J. Appl. Inf. Syst. **5**(4), 16–19 (2013)
23. Prasanth, N., Arul, K.: Converting employee relational database into graph database. Middle-East J. Sci. Res. **22**(11), 1618–1621 (2014)
24. Robinson, I., Webber, J., Eifrem, E.: Graph Databases: New Opportunities for Connected Data. O'Reilly Media, Inc., Sebastopol (2015)
25. Tudorica, B.G., Bucur, C.: A comparison between several NOSQL databases with comments and notes. In: 2011 RoEduNet International Conference 10th edn.: Networking in Education and Research, pp. 1–5. IEEE (2011)
26. Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., Wilkins, D.: A comparison of a graph database and a relational database: a data provenance perspective. In: Proceedings of the 48th Annual Southeast Regional Conference, pp. 42. ACM (2010)
27. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press, Cambridge (1994)

A Contribution of Dynamical Systems Theory and Epidemiological Modeling to a Viral Marketing Campaign

João N.C. Gonçalves[1(✉)], Helena Sofia Rodrigues[2,3], and M. Teresa T. Monteiro[4]

[1] Department of Production and Systems, University of Minho, Braga, Portugal
jncostagoncalves@gmail.com
[2] Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, Aveiro, Portugal
sofiarodrigues@esce.ipvc.pt
[3] School of Business Studies, Viana do Castelo Polytechnic Institute, Valença, Portugal
[4] Algoritmi R&D Center, Department of Production and Systems, University of Minho, Braga, Portugal
tm@dps.uminho.pt

Abstract. Nowadays, the interest in analyze and study the behavior of uncontrollable nature phenomena related to the impact of marketing campaigns is an action of prime importance to prevent chaotic dynamics. In this paper we assess the influence of Dynamical Systems theory and Mathematical Epidemiology on a real viral marketing campaign: Dove Real Beauty Sketches, based on a SIR epidemiological model. Motivated by the overwhelming success of this campaign, we study the mathematical properties and dynamics of the campaign real data - from the parameters estimation and its sensitivity to the stability of the mathematical model, simulated in Matlab. Mathematically, we show not only that the campaign was a viral epidemic, but also that it can be leveraged and optimized by epidemiological and mathematical modeling, which offer important guidelines to maximize the impact of a viral message and minimize the uncertainty related to the conception and outcome of new marketing campaigns.

Keywords: Marketing campaign · SIR epidemiological model · Optimization · Stability and equilibrium

1 Introduction

Viral Marketing (VM) can be seen as an advertising strategy that benefits from the impact of network effects, mainly of word-of-mouth, within a specific set of

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_96

population [1]. Due to the fact that, according recent studies, people tend to trust and be influenced by their social network (e.g. friends, coworkers or relatives) in detriment of standard media, VM is one of the most efficient strategies to face stochastic behavior related to general viral marketing campaigns [1]. Unlike most other marketing techniques, VM is a powerful marketing option to reach the masses with a low cost associated to it [2]. Whereas, in traditional marketing, communication strategies are deeply concentrated in the customer, VM creates mechanisms in order to avoid the implication of the original source, by increasing information exchange among customers and consumers [3]. Moreover, VM has a snowball effect by replicating the message among the susceptible population, increasing the number of infected individuals [4].

However, to be successful, this marketing method requires that the message involved in the process must be attractive to the target audience. Hence, in order to meet the challenge of increasing competition among marketing companies, researchers and marketing professionals have been trying to design marketing advertisements with robustness and resistance to fast followers [5]. In this regard, since the behavior of an epidemic is analogous to VM, the relationship between Epidemiology and Marketing, despite being recent, has been increasingly explored by researchers [6]. As the main motivation, to contribute to the development of this relationship, this paper studies mathematical properties of a specific viral campaign - Dove Real Beauty Sketches - modeled by a SIR epidemiological model under Dynamical Systems and Mathematical Epidemiology theory. As a result, we show some aspects which stand out how can VM campaigns be improved, during its planning phase. In a sense, the choice of this viral campaign is related to the fact that it was a perfect sample of an advertising that combines both solid marketing strategies and resistance to imitation.

Produced in 2013, Dove Real Beauty Sketches is a short movie within the marketing campaign for Real Beauty by Dove. As a viral marketing campaign, Dove Real Beauty Sketches focuses on state the definition of beautiful, promoting self-esteem among women from different age and make themselves safe and confident about their look [7].

The campaign was a huge success across the globe, considered the most viral video advertisement of all time [8], and represents an excellent example of how to build a successful viral marketing campaign. The article is organized as follows. In Sect. 2, supported by Dove Real Beauty Sketches marketing campaign real data, is formulated and discussed the mathematical model. After the model formulation, Sect. 3 presents the parameters estimation, simulated in Matlab. In Sect. 4 we study the dynamics of the model, under Dynamical Systems and Mathematical Epidemiology theory. In addition, also in this section, the mathematical behavior is sustained by marketing concepts related to the campaign. Conclusions are carried out in Sect. 5.

2 SIR Epidemiological Model

Taking into consideration that VM can be modeled under standard epidemic models [5,9], we consider a SIR epidemiological model to analyze, over time, the dynamics of the Dove Real Beauty Sketches marketing campaign.

This model describes the variation, in time, of the susceptible, infected and recovered population by the following system of ordinary differential equations, subject to initial conditions related to the campaign real data [8]:

$$\begin{cases} \frac{dS(t)}{dt} = -\beta \frac{S(t)I(t)}{N} \\ \frac{dI(t)}{dt} = \beta \frac{S(t)I(t)}{N} - \gamma I(t) \\ \frac{dR(t)}{dt} = \gamma I(t)\ , \end{cases} \qquad \begin{cases} S(0) = 10^9 - 30000 \\ I(0) = 30000 \\ R(0) = 0\ . \end{cases} \tag{1}$$

$S(t)$, $I(t)$ and $R(t)$ represent, respectively, the mutually-exclusive compartments of susceptible, infected and recovered population at time t. In terms of marketing, we define compartment $S(t)$ as the audience that marketers want to reach with the marketing message; $I(t)$ as the individuals that transmit the message or recommend a product to friends or relatives, motivated by monetary interests or just for the attractiveness of the intended message, and $R(t)$ as the portion of population who stop transmitting the message because, for instance, it is no longer appealing. Regarding to the parameters in terms of epidemiology, $\beta, \gamma > 0$ traduce, respectively, the infectivity and recovery rate. In the marketing context, β can be seen as the predisposition to deal with the message and be motivated to share it, and γ as the interruption of the message transmission by individuals [5]. Regarding to the construction of the model equations, the rate $\beta S(t)I(t)$ results from the probability of transmit the message, within the susceptible population, by an individual who had contact with a susceptible one $\left(\frac{S(t)}{N}\right)$, multiplied by the number of new transmission cases in time (βN) per infected individual $(I(t))$. Over time, individuals stop to share the marketing message according to the rate $(\gamma I(t))$.

The total host population N, estimated by 10^9 [10], is considered constant in time, i.e.,

$$\frac{dN}{dt} = 0 \Leftrightarrow \frac{d(S+I+R)}{dt} = 0 \Leftrightarrow N = S(t) + I(t) + R(t),\ \forall t \in [0, \infty).$$

Another relevant aspect related to the dynamics of the SIR model is the asymptotic state of the susceptible population. In this sense, marketers should predict the expectable duration time of the success related to the campaign diffusion. Thus, by [11], the evolution of the number of susceptible individuals, over time, can be measured by the following expression

$$\frac{dS}{dR} = \frac{-\beta S}{\gamma N} \Rightarrow S(t) = S(0) \exp\left(\frac{-\beta R(t)}{\gamma N}\right),\ \forall t \in [0, \infty)\ .$$

Since $I_\infty = 0$, we can also deduce that, over time

$$\lim_{\mathbb{R}_0^+ \ni t \to \infty} S(t) = S_\infty = S(0) \exp\left(\frac{\beta(S_\infty - N)}{\gamma N}\right)\ , \tag{2}$$

and, consequently,

$$R_\infty = N - S_\infty \ . \tag{3}$$

These expressions allow marketers to have a global view, in real time, about the evolution of the formulated system. As expected, according to (2), the number of individuals who have not been in contact with the viral message tends to decrease over time, since the number of infected individuals tends to grow up as long as the message continues to be appealing to the diffusion agents. Over time, by (3), when the message is no longer a novelty, the number of people who stop to share it increases. Noting that $R(t)$ derives from $S(t)$ and $I(t)$ and considering $\widehat{S} = \dfrac{S}{N}$, $\widehat{I} = \dfrac{I}{N}$, $\widehat{R} = \dfrac{R}{N}$, the SIR model (1) and its initial conditions can be rewritten as:

$$\begin{cases} \dfrac{d\widehat{S}(t)}{dt} = -\beta \widehat{S}(t)\widehat{I}(t) \\ \dfrac{d\widehat{I}(t)}{dt} = \beta \widehat{S}(t)\widehat{I}(t) - \gamma \widehat{I}(t) \\ \widehat{R}(t) = 1 - \widehat{S}(t) - \widehat{I}(t) \ , \end{cases} \qquad \begin{cases} \widehat{S}(0) = \dfrac{S(0)}{N} \\ \widehat{I}(0) = \dfrac{I(0)}{N} \\ \widehat{R}(0) = 0 \ . \end{cases} \tag{4}$$

To the marketing message be widely diffused, it is clear that the number of infected individuals must increase, i.e., $\dfrac{d\widehat{I}}{dt} > 0$. Considering $\widehat{S} \approx 1$ in this inequality, resulting from the fact that in the beginning of an epidemic almost everyone is susceptible, we get $\dfrac{\beta}{\gamma} > 1$. The basic reproduction number, $\mathcal{R}_0 = \dfrac{\beta}{\gamma}$, is defined as the expected number of secondary infections produced by a single infected individual, in a susceptible population, and governs the behavior of the SIR model, hence its importance. In the marketing context this number represents secondary transmissions sent by a single individual to a specific target audience. Moreover, if $\mathcal{R}_0 < 1$ the message is not widespread. On the other hand, if $\mathcal{R}_0 > 1$, the message is broadly sent among the susceptible population [5]. The basic reproduction number is determinant on the system dynamics, as we shall see later.

Note that, by the side of marketers, an accurate estimation of the parameters β and γ deeply helps to monitor the evolution of the campaign among the social network environment. Hence, in the next section, we estimate β and γ in order to adjust the model to the campaign real data.

3 Parameters Estimation Using an Optimization Approach

The estimation of β and γ has a significant importance in the model dynamics and adjustment to the campaign real data. However, it is not trivial to find strategies or recipes to estimate model parameters. Moreover, in the marketing field, this difficult increases exponentially, due to its erratic behavior. Another important aspect is linked to the fact that, related to the word-of-mouth and

network effects processes, social behavior is not predictable with a desired accuracy. So, find a perfect method to estimate parameters related to stochastic patterns it is quite difficult, and marketers must take these critical aspects into account upon the design of future campaigns, in order to have more control on the evolution of it. Although, there are accurate estimation methods among the literature to fit the model to data from an outbreak. Based on [12], for this particular model, we considered Least Squares Parameter Estimation method and ode45 routine in Matlab. Using these optimization techniques, were obtained $\beta = 20.4843$ and $\gamma = 19.7947$ for optimal parameters. This information allows to conclude that $\mathcal{R}_0 > 1$, which represents an epidemic related to the success of Dove Real Beauty Sketches. In addition, in case of the incorporation of control variables into the mathematical model, parameters estimation assumes a key role in order to understand, in mathematical terms, not only the real impact of an advertisement, but also the costs associated to its launch. In the appendix of [13] it is possible to find the sensitivity equations for the SIR model, which constitute an important indicator to study the effect of β and γ variation. In this context, we performed some simulations to assess and analyze the parameters variation (Fig. 1).

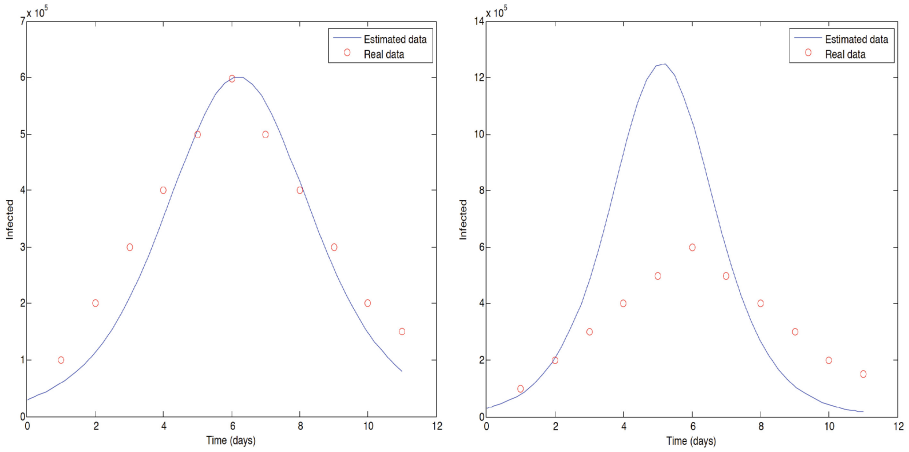


Fig. 1. $\beta = 20.4843$, $\gamma = 19.7947$ (left) and $\beta = 20.1843$, $\gamma = 19.1947$ (right). The effect of β and γ variation in the adjustment to the campaign real data.

Figure 1 (left) displays the estimated data with optimal parameters. Figure 1 (right) presents the estimated data with a smooth variation in β and γ. According to Fig. 1 (left), the optimal values obtained show an almost perfect fitting to campaign real data and it becomes clear that a slight variation on the optimal parameters leads to a misalignment between the estimated and real data (Fig. 1 (right)). So, we can characterize the estimation of β and γ as an ill-conditioned problem. At this point, some guidelines can be given to marketers. With a strong

applicability, the Least Squares Parameter Estimation method leads to accurate results, minimizing the error related to the parameters adjustment to model data. Hence, as well as others, it can be used to manage more efficiently the model dynamics related to future viral campaigns. In this regard, combining with prediction models, a more controlled model dynamic, resulting from an accurate parameters estimation, suggests less costs to the marketers - since the expected margin of error should be residual.

4 Stability and Dynamics of the Model

This section performs a stability and dynamical analysis of the model, by linking it to the Dove Real Beauty Sketches marketing campaign, based on Mathematical Epidemiology and Dynamical Systems results. Henceforth, as a starting point to analyze and discuss mathematical properties related to stability and dynamics of the model, it will be considered the system (4) with its initial conditions, for all $\left(\widehat{S}, \widehat{I}\right) \in [0,1]^2$. Note that the system of ordinary differential equations is autonomous, since $\widehat{S}$ and $\widehat{I}$ do not depend explicitly on time t.

Let $f : [0,1]^2 \to \mathbb{R}^2$ be a C^1 map. It is possible to write (4) in the matrix form

$$\frac{dx}{dt} = f(x(t)),\ \forall t \in [0, +\infty)\ ,$$

where

$$x(t) = \begin{bmatrix} \widehat{S} \\ \widehat{I} \end{bmatrix} \text{ and } f(x(t)) = \begin{bmatrix} -\beta \widehat{S}(t)\widehat{I}(t) \\ \beta \widehat{S}(t)\widehat{I}(t) - \gamma \widehat{I}(t) \end{bmatrix}\ .$$

By solving $f(x(t)) = 0_{\mathbb{R}^2}$ it is concluded that all the points in the form $(S^*, 0)$, where S^* can be any number on $[0,1]$, are equilibrium solutions of the system. However, special focus is devoted to two disease-free equilibrium points: $E_1 = (1,0)$, when the levels of infection are null, and $E_2 = (0,0)$, related to the end of the epidemic, when $\widehat{R} = 1$. In analyzing equilibrium solutions it is pertinent to obtain the linearized matrix of the system (4), in a neighborhood of a critical point (S^*, I^*), for which $I^* = 0$.

$$\begin{aligned} J_f(S^*,0) &= \left[\begin{array}{cc} \dfrac{\partial\left(-\beta \widehat{S}(t)\widehat{I}(t)\right)}{\partial \widehat{S}} & \dfrac{\partial\left(-\beta \widehat{S}(t)\widehat{I}(t)\right)}{\partial \widehat{I}} \\ \dfrac{\partial\left(\beta \widehat{S}(t)\widehat{I}(t) - \gamma \widehat{I}(t)\right)}{\partial \widehat{S}} & \dfrac{\partial\left(\beta \widehat{S}(t)\widehat{I}(t) - \gamma \widehat{I}(t)\right)}{\partial \widehat{I}} \end{array} \right]\Bigg|_{(S^*,0)} \\ &= \begin{bmatrix} 0 & -\beta S^* \\ 0 & \beta S^* - \gamma \end{bmatrix}\ . \end{aligned} \tag{5}$$

Considering the characteristic equation $\det(J_f(S^*,0) - \lambda I_2) = 0$, the eigenvalues of $J_f(S^*,0)$ are $\lambda_1 = 0$ and $\lambda_2 = \beta S^* - \gamma$. In general, by evaluating the eigenvalues at the equilibrium solutions, we can infer its stability. However, for E_2, linear stability analysis may not be sufficient. In this particular case are necessary more advanced mathematical techniques to study the dynamics of the system around E_2. Firstly, considering the optimal parameters obtained in the previous section and the disease-free equilibrium solutions it follows, for E_1 and E_2, respectively, that

$$\lambda_2 = \beta - \gamma > 0 = \lambda_1 \text{ and } \lambda_2 = -\gamma < 0 = \lambda_1 \, .$$

In relation to E_1, since $\Re(\lambda_2) > 0$, according to basic stability properties of linear systems, E_1 is an unstable solution. Establishing a relationship with marketing, this solution corresponds to the case that the number of individuals who recommend the marketing campaign increases faster than it stops being shared (viral marketing epidemic). Considering the equilibrium E_2, is not possible to conclude, from the linear stability analysis by itself, whether or not the equilibrium solution E_2 is stable. In this regard, the behavior of the autonomous systems as well as its equilibrium solutions can be described through a phase-space. Thus, in a two-dimensional space, the orbits of the system (4) can be described by the equation

$$\frac{d\widehat{I}}{d\widehat{S}} = \frac{\beta\widehat{S}\widehat{I} - \gamma\widehat{I}}{-\beta\widehat{S}\widehat{I}} = -1 + \frac{1}{\mathcal{R}_0\widehat{S}} \Leftrightarrow \widehat{I}(\widehat{S}) = -\widehat{S} + \frac{1}{\mathcal{R}_0}\log(\widehat{S}) + C, \; C \in \mathbb{R} \, . \quad (6)$$

This expression enables marketers to predict the final state of the epidemic, from a specific initial state. In this context, through $\dfrac{d\widehat{I}}{dt} = 0$ we can calculate the maximum number of infections, at a time t. This number is obtained when $\widehat{S} = \dfrac{1}{\mathcal{R}_0}$. Thus, by substituting $\widehat{S} = \dfrac{1}{\mathcal{R}_0}$ and $\widehat{I} = \widehat{I}_{\max}$, the maximum number of infections in time is given by

$$\widehat{I}_{\max} = \widehat{I}(0) + \widehat{S}(0) + \frac{1}{\mathcal{R}_0}\left(\log\left(\frac{1}{\mathcal{R}_0}\right) - \log\left(\widehat{S}(0)\right) - 1\right) \, .$$

Generated by (6), Fig. 2 shows one orbit of the SIR system and traduces the impact, in time, of the Dove Real Beauty Sketches marketing campaign.

Following an assertive analysis of the Fig. 2, $\widehat{S} = 1$ represents the launching of the marketing campaign, where almost nobody knows about it ($\widehat{I} \approx 0$). Over time, individuals begin to spread the marketing campaign through network effects and the number of people who have contact with the marketing message reaches a maximum level (viral epidemic), that is

$$\lim_{\widehat{S} \to \frac{1}{\mathcal{R}_0}^+} \widehat{I}(\widehat{S}) = \widehat{I}_{\max} \, .$$

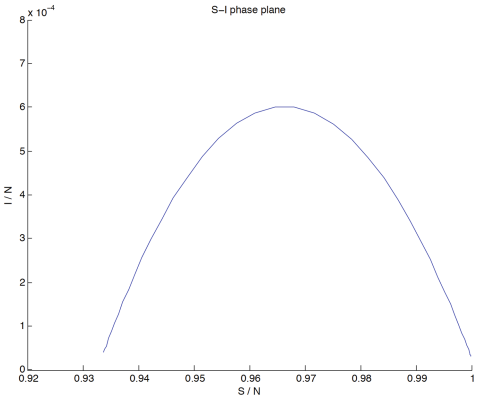


Fig. 2. Orbit of the nonlinear SIR model.

Thus, when a large portion of people have already contact with the campaign, the number of infected people tends to decay to zero, and the number of people who can have contact with the marketing message tends to stabilize around $\widehat{S} \approx 0.93$. It should be noted that the greater the sharing of the message, the greater the number of infected individuals. In the figure below, are illustrated the mathematical results previously analyzed.

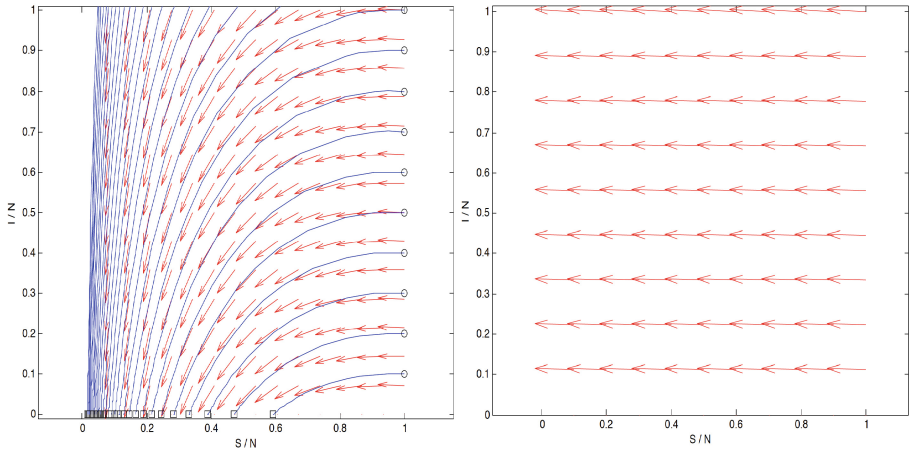


Fig. 3. Dynamics of the nonlinear SIR model (left) and linearized model at E_1 (right).

Figure 3 (left) refers to the nonlinear model (4) and, for each trajectory considered, the initial and final point are marked with a *circle* and a *square*, respectively. Figure 3 (right) displays the vector field of the linearized model, generated by (5) at the equilibrium E_1. From the phase-plane analysis we infer that E_2

is stable, since all trajectories starting in an arbitrary $a \in B(E_2, \delta)$, for $\delta > 0$, move about a finite range of distance, converging to an ending point $b \in \mathbb{R}^2_+$ nearby E_2. In other words, the solutions contract back in the direction of E_2. In what concerns to the vector field showed in Fig. 3 (left), the vectors behavior is perfectly expectable due to the fact that, by the first equation of (4), $\frac{d\widehat{S}}{dt} < 0$, since $\widehat{S}, \widehat{I} > 0$. At this point, $\widehat{S} = \frac{1}{\mathcal{R}_0}$ is crucial to understand the dynamics of the system. For starting points $(a_0, b_0) \in \mathbb{R}^2_+$ in a neighborhood of $(S^*, 0)$, for which $S^* \ll \frac{1}{\mathcal{R}_0}$, the trajectories tend to return to a nearby solution $(S_f, 0) \in B(E_2, \delta)$, for $\delta > 0$. Relating to the marketing context, this means that if a small group of persons have contact with the marketing campaign, then it will not be widespread (no viral epidemic).

Concerning to E_1, as calculated previously, this solution is unstable, since the trajectories grow away from the equilibrium, and computing simulations confirm this findings (Fig. 3 (left)). In other words, supposing that starting points $(a_0, b_0) \in \mathbb{R}^2_+$ are nearby $(S^*, 0)$, for which $S^* > \frac{1}{\mathcal{R}_0}$, the trajectories tend to diverge from the equilibrium, causing a viral epidemic. Hence, these equilibrium solutions are unstable. In terms of marketing, this means that the number of infected individuals - who recommend the campaign - tends to, gradually, grow up, until the maximum level of infection be reached. Then, over time, this number tends to fall down to nearby zero. According to Fig. 3 (right), the instability of E_1 is even more evident, through the vector field, which corroborates that Dove Real Beauty Sketches marketing campaign was a viral epidemic.

5 Conclusions

In this paper, we studied the impact of Dynamical Systems and Mathematical Epidemiology on the dynamics of a specific viral campaign. We considered a SIR epidemiological model and performed a parameters estimation. Some numerical simulations showed not only that Dove Real Beauty marketing campaign was a viral success in terms of marketing strategy, but also that mathematical analysis, applied to viral campaigns, can provide a key to differentiation on the marketing market. We also studied the stability of the two disease-free equilibrium solutions related to the SIR epidemic model. To the marketing business, this model creates an interesting application since it is a cheap and efficient way to measure the stochastic behavior of a marketing campaign. However, some drawbacks can be pointed out to the model: the parameters estimation is not a trivial task, due to the challenge in predict the behavior and dynamics of a viral campaign, and a slight variation on the input data can lead to huge variations on the output.

All in all, the use of Dynamical Systems and Mathematical Epidemiology theory has a very positive impact on the orientation and analysis of successful and general viral marketing campaigns.

Acknowledgments. This work was supported in part by the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), through CIDMA - Center for Research and Development in Mathematics and Applications, within project UID/MAT/04106/2013; and through Algoritmi R&D Center, under COMPETE: POCI-01-0145-FEDER-007043 within the Project Scope: UID/CEC/00319/2013.

References

1. Long, C., Wong, R.C.W.: Viral marketing for dedicated customers. Inf. Syst. **46**, 1–23 (2014)
2. Mochalova, A., Nanopoulos, A.: A targeted approach to viral marketing. Electron. Commer. Res. Appl. **13**(4), 283–294 (2014)
3. Woerndl, M., Papagiannidis, S., Bourlakis, M., Li, F.: Internet-induced marketing techniques: critical factors in viral marketing campaigns. Int. J. Bus. Sci. Appl. Manage. **3**(1), 33–45 (2008)
4. Haryani, S., Motwani, B.: Discriminant model for online viral marketing influencing consumers behavioural intention. Pac. Sci. Rev. B: Humanit. Soc. Sci. **1**(1), 49–56 (2015)
5. Rodrigues, H.S., Fonseca, M.J.: Can information be spread as a virus? Viral marketing as epidemiological model. Math. Methods Appl. Sci. **39**(16), 4780–4786 (2016)
6. Gardner, J., Sohn, K., Seo, J., Weaver, J.: A sensitivity analysis of an epidemiological model of viral marketing: when viral marketing efforts fall flat. J. Mark. Dev. Competit. **7**(4), 25–46 (2013)
7. Breen, C.: "Real Beauty" Campaigns in the Media: Dove Campaign for Real Beauty. http://www.writing.ucsb.edu/sites/secure.lsit.ucsb.edu.writ.d7/files/sitefiles/news/2015%20Breen.pdf. Accessed 28 Sept 2016
8. Stampler, L.: How Dove's 'Real Beauty Sketches' Became The Most Viral Video Ad Of All Time. http://static4.businessinsider.com/image/519cd1336bb3f76344000015-960/screen%20shot%202013-05-22%20at%209.48.32%20am.png. Accessed 28 Sept 2016
9. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Trans. Web (TWEB) **1**(1), 1–39 (2007)
10. Schroeder, S.: YouTube Now Has One Billion Monthly Users. http://mashable.com/2013/03/21/youtube-one-billion/#Wv1EUzqZGkqx. Accessed 28 Sept 2016
11. Britton, T.: Stochastic epidemic models: A survey. Math. Biosci. **225**(1), 24–35 (2010)
12. Lloyd, A.: Sensitivity Analysis and Least Squares Parameter Estimation for an Epidemic Model. http://www4.ncsu.edu/~allloyd/RTG/RTG_Workshop_Parameter_Estimation.pdf. Accessed 28 Sept 2016
13. Capaldi, A., Behrend, S., Berman, B., Smith, J., Wright, J., Lloyd, A.: Parameter estimation and uncertainty quantification for an epidemic model. Math. Biosci. Eng. **9**(3), 553–576 (2012)

Developing an Ambient Intelligent-Based Decision Support System for Production and Control Planning

Marco Gomes(✉), Fábio Silva, Filipa Ferraz, António Silva, Cesar Analide, and Paulo Novais

Department of Informatics, University of Minho, Braga, Portugal
{marcogomes,fabiosilva,analide,pjon}@di.uminho.pt,
filipatferraz@gmail.com, antoniosilva9116@gmail.com

Abstract. Scheduling production instructions in a manufacturing facility is key to assure a efficient process that assures the desired product quantities are produced in time, with quality and with the right resources. An efficient production avoids the creation of downstream delays, and early completion which both can be detrimental if storage space is limited and contracted quantities are important. Therefore, the production, planning and control of manufacturing is increasingly more difficult as family products increases. This paper presents an ongoing Ambient Intelligent decision support system development that aims to provide assistance on the creation on standard work procedures that assure production quantity and efficiency by means of ambient intelligence, optimization heuristics and machine learning in the context of a large organization.

Keywords: Ambient intelligence · Industry 4.0 · Standard work decision support system

1 Introduction

This project fits in a strategic partnership between our a research laboratory and a large Portuguese manufacture factory, which the main business is the development and production of infotainment and instrumentation systems for the automotive industry. These synergies were promoted to answer the challenge of improving the existent methodologies in order to increase the competitiveness, efficiency and productivity through systematic innovation.

Involved in such initiative, the work portrayed in this document outlines the phase of requirements specification, adaptation of ambient intelligence in the context of industry 4.0 and elaboration of intelligent software architectures. The main objective is to answer each requirement from the stakeholder of the

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_97

project and propose intelligent services that can be used to ease tasks and procedures. Included in these intelligent services are the development of heuristics for optimization, pattern discovery through the analysis of inputs and process optimization by process mining.

The main requirements requested deal with the improvement of the elaboration of Working Instructions (WI) and the planning of Standard Work (SW), accordingly to the Lean Thinking principles and in the context of the Industry 4.0 framework. Work Instructions (WI) and Standard Work Sheets (SWS) are tools that support production and play key roles in the engineering process of the organization, impacting performance and the meeting of customer's requirements.

WI provides information about the sequence of operations needed to perform a task correctly by an operators in a workstation. It is provided to operator when its creation is completed and aid the operator with instruction for the manufacturing process. SW is made from the planning of a set of WI that compose a manufacturing line and allow effective line balancing, allocating tasks to operators which are developed in or across workstations. The main problem in SW is task allocation by each operator. The non-existence of decision support system to aid this planning motivates the development of this project which aims to help analyse target quantities, number of operators and time available for production while reducing errors, alerting users and promoting continuous improvement.

The target company aims to have access to real-time updated information, integrated with other information systems already in use, to support the decision-making process and automatically aid the generation of WI and SW according to a business process.

This article will be organised in four sections to introduce describe the main phases of the work developed. In the production and control processes section a brief overview of existing state of the art contextualized with the target company is made, as well as, an overview of the solution provided by this research. In an ambient intelligence-based decision support system to enhance production and control planning a detailed description of the research results is done, outlining the system architecture, decision support capabilities and the influence and adequacy of ambient intelligence towards this problem. The discussion of results is made in a separate section prior to the conclusion.

2 The Production and Control Processes

The production and control processes in large companies demand high complexity, for which technology may help increase performance and competitiveness. One example is the standardization of the supply chain. It is essential to have well defined procedures, trained workers and available resources. This a requirement to work properly, as well as, improve on business processes, increase and adjust the capacity required by the customer and decrease wastes, costs and useless resources [4]. Therefore, the manufacturing process is a crucial phase in the supply chain because it assures that the operations and tasks are performed

in controlled and standardized terms, adding to this process a higher value and guaranteeing the same consist quality across manufactured objects [6].

Due to the high complexity of the processes, companies try to standardize every process in its supply chain, in order to achieve a higher performance and competitiveness. It is crucial to have well-defined procedures, so that the processes can be improved.

As presented in the previous section, this work pretend to standardize the fabrication and control processes, using a lean line tool, like Standardized Work, that consists in three components:

- Takt Time - is the demand rate and serves as a reference to set the pace of production;
- Work Flow - the sequence in which an operator performs tasks within takt time;
- Standard Inventory - includes units and machines, required to keep the process operating smoothly.

The Standardized Work allows an effective line balancing, allocating tasks to operators and workstations [3]. It is possible to establish an optimized standard work sequence taking into account several constrains, such as the Takt Time, operation time and movements. Additionally, the Standard Work provides relevant information for the manufacturing process that helps the decision-making process.

2.1 Modelling and Mining the Process

Process optimization is a beneficial task that most organizations do in order to achieve operational gains and increase the efficiency of their processes. The area for process optimization in the field of information Technologies (IT) may be called process mining [2]. The optimization is made through structures such as Work-flow Management (WM) Systems which register the start and end of activities [1]. Each organization often has log of operations often referred as event log, history or audit trail. A process instance is referred to the task being processed such creation of Work Instruction (WI) or Standard Work (SW) computation. Within each process, there are activities performed and associated with the process instance in these logs. From activities and process logs it is possible to represent process work-flows in graph structures assessing dependencies and where possible. Special analysis on the process is made through the use process mining algorithms in order to optimize each work-flow. A work-flow engine, has the task to automate the tasks within the work-flow creating automations for each task or waiting on input to conclude other in order to manage an entire work-flow. These engines suffer limitations due to activities performed that are unaware of the process, therefore they may not be involved in the management and planning of the process by themselves. Systems where each task is aware of its process are called Business Process Management Systems (BPMS) or Process Management Systems (PMS). Either terms define systems that rely

on process models to control process work flow. Work-flows contain ordered activities describing their casual dependencies, temporal properties and creation and use of data.

In order to better understand the whole process, it was carried out a process analysis supported by BPMS aimed to map the process and identifying the problems and non-standardized procedures. Through the analysis of the outcome was possible to identify six main problems that result from the use of non-standard tools:

- Low productivity;
- Low efficiency;
- High time consumption;
- High resources consumption;
- High errors occurrence;
- Orders delays.

Regarding to the main causes that lead to the effects mentioned, they ranged from human errors (due the lack of training and lack of information of decision-making process), software integration variability, to Work Instruction and SW procedures that require a lot of manual operations and are supported by several non-standard tasks. Moreover, it was pointed out that the process of creation of a Standardized Work usually includes inputs from different departments, under different formats, making this process harder to streamline. Regarding the users, it is emphasized the SW creator and the team leader line, who are responsible to check if the production process occurs accordingly to the predicted SW in the assembly line. From the analysis of the requirements, we concluded that to manage and create a Standardized Worksheet, it is required inputs from several departments and systems:

- the precedence table of the workstations from the testing process (an Excel file);
- an assembly line layout in order to figure out the operators movements (a CAD file);
- some data, like the percentage of machine usage and minutes per shift from the company Intranet (an Excel file);
- a set of tasks from the Work Instructions (an Excel file);
- MTM measurements (a PDF file);
- the product to be produced quantities;
- machine times to calculate the time per shift (an Excel file).

About the SW management process, it involves its creation, its update, its deletion and search possibilities. Eventually, it is important to notify the team leader line that a new SW is ready, as well as receive any comment based on the SW consulted by the team leader line that might change it. The manufacturing and control processes are bounded to use the Standardized Work, since it brings optimization of the procedures, as well as services with higher quality and productivity [5].

2.2 The Solution Overview

Presently, the processes of WI and SW creation are supported by Excel files, which require a lot of manual operations and time consumption. Additionally, these are paper-based processes that lead to order delays, low efficiency and productivity. Moreover, changes in WI and SW cannot be visualized in real-time which implies delays and potential errors updating a large database of files, transmitting files across departments or consulting different versions of the same WI and SW. This situation can be improved with an integrated and intelligent Decision Support System (DSS) to enhance productivity, planning efficiency and response capability to an high variability context and uncertain demand by reducing time, resources consumption and errors events. In this sense, there can be stated key purposes that the proposed system addresses:

- Process automation: the procedures must be supported by an system that minimizes the manual data insertion and the human-error;
- Information real-time update: the information flow must be automatic and provided by the system, in order to ensure the data accuracy;
- WI visualization in displays: paper-based process will be avoided and the WI will be visualized through displays that provide relevant information about the assembly process using animations for a better understanding;
- Support in the decision-making process: During the SW development, there are some constraints that must be fulfilled. The system should be able not only to support the decision-making process, but also through the optimization and the choosing of the best scenarios;
- Provide analytical support: having the access to the manipulation of a time-series of internal historical company data, real-time operational data and sometimes external data the analytical model of the system will provide a highest level of decision support linked to analysis of large collections of historical data and real-time data.
- Applying Ambient Intelligence approaches: a set of AI methods will be available for reasoning under uncertainty, to support complex decision analysis regarding core planning procedures and also learning, in order to adapt in a dynamic environment.

The analysis of both current business process and the functional and non-functional requirements resulted in the conceptualization of use cases for each scenario, WI and SW and potential clients of the Decision Support System (DSS).

In the case of WI creation, update and availability the use case diagram in Fig. 1 entails the list of operations required, as well as, datasources and client permissions.

Using the same strategy, the creation, update and availability of SW, in Fig. 2 details client permissions, datasources and the list of operations required.

Bearing in mind that the current state is characterized by the lack of integration with other IT systems, the proposed system was designed after a detailed

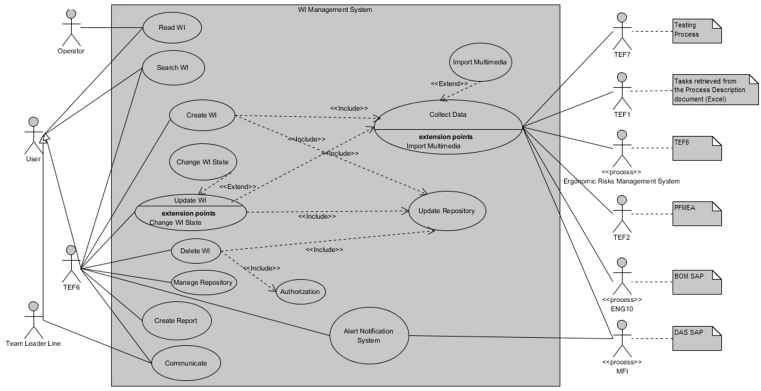


Fig. 1. Work instructions use case

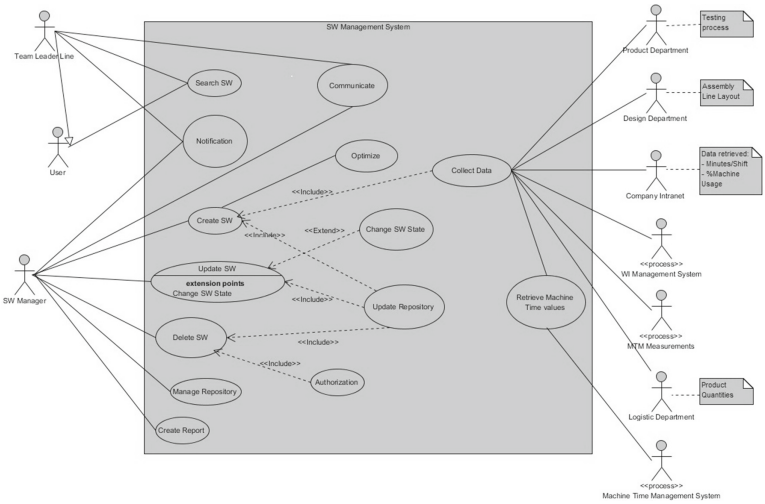


Fig. 2. Standardized work use case

analysis of functional and non-functional requirements captured by direct interviews with stakeholders with the objective of circumventing this issue. Moreover, it was also taken into account the business objectives related to the need of improving production outcomes and the decision-making performance by implementing production optimization.

Therefore, an intelligent Decision Support System is proposed to answer these issues and support managers making timely and optimal decisions. These systems are designed specifically to facilitate decision processes. They support rather than automate decision making and are able to respond quickly to changing needs of decision makers. Traditionally, academics and practitioners have discussed building DSS regarding four major components – (1) the user

interface, (2) the database, (3) the models and analytical tools, and (4) the DSS architecture and network. This traditional list of components remains useful because it identifies similarities and differences between categories or types of DSS and it can help managers and analysts build new DSS. In the following section, we depict our solution using this approach.

3 An Ambient Intelligence-Based Decision Support System to Enhance Production and Control Planning

After the process of requirements specification, we designed an intelligent system comprising several building blocks, including: architecture and networking design, refer to how hardware is organized, how software and data are distributed in the system, and how components of the system are integrated and connected; the model-driven module, emphasize access to and manipulation of a multi-criteria model to provide decision support; and a ambient intelligence module, use artificial intelligence methods to produce relevant knowledge, avoid error and create intelligent services through software agents.

The objective of the system is to change the actual process, which includes automate parts of the process with information systems, optimize and improve efficiency in the processes, as well as reduce errors. Moreover, it is intended that the system should not only optimize the standard work and work instructions by providing dynamically allocation of tasks, but also provide a tool that assist and support the control of production processes.

3.1 The System's Architecture

It is intended that the application, not only should optimize the work instructions in the assembly lines and the standard work, providing dynamically the allocation of tasks, but also should provide a tool that help and support the control and production processes. To achieve these goals, was created the architecture depicted in Fig. 3. It consists of a Core Server that is composed by four components: *Database Handler*, *Authentication Agent*, *Standard Worksheet Handler* and *Standard Worksheet Engine*. The *Database Handler* is responsible to manage the access to the database, containing the constraints of production and control planning, the product information, the work instructions, the table of precedence (which gives the flow between the stations), the tasks to be performed by the operators, etc. On the other side, the *Standard Worksheet Handler* manages the access to standard worksheets information in Production and Control Planning Repository, and the *Authentication* component is responsible for authenticating the user into the system. The *Standard Worksheet Engine* interprets all the control and planning constraints, product information, the work instructions and provide standard worksheets alternatives. This component consists in a set of rules and optimization models that can constrain possible solutions, as well as alternative solutions and methods for evaluating them. These standard worksheets alternatives can be access through RESTful web services in

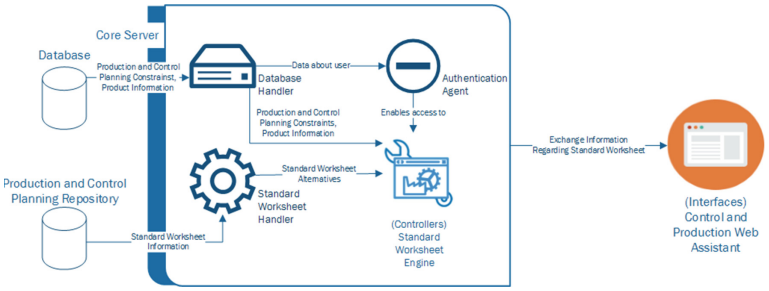


Fig. 3. Architecture of the production and control planning system.

order to ensure they can be easily integrated into any type of application. The Core logic is implemented with web languages in a client server model where information is stored in a central server that provides the information to thin clients through web pages.

The *Control and Production Web Assistant*, which uses the web services available in the Core Server, will be develop as a web application following the Model-View-Control (MVC) pattern. This assistant provides the standard worksheets alternatives for the user to evaluate them. In order to allow the users interaction, this application offers the possibility of changing any attribute of the standard worksheet and the possibility of recalculate the standard worksheet to get the viable solution. The application also will provide a message box to allow and facilitate the communication between the different participants in this process. These alerts are shown as side messages and are gathered in a notification stack.

3.2 Model-Driven Representation

Model (a mathematical representation that relates variables) and analytical models are a dominant component in an intelligent Decision Support-System. When a model is need to understand a situation, then is needed to deliver a representation to the users of the system. In many industrial engineering applications the final decision is based on the evaluation of a number of alternatives in terms of a number of criteria. This problem may become a very difficult one when the criteria are expressed in different units or the pertinent data are difficult to be quantified. The Weighted Sum Model is an effective approach in dealing with this kind of decision problems. Moreover, it is a simple model used to identify and choose alternatives based on the values and preferences of the decision maker. It combines the different objectives and weights corresponding to

those objectives to create a single score for each alternative to make them comparable. On the other hand, using a mixed integer linear programming model for the production line balancing and optimization, which will allow to reduce the resources used, the errors introduced by the manual operations, increases automation and human independence process.

We purpose, firstly, to develop a simple mathematical model to support the decision-making regarding the manufacturing process. The objective is rather expose the data patterns and relationships to have a automated mean to extract information and knowledge that is useful for decision support. Also this strategy will allow asynchronous processing of the available inputs, provide greater flexibility in modelling, and allow adaptive channel weighting between different conditions [7]. Secondly, to combine the previous model (that provides a basic functionality) with the incorporation of optimization models to analyse complex data. This approach is oriented towards finding the best decision alternative to generate optimal solutions that are consistent with time and resource constraints. Furthermore, we intent to develop a catalogue of different classes of models that seems useful to solve a specific class of problems like a scheduling problem or a combinatorial search problem. The intent is to develop a catalogue of models including all the models that allow an effective line balancing, allocating tasks to operators and workstations and yield relevant information for decision-making in the manufacturing process.

3.3 The Ambient Intelligence Approach

An intelligent environment acts on behalf of its users anticipating needs and easing tasks. Developing projects with this mentality require inputs from environment sensors, detection of tasks and actuators. In case of an industrial environments, this vision is kept, though with slight modifications to adequate the norms to a professional space.

An abstract notion of sensors and actuators can be used to include software agents as both sensors and actuators in a information systems approach. These agents obtain relevant data to each process and deliberate actions based on heuristics from artificial intelligence, expert systems and machine learning autonomously on behalf of human collaborators in hope to provide meaningful suggestions, action or plans of action.

In the case of the work proposed in this article, current data is compared to historical data to discern how a process should be run. Analytics provide recommendations, enhancements and warnings to support decision-making. This results in effective, productive, environmentally friendly and cost-effective production.

Experience from past known approaches are used as input in a machine learning algorithm, the ANN, which is trained to learn patterns on the inputs and relate them with known outcomes. After the learning phase the ANN is able to provide answers to inputs on a online system which seeks optimizations and act faster than traditional heuristics which can take, hours, days or weeks to provide the desirable results. This system combines the intelligent environment

approach by capturing initial data from the collaborator traditional work, and on later stages agents propose working solutions to the collaborator anticipating his work.

4 Discussion

Results in this approach come from the application of a requirement analysis with interviews to the stakeholders of the project and the detailed mapping of requirements and their adoption to the ambient intelligence patterns increasing efficiency and value to the business process. This business process transformation is inspired in the industry 4.0 moto and intelligent manufacturing.

Intelligent services are also provided by software agents that accompany traditional business tasks such the planning of the standard work and provide suggestion based on previous approaches and their results. Such is the case with the ANN detailed in this proposal. More services shall be detailed in future iterations of this project but this first one lays the ground work to acquire a new mentality supported by the concepts of ambient intelligence.

The design used and sensor agent may be useful towards other tasks inside process mining, such as finding recurrent patterns in the standard work design which can evidence strategies to speed up optimization tasks. For instance, discovering that a certain layout in similar standard work provides local optimal results may be used on new similar designs. This similarity may be again calculated by software agents and detailed to the users later.

We aim to provide a proof-of concept application intelligent services and ambient intelligence patterns at the disposal of business processes and how they can be related. This approach successfully demonstrates this strategy and outlines its potential gains in terms of work automation and increase in efficiency. All this then be translated into added value to the manufacturing process which is the desire of the stakeholders.

5 Conclusion

This work details the transformation of a business process, SW and WI planning, inside a manufacturing company. Our approach uses the ambient intelligence, machine learning and multi-agent systems to manage the process flow and en results. It was proposed an ongoing project to design and develop an intelligent decision support system for users to make effective decisions regarding the process of Work Instructions and Standard Work procedures. A vision of the desired solution, as well as, different approaches were provided in their early stages and seeking potential modifications.

The future research should continue to explore the potential improvement of the decision-support system in terms of including non-quantifiable criteria, the meaningful information from additional ambient intelligent services and the ways to link and relate the historical data to new data gathered from a wide range of sources with automatic data feeds. Other opportunities include using

the ambient intelligent approach to make autonomous notification of collaborators of new solutions on prolonged tasks such as the ones derived from linear optimizations. In this light, agent may also derive a working solution through machine learning and autonomously search optimal ones using the traditional algorithms, or evolutionary algorithms. In the event of obtaining a better solution it may then autonomously update the plan with the new found solution or alternatively notify the collaborator of the new solution available.

Acknowledgements. This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within the Project Scope UID/CEC/00319/2013. This research is also sponsored by the Portugal Incentive System for Research and Technological Development. Project in co-promotion no 002814/2015 (iFACTORY 2015-2018).

References

1. van der Aalst, W.M.P., Reijers, H.A., Weijters, A.J.M.M., van Dongen, B.F., Alves de Medeiros, A.K., Song, M., Verbeek, H.M.W.: Business process mining: an industrial application. Inf. Syst. **32**(5), 713–732 (2007). http://dx.doi.org/10.1016/j.is.2006.05.003
2. Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes, 1st edn. Springer Publishing Company Incorporated, Heidelberg (2011)
3. Emiliani, M.: Standardized work for executive leadership. Leadersh. Organ. Dev. J. **29**(1), 24–46 (2008)
4. Productivity Press Development Team: Standard Work for the Shopfloor (2002)
5. Ungan, M.C.: Standardization through process documentation. Bus. Process Manag. J. **12**(2), 135–148 (2006)
6. Williams, B.A.: Standard work-lean tools and techniques. SAE Technical Paper (2001)
7. Wu, L., Oviatt, S.L., Cohen, P.R.: Multimodal integration - a statistical view. IEEE Trans. Multimed. **1**, 334–341 (1999)

CobWeb Multidimensional Model: Visualizing OLAP Query Results Using Tag-Cloud Operators

Omar Khrouf[1(✉)], Kaïs Khrouf[1], and Jamel Feki[2]

[1] MIR@CL Laboratory, University of Sfax, Sfax, Tunisia
Omar.khrouf@yahoo.fr, Khrouf.Kais@isecs.rnu.tn
[2] FCIT, IS Department, University of Jeddah, Jeddah, Saudi Arabia
jfeki@uj.edu.sa

Abstract. OLAP (On-Line Analytical Processing) systems provide decision makers with multidimensional analyses of large databases by generating an aggregated vision of data. Nowadays, these systems face growing non-numeric data. In this context, we propose in this paper a new generic multidimensional model called *CobWeb* dedicated to the OLAP of XML documents; it is based on the concept of facet. The *CobWeb* model aims to ease the expression of queries; also it offers an appropriate vision of the document warehouse. In this context and for manipulation purposes, we propose new visualization operators for OLAP query results by using the concept of Tag clouds as a means to help decision-makers to see the content in an efficient manner and then to focus on the knowledge in results.

Keywords: XML documents · OLAP · *Cobweb* multidimensional model · Tag clouds

1 Introduction

Nowadays, documents contain pertinent data for the decision-making process and, therefore, their analysis can help decision makers to better understand the business processes of their organizations and take well founded decisions they cannot find by examining a conventional (numerical) data warehouse. However, in practice, most of the textual data are not included in decision-making systems or not considered during the decision-making process; this is due to several reasons as the lack of specific features for document warehouses. To alleviate this problem, it became necessary to integrate textual data into the decision-making information systems; this will help decision makers to identify hidden aspects and then improve their decisions.

In this context, several efforts have been invested; more accurately, most of them were interested in the exploitation of information stored in documents by using a set of facets [5] in order to model the documents according to several viewpoints issued from users and to improve the On-Line Analytical Processing (OLAP) of documents [11]. In this paper, we focus on the multidimensional modeling of documents and the OLAP analysis of textual data.

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_98

According to our review of the literature works related to document warehousing and OLAP of documents, the existing contributions could be classified into two main categories: (1) Works which adapted the classical multidimensional model (e.g., star, snowflake or constellation schemas) by enriching it with extensions specific for textual processing ([2, 3] for data-centric documents; [6] for document-centric documents); and (2) Works which proposed specific models for the OLAP of documents, such as galaxy model [8] and diamond model [1].

The purpose of our work is to propose a new multidimensional model called *CobWeb*, as an extension of the galaxy model [9] dedicated to the OLAP of documents. The *CobWeb* model is based on the use of the concept of facets, these facets are standard. At the conceptual level a facet could be seen as a viewpoint since it groups a set of data describing of similar documents. At the external level (i.e., query level), a facet is considered as a means of expression for users' requirements. That is why we transform every facet into a multidimensional component of the *CobWeb* document warehouse model; more accurately a facet is transformed into a dimension as a potential axis of analyses. We have ensured that our *CobWeb* model differs from the existing models by a set of significant extensions such as the exclusion constraint between dimensions, the ability to define recursive parameters, duplicated dimension, and correlated dimensions.

From the other hand, and for performing analyses of the contents of documents, we have proposed a set of four operators for the visualization of results of OLAP queries by using the concept of Tag clouds in order to help decision-makers to better interpret the results of their queries.

We have organized this paper as follows. Section 2 presents the related works dealing with the OLAP of documents and highlights their main drawbacks. Section 3 describes the *CobWeb* multidimensional model focusing on its specifics. Then, we present, in Sect. 4, our new OLAP operators for document warehouses. Finally, Sect. 5 is reserved for the conclusion.

2 Related Work

In this section, we focus on the related works dealing with the OLAP of documents. Among these works the authors in [10] proposed a framework for multidimensional analysis of *XML* documents called *XML-OLAP*. They defined a multidimensional expression language over *XML* cubes called *XML-MDX*. They also specify text mining operators for aggregating text constituting the measure data such as summarization, classification, and top keyword extraction.

Other studies suggest using the text mining techniques and information retrieval in order to aggregate textual data. In this sense, [9] developed two aggregation functions for the galaxy model: (1) *AVG_KW* summarizes a set of pseudo-average keywords replacing the initial set with a smaller and a general set; and (2) *TOP_KW* function that returns the k main keywords from a set of keywords. To do so, first the authors calculate the frequencies of terms by using the *Tf-Idf* function, secondly they select the first term among the k-most-frequent terms.

In [6] a textual dimension has a hierarchy that specifies the semantic between the terms of documents; the hierarchy enables a semantic navigation through two operators: *PULL-UP* and *PUSH-DOWN* in order to help users in the multi-semantic-level analysis of textual data. The authors define two aggregation measures adapted to textual data: Terms Frequency and the Inverted Index.

[6, 9] have used analytical measurements based on the statistical method *Tf-Idf*. However, this method does not allow a consideration of the semantics content of documents. In order to integrate this aspect, the authors of [7] propose a new textual analysis measure for *CXT-Cube*, based on an adapted vector space model to represent textual data. In order to calculate the document terms' weights, [7] developed a relevance propagation technique on a concept hierarchy. Also, the authors provide an aggregation operator *Orank* (OLAP rank) that aggregated a set of documents by ranking them in a descending order using a vector space representation.

[1] present new aggregation operators that take into account the specificities of textual data from documents: *List_Concept*: returns a list of the most used concepts from a set of concepts in order to aggregate them into the corresponding cell of the multidimensional table; *G_Concept*: extracts the most used generic concepts; *S_Concept*: extracts the most used specific concepts; *Top_Concept*: groups the operators *List_Concept, G_Concept* and *S_Concept* to display the first concept of each of these operators.

We have learned from this study that most of works related to OLAP of documents proposed aggregate functions to deal with the textual content of documents. The result of these aggregate functions neglects, by eliminating, less frequent words which could be significant for the decision-maker analyses. To alleviate this limit, we propose, in this paper, a new visualization technique based on the Cloud of Tags format for the multidimensional result; this format highlights the most frequent concepts to the decision-maker without eliminating those are less frequent. In addition, the decision-maker will be able to examine less frequent concepts by hiding those are highly sized.

3 CobWeb Multidimensional Model

In order to integrate the concept of *facet* in our multidimensional model, we define a set of five facets namely *Content*, *Structural*, *Metadata*, *Keyword* and *Semantic* [4]. Each facet describes a useful aspect to the exploitation of documents according to a viewpoint in order to allow the user to see documents from multiple views and then to have a more targeted access to information as needed. In our approach, the five proposed facets should be: (1) *Standard*, i.e., independent of any specific domain of application, (2) *Document-Structure Free*, i.e., not limited to a set of predefined documents or to documents of the same structure, (3) Complete, i.e., cover all the information that describe XML documents and which are useful to the decision-maker to satisfy their OLAP needs, and finally (4) Automatically extracted from documents.

Based on these facets, we propose the *CobWeb* multidimensional model dedicated to the OLAP of documents in order to provide more opportunities for the expression of analytic queries and a vision more targeted of the data to decision-makers. To build this

model, the main idea consists to transform every facet into a dimension since these facets may represent a means of expressing the users' viewpoints and therefore, cover the required data for their requirements. Furthermore, we have enriched the model with the *Document* dimension in order to link the information from different facets to their original documents.

In our work, we are inspired from the galaxy model [9] where a dimension could act as a subject of analysis (i.e., fact in the multidimensional terminology). This model connects several compatible dimensions by a node. However, the galaxy model is convenient for a collection of document which must be structurally homogeneous; this represents a restriction to decision-makers when they want to analyze data belonging to different collections. To alleviate this limit, our *CobWeb* model (Fig. 1) allows to group factual and textual data extracted from heterogeneous documents (Structure and content). *CobWeb* differs from the existing models by the following extensions:

- **Correlated Dimensions:** The conventional multidimensional operators (i.e., specific to numeric data warehouses) *Drill Down* and *Roll Up* allow detail/aggregate the results of OLAP analysis; they apply on parameters belonging to a same hierarchy. However, in the OLAP of documents, it would be interesting for example to move from the Content facet to the Concept facet or from Content to Keywords facet. In the classical multidimensional modeling, moving across dimensions is prohibited because of the absence of inter-dimensional relationships. In the *CobWeb*, we propose the concept of correlated dimensions, which allows a same query to move between dimensions (Fig. 1). The shifting from one dimension to another is accepted under condition: when we respect the direction of the dashed arrow linking dimensions. For example, it is possible to move from the *D_Content* dimension to the *D_Semantic* dimension, the reverse is not allowed.

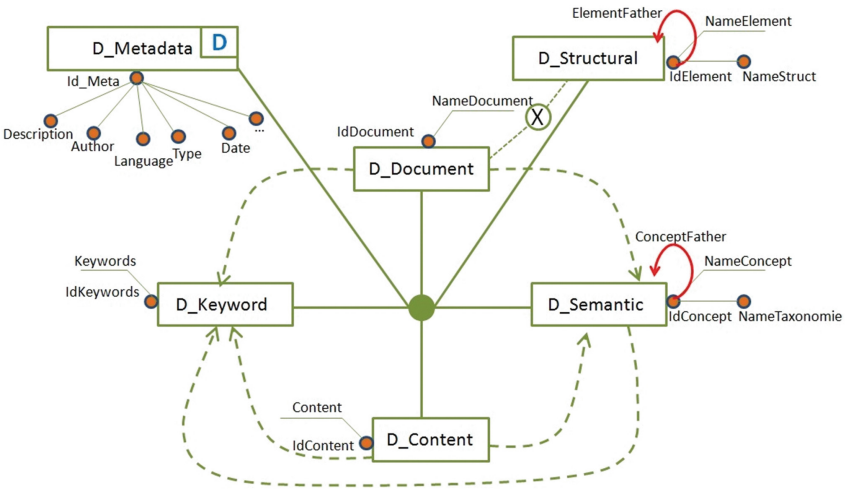


Fig. 1. CobWeb model.

- **Duplicated dimension:** It is a dimension that could be used more than once in the same OLAP analysis (multidimensional query). This type of dimension is rarely used in classical modeling (such as the star model); when used it requires high skilled user. In order to illustrate the duplicated dimension, suppose we want to perform analyses of documents published and sold within the same period (i.e., year for instance). Thus, the Date dimension must be used twice (once as a Publish date and once as a Sale date). Such a request requires a level of expertise raised for the decision-makers, who are generally not computer scientists.

In our works, we have proposed a specific dimension namely *Metadata* characterized by a set of additional parameters that can be used both in the same query, such as Author or Language. Graphically, a duplicated dimension is symbolized by the letter **D**. In the *CobWeb* model, we have only one duplicated dimension, called Metadata (Fig. 1).

- **Recursive Parameter:** In classical data warehouse schemas, parameters and hierarchies of dimensions are known in advance. However, in our work:

– The structure of a document may differ from one collection to another; it is hierarchically organized.
– We determine the semantic structure of documents by projecting the document on taxonomy (a hierarchy of concepts); this helps describing the textual content of documents. The number of concepts and levels varies from one taxonomy to another.

For these two dimensions define a new type of parameter called *recursive parameter*, since the documents and the taxonomy used in our model is represented in a hierarchical manner.

– The structural dimension helps us to move between levels (Content < Section < Subsection < Paragraph) using the conventional OLAP operators namely Roll Up and Drill Down.
– The semantic dimension allows the movement between concepts (e.g., from "Information System" to "Data Warehouse" to "Cube" and reversely).

Graphically, a recursive parameter is shown diagrammatically by a directed loop (Fig. 1).

- **Exclusion Constraint between Dimensions:** The exclusion constraint requires that a couple of dimensions cannot be used simultaneously in the same analysis. In *CobWeb*, the exclusion constraint concerns the Document and the Structural dimensions because an analysis must concern the documents or parts of the documents (title, section, etc.), but not both at the same time. Graphically, this exclusion constraint is denoted by a circle containing the letter **X** connected to the involved dimensions, such as: *D_Document* and *D_Structural* in Fig. 1.

4 New Visualization Operators for OLAP Query Results in Document Warehouses

For the OLAP of documents, the content of the multidimensional tables can be numerical (e.g., the number of documents) or textual (e.g., list of keywords, concepts). When the content of these multidimensional tables become very congested it complicates the task of the decision-maker. In the literature, most of works have proposed aggregation functions (e.g., *Top_Kewyord* [8] or *Top_Concept* [1]) in order to display the N most important keywords or concepts in the documentary collection. In our work, we propose to use the concept of tag clouds in order to represent the result of an OLAP query performed on a document warehouse. The tag cloud helps users (i.e., decision-makers) better see the result of a query and then better understand the result.

To do so, we propose a set of four operators for the visualization of results of OLAP queries as a means to help decision-makers focusing on knowledge incarnated in results. These operators are called *TableCloudTags, Filter_Tag, CellCloudTags* and *Agg_Tag*. They are detailed hereafter.

4.1 *TableCloudTags* Operator

This operator accepts the content of a multidimensional table resulted from an OLAP query on a document warehouse document and displays this content as a single cloud of words where every word has a font size proportional to its frequency in the table.

Definition. $Cube_{TC}$ = *TableCloudTags (Cube)* where:

Input

- *Cube* is the multidimensional table on which the *TableCloudTags* operator is applied,

Output

- $Cube_{TC}$ is the resulting cube composed of a single cell of Clouded tags.

Example. Query Q1: Analyze the keywords by Author and Date. Figure 2 depicts the multidimensional table which results from the execution of this OLAP query as a tag cloud. Note that keywords such as **XML**, **Cube** and **Modeling** have a higher frequency

		Dimension 1: Authors		
		Omar Boussaid	Gilles Zurfluh	Frank Ravat
Dimension 2: Date	Keywords			
	2014	OLAP, Semantic	XML, Data	Modeling
	2013	*	*	Galaxy model



Fig. 2. Result of query Q1

compared to remaining keywords. Therefore, the analyzed documents treat more particularly these concepts.

4.2 *Filter_Tag* Operator

The number of words generated by the tag cloud could be very important for a multidimensional query. For this reason, we propose to filter the tag clouds by:

- **User profile:** The user can specify his profile that describes his interests. In our work, the user profile is composed of a set of topics; each topic is associated with a set of keywords.
- ***Filtering according to a threshold***: To reduce the number of words generated by the tag cloud, the user can set a threshold that represents the percentage of words to be displayed (for example, 30%). In this case, the tag cloud shows only the most significant words based on this percentage.
- ***The specification of values from data of dimensions:*** decision-makers can filter the tag cloud by choosing some values from dimension data of the multidimensional table.

Definition. $Cube_{FT}$ = *Filter_Tag (Cube, Type)* where:

Input

- Cube is the multidimensional table on which the *Filter_Tag* operator is applied,
- The selected type in order to filter the tag cloud

Output

- $Cube_{FT}$ is the resulting cube having the same number of cells as the input cube

Example. Figure 3 shows the tag cloud filtered by the *Author* dimension **(Frank Ravat)** and *Date* dimension **(2013)**.

		Dimension 1: Authors		
		Omar Boussaid	Gilles Zurfluh	Frank Ravat
Dimension 2: Date	Keywords			
	2014	OLAP, Semantic	XML, Data	Modeling
	2013	*	*	Galaxy model



Fig. 3. Example of filtering the tag cloud by data dimension

4.3 *CellCloudTags* Operator

In the multidimensional table, every word is displayed only once in the cell even if it was repeated (where is it repeated) many times. This representation imposes a restriction to decision-makers if they want to know the most analyzed keywords in every cell. To

help, we propose a new OLAP operator: *CellCloudTags* in order to combine the OLAP and the tag cloud in order to represent the result of OLAP querying as a tag cloud in every cell of the multidimensional table. This representation allows us to know the most important data in each cell.

This operator accepts the content of a multidimensional table resulted from an OLAP query on a document warehouse document and displays this content as multiple clouds of words; in fact as many clouds as cells in the table. Naturally, every word in a cell has a font size proportional to its frequency in its cell.

Definition. *Cube_TagC= CellCloudTags (Cube, Cij)* where:

Input

- Cube is the multidimensional table, on which the Coup_Tag operator is applied,
- C_{ij} is the intersection of i^{th} line and the j^{th} column of the Cube

Output

- *Cube_TagC* represent the tag cloud in every cell of the multidimensional table.

Example. Figure 4 shows the result of query Q1 by using the operator *CellCloud-Tags*. We note that **Galaxy** has font size bigger than **XML**; therefore, the author Frank Ravat investigates more in the **Galaxy** than **XML**.

		Dimension 1 : Authors		
		Omar Boussaid	Gilles Zurfluh	Frank Ravat
Dimension 2 : Date	Keywords			
	2014			
	2013	*	*	

Fig. 4. Example of coupling multidimensional table and tag cloud

4.4 *Agg_Tags* Operator

This operator aggregates the textual data represented by our tag cloud in every cell of the multidimensional table in order to help the decision-maker to draw his attention on the most frequent data by dimensions axis. The aggregation function we are using calculates the number of occurrences of every word in each cell of the cube according

to one among the two displayed dimensions. The relevance of each word in a dimension is proportional to its frequency in the set of words displayed in that dimension.

The navigational analysis through OLAP cubes massively uses aggregations through drilling operations such as the Roll Up and the Drill Down operations. In order to help the decision-maker to easily interact with an OLAP system, we have proposed a new OLAP operator: *Agg_Tag* in order to allow the aggregation of textual data in OLAP environments. The proposed aggregation operator has to combines different tag cloud by axis dimensions.

Definition. *Cube_Agg, = Agg_Tag (Cube_TagC, Method)* where:

Input

- *Cube_TagC* is the cube result of the operator *Coup_Tag*
- *Method ∈ {line, Column}* allows the user to aggregate by dimension 1(line) or dimension 2 (Column).

Output

- *Cube_TagC* represent the aggregation tag cloud by axis dimensions.

Example. Figure 5 shows the result of query Q1 by using the operator *Agg_Tag*.

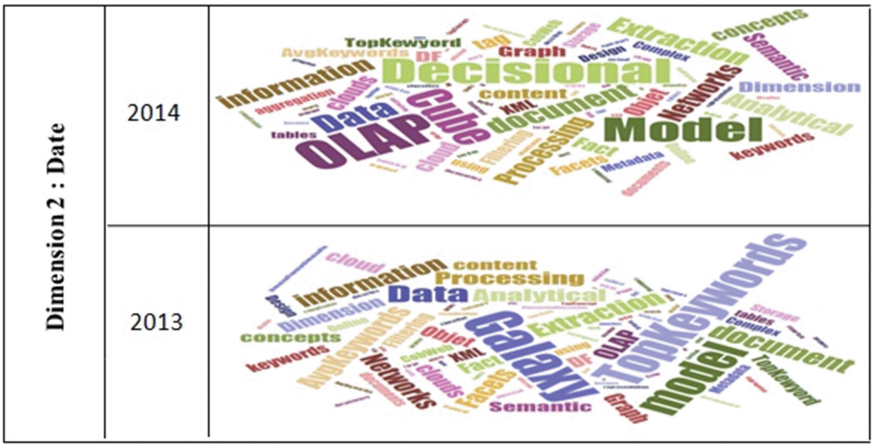


Fig. 5. Aggregation tag cloud by Date dimension

5 Conclusion

In this paper, we presented a framework for a multidimensional OLAP text analysis. We proposed a new generic multidimensional model dedicated to the On-Line Analytical Processing (OLAP) of XML documents, called *CobWeb* model. This model differs from existing models by the following extensions: the exclusion constraint, the recursive parameters, the duplicate dimension and the correlated dimensions. Thus, we proposed

a new visualizing OLAP query results based on tag clouds in order to facilitate the interpretation of the results of the multidimensional analyses on the textual data. In perspective, we plan to conduct experiments to measure the quality of the result extracted by our OLAP operators. Finally, we intend to introduce the collaborative aspect which allows the sharing of OLAP analyses between users working in the same organization.

References

1. Azabou, M., Khrouf, K., Feki, J., Vallès, N., Soulé-Dupuy, C.: Diamond multidimensional model and aggregation operators for document OLAP. In: IEEE Ninth International Conference on Research Challenges in Information Science 2015, Athens, Greece, pp. 363–373 (2015)
2. Feki, J., Ben Messaoud, I., Zurfluh, G.: Building an XML document warehouse. J. Decis. Syst. **22**(2), 122–148 (2013). Taylor & Francis
3. Hachaichi, Y., Feki, J.: An automatic method for the design of multidimensional schemas from object oriented databases. Int. J. Inf. Technol. Decis. Making **12**(2), 1223–1259 (2013)
4. Khrouf, O., Khrouf, K., Altalhi, A., Feki, J.: CobWeb multidimensional model: filtering documents using semantic structures and OLAP. In: ICIW 2015: The Tenth International Conference on Internet and Web Applications and Services, IARIA (2015). ISBN: 978-1-61208-412-1
5. Kumar, S., Morstatter, F., Marshall, G., Liu, H., Nambiar, U. Navigating information facets on Twitter (NIF-T). In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, pp. 1548–1551 (2012)
6. Lin, C.X., Ding, B., Han, J., Zhu, F., Zhao, B.: Text cube: computing in measures for multidimensional text database analysis. In: Eighth IEEE International Conference on Data Mining, vol. 54, pp. 905–910 (2008)
7. Oukid, L., Asfari, O., Bentayeb, F., Benblidia, N., Boussaid, O.: CXT-cube: contextual text cube model and aggregation operator for text OLAP
8. Ravat, F., Teste, O., Tournier, R., Zurfluh, G.: Designing and implementing OLAP systems from XML documents. In: Kozielski, S., Wrembel, R. (eds.) New Trends in Data Warehousing and Data Analysis. AIS, vol. 3, pp. 1–21. Springer, Heidelberg (2008)
9. Ravat, F., Teste, O., Tournier, R., Zurfluh, G.: Finding an application-appropriate model for XML data warehouses. Inf. Syst. **36**(6), 662–687 (2010)
10. Park, B.-K., Han, H., Song, I.-Y.: XML-OLAP: a multidimensional analysis framework for XML warehouses. In: Tjoa, A.M., Trujillo, J. (eds.) DaWaK 2005. LNCS, vol. 3589, pp. 32–42. Springer, Heidelberg (2005). doi:10.1007/11546849_4
11. Zhang, D., Zhai, C., Han J.: Topic cube: topic modeling for olap on multidimensional text databases. In: SDM 2009: Proceedings of the 2009 SIAM International Conference on Data Mining, Sparks, NV, USA, pp. 1124–1135 (2009)

A Time Delay Neural Network for Online Arabic Handwriting Recognition

Ramzi Zouari[1(✉)], Houcine Boubaker[1], and Monji Kherallah[2]

[1] National School of Engineers of Sfax (ENIS), University of Sfax, Sfax, Tunisia
ramzi.zouari@gmail.com, boubaker-houcine@ieee.org
[2] Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia
monji.kherallah@enis.rnu.tn

Abstract. Handwriting recognition is an interesting part in pattern recognition field. In the last decade, several approaches are focused on online handwriting recognition because the very rapid growth of new technologies in the field of data entry. In this paper, we propose a new system for online Arabic handwriting recognition based on beta-elliptic model which allow to segment the trajectory into segments called strokes by inspecting the extremums points of velocity profile and extract their dynamic and geometric profiles. These strokes are used to train the Time Delay Neural Network (TDNN) which is able to represent the sequential aspect of input data. To evaluate our method, we have used a total of 25000 Arabic letters from the LMCA database. Our experimental results demonstrate the effectiveness of our proposed method and show recognition rates exceeds the 95%.

Keywords: Time delay neural network · Beta-elliptic model · Beta impulse · Velocity · Receptive fields · Shared weights

1 Introduction

Since few decades, several researches have focused on the study of Arabic handwriting because it has different properties compared to other languages. This is a cursive script that contains 28 basic letters where the majority of them change their shape according to their position in the word (ـهـ, ـهـ, ـه, ه) [1]. Handwriting analysis can be made in order to recognize the script, identifying the writer and identifying the emotional state of the writer. There are two branches of research in handwriting recognition. The first one is off-line where the handwriting is obtained from a digital input source (scanner, digital camera); in this case we obtain a binary image where the pixels values are either 1 or 0. The second branch, named online, handwriting script is captured by a touch-sensitive device (tablet). Then, dynamic information's are available such as (x, y) coordinates, velocity profile, pen pressure and temporal order of trajectory [2]. In our study, we deal with online handwriting recognition systems.

In order to model the sequential input data, several works have used Hidden Markov Models (HMM) and Recurrent Neural Networks (RNN). We cite the system made by Bezine et al. [3] that consist on segmenting the handwriting into strokes and assign to each one an Elementary Perceptual Code (EPC). These EPCs present the

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_99

observations sequences used to train a HMM. The proposed model achieves a higher rate with samples from Mayastroun Database. Abdelaziz et al. [4] introduced a novel method that handles delayed strokes. They extracted some geometric features as chain code, aspect ratio, etc. These features are trained on fully-fledged Hidden Markov Model (HMM) and showed recognition accuracy of 99% on a large vocabulary database. In [5] local and vicinity features are extracted from ADAB database. These features are trained and tested with a continuous HMM and reached a rate of 96.03% in top10. Among the systems that have used RNN, we can cite the work of Maalej et al. [6], who presents an online system based on deep recurrent neural networks with a Long Short Term Memory. This network is feed by a sequence of trajectory coordinates (x, y) with a dropout technique to improve the performance of its system. The idea of using the (x, y) coordinates is also proposed in [7] who used it to calculate the Fourier descriptor and the chain code. The proposed method is tested on Support Vector Machine (SVM) classifier and achieved a recognition rate of 92.43%.

Another study was made by Kherallah et al. [8] consists on segmenting the handwriting into strokes and on converting it into visual codes based on freeman method. In these codes, they applied the basic operations of genetic algorithm (GA) such as mutation and cross-over. This approach is assessed on LMCA database and showed an identification rate of 97%. The same technique of segmentation used in [8] is reproduced in [9] to develop a recognition system based on Graph-matching. This approach is tested on Postal Addresses extracted from ADAB database and reached a recognition rate about 98%.

In literature, several researches combine both on line and offline features of handwriting trajectory. Elleuch et al. [10] used this technique on Convolutional Deep Belief Network (CDBN) and obtained recognition rate up to 97% with LMCA database. Gargouri et al. [11] extracted the freeman chain code from the online coordinates and Zernike Moments from the offline handwriting. These features are tested using three different classifiers (DTW, LD and SVM). The results of tests on ADAB database gave a recognition rates near to 97%.

In other hand, few studies have chosen to hybridize multiple classifiers. In this context, Tagougui et al. [12] proposed a hybrid MLP/HMM modeling approach. The outputs of the MLP are decoded using a Discrete HMM with a codebook of size 512 which is used only in test phase. This architecture is evaluated on ADAB database and achieved recognition rate about 96.45%. Hamdani et al. [13] combine multiple HMMs using both online and offline features. They system is evaluated using the IFN/ENIT database and improved the recognition accuracy from 63.8% to 81.93%.

In this work, we present a new method for online Arabic handwriting recognition. The Beta-elliptic approach is used to model the handwriting trajectory, which is widely used in many fields on online handwriting [14]. Our proposed system is the first to use a Time Delay Neural Network (TDNN) in Arabic handwriting and also the first to use the beta elliptic parameters with TDNN, which present our contribution in this study.

The rest of this paper is structured as follows: Sect. 2 present the beta-elliptic model. Section 3 we give a description of the TDNN architecture and the applied training algorithm. Section 4 describes the architecture of our proposed methods. In Sect. 5, we evaluate the results of the experiments and we present the conclusion and future work.

2 Beta-Elliptic Model

The beta-elliptic approach allow to combining the two aspects of writing action: velocity and the trajectory [15]. The curvilinear velocity $V_\sigma(t)$ is given by the following formula (1):

$$V_\sigma(t) = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} \tag{1}$$

The curvilinear velocity of a continuous cursive handwriting shows a signal profile starting from a zero speed, then alternate minimum and maximum speed and finish by an almost zero speed. The points of velocity extrema correspond to the maximum values of deviation angle of the tangent to the trajectory. The handwriting trajectory is then segmented into simple movements called strokes which they can be determined by an inspection of maximums, minimums and inflexion points of the curvilinear velocity signal [15] (see Fig. 1).

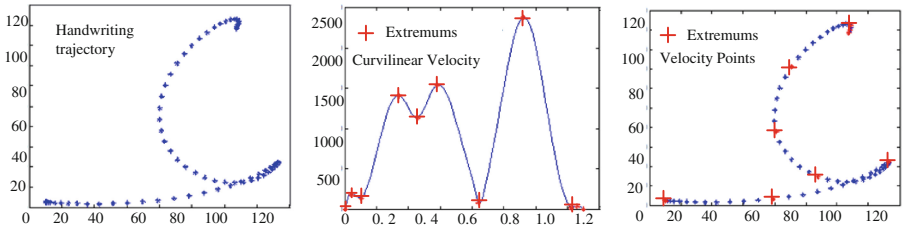


Fig. 1. Inspection of extremums points of velocity profile from trajectory

2.1 Velocity Profile Modeling

Plamondon et al. [17] has proved that curvilinear velocity of one stroke can be approximated by the Beta model (see Fig. 2). Therefore the curvilinear velocity is the result of an algebraic addition of the overlapped Beta profiles of its successive segmented strokes (see Fig. 3).

$$V(\mathrm{t}) \approx \sum_{i=1}^{n} K.\beta(t, p, q, t0, t1) \tag{2}$$

$$\beta_i(t_i, q_i, p_i, t_{0i}, t_{1i}) = \left\{ \begin{array}{ll} \left(\frac{t-t_{0i}}{t_{Ci}-t_{0i}}\right)^{p_i} \cdot \left(\frac{t_{1i}-t}{t_{1i}-t_{Ci}}\right)^{q_i} & \text{if } t \in [t_{0i}\,,\, t_{1i}] \\ 0 & \text{elsewhere} \end{array} \right\} \tag{3}$$

With

- K_i : amplitude of the i^{th} beta impulse
- p, q are intermediate parameters

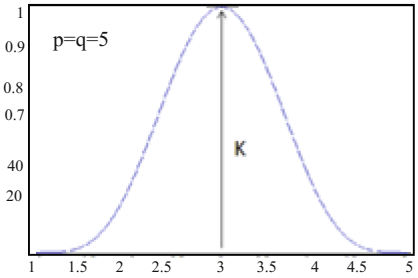


Fig. 2. Beta model

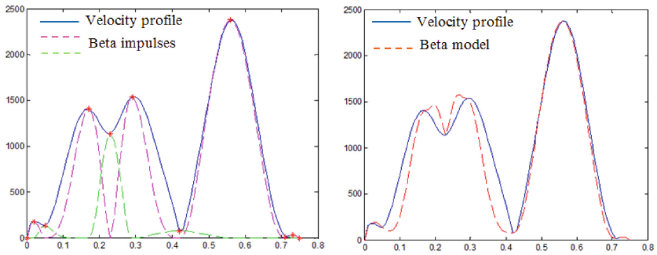


Fig. 3. Velocity signal modeling by overlapping Beta impulses

– t_0, t_c, t_1 are the moments witch correspond respectively to the start, the maximum amplitude and the end of the Beta function ($t_0 < t_c < t_1$).

$$t_C = \frac{(p \times t_0) + (q \times t_1)}{p + q} \tag{4}$$

2.2 Trajectory Profile Modeling

In the space domain, each stroke delimited between two consecutive points of velocity extremum is assimilated to an elliptic arc [18]. From each arc, we calculate three parameters (a, b and θ) that corresponds respectively to the half ellipse axis major length, half ellipse axis small length and major axis ellipse inclination angle (see Fig. 4).

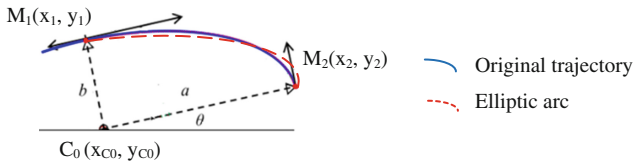


Fig. 4. Trajectory stroke fitting by elliptic arc.

3 Time Delay Neural Network

The TDNN, Time Delay Neural Network is a convolution neural network (CNN) which was first introduced for speech recognition [19]. The architecture of this network is derived from a Multi-Layer Perceptron (MLP) with back-propagation learning. In MLP, the order of input features doesn't affect the result of training because each neuron in hidden layer is fully connected to all neurons of the input layer simultaneously (global view), whereas in TDNN each neuron in hidden layer is connected to a subset of neurons from the preceding layer that represent a receptive field. Therefore TDNN is very useful for a sequential data [20]. TDNN incorporate a very important notion which is weight sharing. Then the neurons in different receptive fields shared the same weights. This constraint reduces the number of parameters in the system and facilities the generalization process. To build a TDNN we have to define the number of layers, the size of receptive fields that must be the same between two consecutive layers, and finally the value of temporal shift between receptive fields called Delay.

The architecture of TDNN consists of two parts (see Fig. 5). The first present the lower layers and implements the convolutions. It permits to transform the input features into another sequence of higher order features. The second part is similar to MLP and contains the outputs of network. This architecture was used by Poisson et al. [21] and was tested on IRONOFF and UNIPEN databases and obtained significantly higher performances near to 98%. In other hand, TDNN allow them to reduce the number of coefficients from 36,110 for a MLP to 17,930 with MLP.

4 Proposed System

In this part, we present the different steps of our recognition system. The first is the preprocessing, then the feature extraction is described and finally we detailed the recognition engine with TDNN (see Fig. 6).

4.1 Preprocessing

Before extraction the features, some processes on online handwriting are necessary. In our case all handwriting shapes are normalized in size and we applied a Chebyshev second type low pass filter on the normalized trajectory to eliminate the noise introduced by temporal and spatial sampling [22].

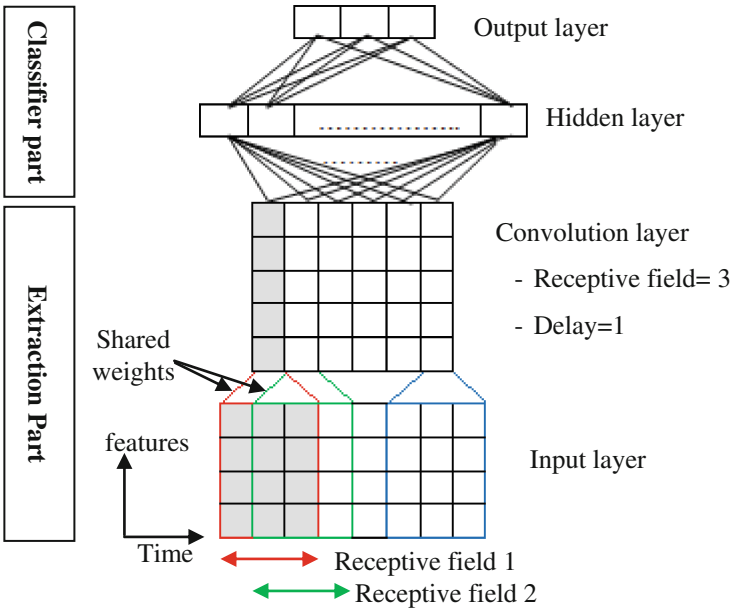


Fig. 5. TDNN architecture

4.2 Features Extraction

In this step, we extract the dynamic and geometric features of normalized handwriting. This is made by a beta-elliptic model as described in Sect. 2. For each stroke, we extract 4 dynamic parameters (K, Δt, Rap and p) resulting from the beta modeling of trajectory and 3 geometric profiles (a, b and θ) which represent the parameter of elliptic arc that modeled the stroke. The resulting feature vector is then composed of 7 features explained in the following table [23] (Table 1):

4.3 Recognition Engine

The proposed recognition system is based on Time delay neural network (TDNN). The choice of this network is justified by this ability to represent the sequential input data thanks to the principle of receptive field. Since our proposed system segment the handwriting trajectory into strokes, we can considerate the set of strokes as a sequential data (see Fig. 5). We note that the number of neurons in the convolution layer can be calculated through the following formula:

$$nb_{conv} = \frac{nb_{strokes} - receptive_field}{delay} + 1 \tag{5}$$

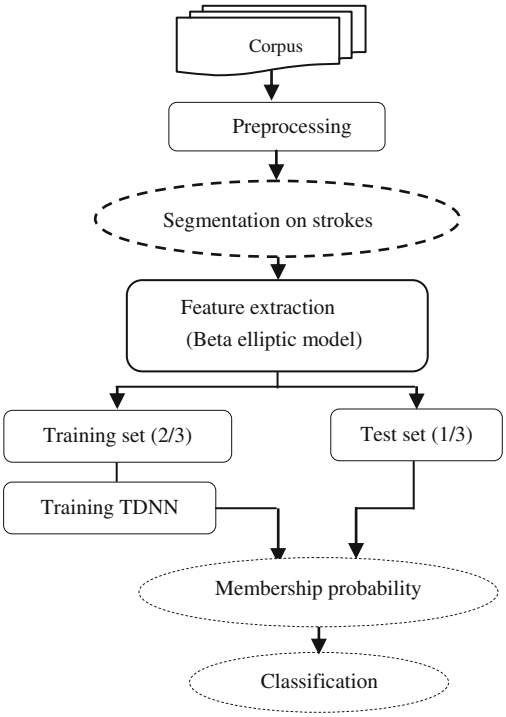


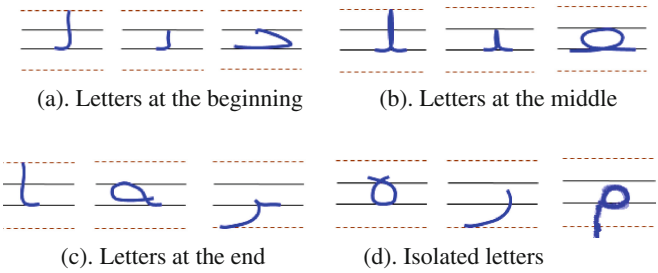
Fig. 6. Architecture of proposed system

Table 1. Component of beta-elliptic features vector.

Features	Parameters	Explanation
	K	Beta impulse amplitude
Dynamic profile	$\Delta t = (t1 - t0)$	Beta impulse duration
	$Rap = \frac{\Delta t_c}{\Delta t} = \frac{p}{p+q}$	Rapport of Beta impulse asymmetry or culminating time
	P	Beta shape parameters
Geometric profile	a	Ellipse major axis half length
	b	Ellipse small axis half length
	θ	Ellipse major axis inclination angle

5 Experiments and Results

To test and assess our proposed system, we have used the LMCA (Letter Mot Chiffres Arabes) database for our experiments [24]. LMCA contain the different letters of Arabic language in their different positions in the word (beginning, middle, end or isolated). It is composed of 212 files containing the (x, y) coordinates of 25000 handwriting letters belonging to 56 classes. The delayed strokes and diacritics are ignored. The following figure shows some letters from the LMCA database [18] (Fig. 7).



(a). Letters at the beginning (b). Letters at the middle

(c). Letters at the end (d). Isolated letters

Fig. 7. Letters from LMCA database

The LMCA database is divided into two parts. Two thirds of letters has been used for the training phase, and the rest has been used for the identification. The TDNN is training with back propagation and gradient descent. The learning rate is fixed at 0.0001 and the momentum at 0.6. The best recognition rate obtained by our system is 95.16% with a number of two convolution layers and one hidden layer. These results are obtained with a receptive field equal to 3 and temporal delay set to 1 (see Table 2).

Table 2. Results of experiment tests

Number of layers	Nbr of epochs	Recognition rate	Duration (mn)
- 1 convolutional layer - 1 hidden layer	40	91.89%	5
- 1 convolutional layer - 2 hidden layer	40	91.1%	9
- 2 convolutional layer - 1 hidden layer	40	94.1%	8
- 2 convolutional layer - 1 hidden layer	60	**95.16%**	12

The performance achieved by our approach using only the beta-elliptic features are better than results reported by Elleuch [10] and Boubaker [25]. This is due to the discriminatory power of features and the suited architecture of TDNN to process online handwriting signals. We notice that our proposed system reach a good recognition rate with a small number of epochs in a short time. This speed is due to the constraint of shared weights that allow reducing the number of parameters of network (see Sect. 3) (Table 3).

Table 3. Performance comparisons with other systems on LMCA database

Authors	Approach	Recognition rate
Elleuchet al. [10]	SVM	91.8%
Boubaker et al. [25]	MLP	94.14%
Kherallah et al. [8]	Genetic algorithm	96%
Our approach	**TDNN**	**95.16%**

6 Conclusion

We presented a new method for Arabic handwriting recognition. This method is based on Time delay neural network (TDNN). This network is widely used in speech recognition, and rarely used for online handwriting. The unique system based on TDNN is used for Latin handwriting recognition. In literature, servals researches uses a HMM and a Recurrent Neural Network to represent the sequential input data. These approaches require a big lot of computation. In contrast, TDNN requires less storage capacity due to its constraint of the weight sharing. The proposed system is the first to use TDNN in online Arabic handwriting. We have tested this network with beta-elliptic parameters and have showed this effectiveness in a short time of execution.

As perspective, we have to add other geometric parameters to the beta-elliptic features and test our system on Arabic word databases such as ADAB and IFN/ENIT.

References

1. Tagougui, N., Kherallah, M., Alimi, A.: Online Arabic handwriting recognition: a survey. Int. J. Doc. Anal. Recogn. (IJDAR) **16**(3), 209–226 (2013)
2. Chaabouni, A., Boubaker, H., Kherallah, M., Alimi, A.M.: Combining of off-line and on-line feature extraction approaches for writer identification. In: International Conference on Document Analysis and Recognition, pp. 1299–1303 (2011)
3. Bezine, H., Ghanmi, W., Alimi, A.M.: A HMM model based on perceptual codes for on-line handwriting generation. In: The Sixth International Conference on Advanced Cognitive Technologies and Applications, pp. 126–132 (2014)
4. Abdelaziz, A., SherifAbdou, A., Al-Barhamtoshy, H.: Large vocabulary arabic online handwriting recognition system. Patt. Recogn. Lett. **19**(4), 1129–1141 (2015)
5. Abdelazeem, S., Hesham, M.: On-line Arabic handwritten personal names recognition based on HMM. In: 11th International Conference on Document Analysis and Recognition System, pp. 1304–1308 (2013)
6. Maalej, R., Tagougui, N., Kherallah, M.: Online Arabic handwriting recognition with dropout applied in deep recurrent neural networks. In: The 12th IAPR Workshop on Document Analysis Systems, pp. 417–421 (2016)
7. Nakkach, H., Haboubi, S., Amiri, H.: Online Arabic character recognition using global and local features. In: The 3th International Conference on Automation, Control, Engineering and Computer Science, pp. 120–124 (2016)
8. Kherallah, M., Bouri, F., Alimi, A.M.: On-line Arabic handwriting recognition system based on visual encoding and genetic algorithm. Eng. Appl. Artif. Intell. **22**, 153–170 (2009)
9. Charfi, M., Kherallah, M., El Baati, A., Alimi, A.M.: A new approach for arabic handwritten postal addresses recognition. Int. J. Adv. Comput. Sci. Appl. **3**(3), 1–7 (2012)
10. Elleuch, M., Zouari, R., Kherallah, M.: Feature extractor based deep method to enhance online arabic handwritten recognition system. In: The 25th International Conference on Artificial Neural Networks, pp. 136–144 (2016)
11. Gargouri, M., Masmoudi, S.: Online Arabic handwriting recognition based on classifier combination. In: The 2nd International Conference on Automation, Control, Engineering and Computer Science, pp. 189–193 (2015)

12. Tagougui, N., Boubaker, H., Kherallah, M., Alimi, A.M.: Hybrid MLPNN/HMM recognition system for online Arabic handwritten script. Computer and Information Technology **4**(3), 107–118 (2013)
13. Hamdani, M., El Abed, H., Kherallah, M., Alimi, A.M.: Combining multiple HMMs using on-line and off-line features for off-line Arabic handwriting recognition. In: The 10th International Conference on Document Analysis and Recognition, pp. 201–205 (2010)
14. Boubaker, H., Kherallah, M., Alimi, A.M.: Handwriting and hand drawing velocity modeling by superposing beta impulses and continuous training component. Int. J. Comput. Sci. Issues **10**(5), 57–63 (2013). No. 1
15. Boubaker, H., Kherallah, M., Alimi, A.M.: New strategy for the on line handwriting modeling. In: The 9th International Conference on Document Analysis and Recognition, vol. 2, pp. 1233–1247 (2009)
16. Kherallah, M., Haddad, L., Alimi, A.M.: On-line handwritten digit recognition based on trajectory and velocity modeling. Pattern Recogn. Lett. **29**, 580–594 (2008)
17. Plamondon, R., Alimi, A.M.: Modelling velocity profiles of rapid movements: a comparative study. Biol. Cybemetics **69**, 119–128 (1993)
18. Dhieb, T., Ouarda, W., Alimi, A.M.: Online Arabic writer identification based on Beta-Elliptic model modelling. In: The 15th International Conference on Intelligent Systems Design and Applications, pp. 74–79 (2015)
19. Waibel, A., Hinton, G.: Phoneme recognition using Time delay neural network. IEEE Trans. Acoust. Speech Signal Process. **37**(3), 328–339 (1989)
20. Ferrat, F., Guerti, M.: Classification of the Arabic emphatic consonants using time delay neural network. Int. J. Comput. Appl. (0975 – 8887) **80**(10), 106–112 (2013)
21. Poisson, E., Lallican, P.: Multi-modular architecture based on convolutional neural networks for online handwritten character recognition. In: Proceedings of the 9th International Conference on Neural Information Processing, vol. 5, pp. 2444–2448 (2002)
22. Tagougui, N., Boubaker, H., Kherallah, M., Alimi, A.M.: A hybrid NN/HMM modeling technique for online Arabic handwriting. Computer Vision and Pattern Recognition, pp 1–6 (2013)
23. Boubaker, H., Kherallah, M., Alimi, A.M.: Spatio-temporal representation of 3D hand Trajectory based on Beta-Elliptic models. Int. J. Inf. Assur. Secur. **18**, 1632–1647 (2016)
24. Elleuch, M., Maalej, R., Kherallah, M.: A new design based-SVM of the CNN classifier architecture with dropout for offline Arabic handwritten recognition. In: International Conference on Computational Science, vol. 80, pp. 1712–1723 (2016)
25. Boubaker, H., Elbaati, A., Tagougui, N., Kherallah, M., Alimi, A.M.: Online Arabic databases and applications. In: Guide to OCR for Arabic Scripts, pp. 541–557 (2012)

A Proposal to Model Knowledge Dimension in Sensitive Business Processes

Mariam Ben Hassen(✉), Mohamed Turki, and Faïez Gargouri

ISIMS, MIRACL Laboratory, University of Sfax,
P.O. Box 242, 3021 Sfax, Tunisia
mariem.benhassen@isisms.usf.tn,
mohamed.turki@isetsf.rnu.tn,
faiez.gargouri@isims.rnu.tn

Abstract. Knowledge development in organizations relies on Sensitive Business Processes (SBPs), which are characterized by a high complexity and dynamism in their execution, high number of critical activities with intensive acquisition, sharing, storage and (re)use of very specific crucial knowledge, diversity of knowledge sources, and high degree of collaboration among experts. In this paper, we propose a semantically rich conceptualization for describing a SBP organized in a new Business Process Meta-model for Knowledge Identification (BPM4KI), in order to develop a rich and expressive graphical representation of SBPs to identify and localize the crucial knowledge. BPM4KI covers all aspects of business process modeling: the functional, organizational, behavioral, informational, intentional and knowledge perspectives. We focus more specifically on Knowledge Perspective which has not yet evolved into BP models. This perspective is semantically rich and well founded is on the «core» domain ontologies. Besides, we evaluate the relevance of some proposed concepts through a real SBP scenario from medical domain in the context of the organization of protection of the motor disabled people of Sfax-Tunisia.

Keywords: Knowledge management · Knowledge identification · Sensitive business process · Business process modeling · «core» domain ontologies

1 Introduction

Nowadays, modern organizations become increasingly conscious of the necessity to formalize and capitalize knowledge produced and used by their business processes (BPs). In this context, business process modeling (BPM) has become crucial concern for any successful organizations to improve the identification, acquisition, storage, dissemination, sharing, creation and (re) use of their individual and organizational knowledge. Considering the large amount of knowledge to be preserved and enhanced, such organizations must first identify and model the Sensitive Business Processes (SBPs) which are likely to mobilize crucial knowledge on which it is necessary to capitalize. In fact, the more organization's BPs are sensitive, the more they can mobilize crucial knowledge. Few existing research on Knowledge Management

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_100

(KM)-BPM focusing on the identification, analysis and modeling of SBPs in order to localize and identify the crucial knowledge[1]. We quote: the Global Analysis METHodology (GAMETH) [1], the identifying crucial knowledge methodology [2] and the Sensitive Organization's Process Identification Methodology [3]. However, the critical phase of « SBPs modeling » has not been explicitly addressed and studied in depth. Three major limitations can be emphasized. In particular, we have noted the absence of: (i) a rigorous scientific approach of BPM for knowledge identification, (ii) a rigorous conceptual specification for the SBP notion and clarity in the representation of its important features, and (iii) expressiveness of BPM formalisms and BP models with a knowledge dimension and other BPM aspects. In fact, the knowledge dimension (i.e. the knowledge required to perform activities, the knowledge created as a result of BP activities, the sources of knowledge, the explicit knowledge, the tacit knowledge, individual and collective dimension of knowledge/activities, the knowledge flows between knowledge sources and activities, the different opportunities of knowledge conversion, etc.) needed for performing SBP is not explicitly represented, integrated and implemented in BP models and BPM approaches. So, in order to remedy for these lacks, this paper aims to extend and consolidate previous work [2, 3], mainly to reduce the gap between BPM and KM and address an important problem that is not often dealt with by KM methodologies. Exactly, our mission aims to enrich and optimize the operation of « modeling and representation of identified SBPs » in order to increase the probability of localizing and identifying the crucial knowledge that requires capitalization. This reduces the cost of the operation of capitalizing on knowledge.

This research work presents a conceptual specification of SBP, by proposing a new multi-perspective meta-model of BPs representation with a view to identifying the knowledge, entitled « BPM4KI: Business Process Meta-Model for Knowledge Identification » . BPM4KI intends to explicit and organize the key concepts and relationships that characterize a SBP. It integrates all relevant aspects/dimensions relating to BPM-KM, i.e. the functional perspective, the organizational perspective, the behavioral perspective, the informational perspective, the intentional perspective and the knowledge perspective. In this paper, we focus more on the « Knowledge Perspective » (or « Knowledge Dimension »), which is not yet explicited and integrated into the BPs models. This perspective, modeled as an ontological conceptual pattern (OCP) [4, 5], is semantically rich and well-based on «core» domain ontologies (which are based on top of the DOLCE foundational ontology [6]. These ontologies provide a set of referential of generic concepts and relationships semantically rich and consensual which we reused, firstly, to broaden and deepen the elements of definition of Knowledge dimension, and on the other hand, to characterize the useful concepts for a rigorous specification and an enriched modeling of the SBPs.

[1] The first facet of knowledge capitalization process [1] concerns problems bound to the identification and localization of crucial knowledge, that is knowledge (explicit knowledge) and knowhow (tacit knowledge) which are necessary for decision-making processes and for the progress of the essential processes which constitute the heart of the activities of the company: it is necessary to identify them, to localize them, to characterize them, to make cartographies of them, to estimate their economic value and to organize them into a hierarchy.

The remainder of this paper is organized as follows: Sect. 2 presents a specification of SBP, describing its main features. Section 3 presents the central concepts that describe the knowledge perspective of SBP necessary for the BPM. Section 4 illustrates the application and the relevance of these concepts, based on a real case study. Section 5 concludes the paper and underlines some future research topics.

2 Sensitive Business Processes

A SBP is a particular type of BP. It has its own characteristics that distinguish it from classical BPs. Indeed, a SBP represents the core process of organization which constitutes the heart of the organization's activities [7–9].

2.1 Main Characteristics of SBPs

A BP is described as « sensitive », if at least one of the following requirements is fulfilled:

- It mobilizes « crucial knowledge », i.e. the most valuable/important knowledge on which it is necessary to capitalize, in the sense that the risk of their loss and the cost of their (re)creation is considered to be important, also their contribution to reach the firm objectives is very important and their use duration is long.
- It contains activities that valorize the acquisition, storage, dissemination, sharing, and creation and (re) use of individual and organizational knowledge, in the sense that it mobilizes a large diversity of knowledge sources consigning a great amount of very important heterogeneous knowledge.
- It is heavily dependent on the tacit knowledge embedded in the stakeholders' minds (i.e. experts). In fact, this knowledge is mainly implicit, rarely explicit and disseminated the by experts who hold it, and will therefore be difficult to identify, exploit and valorize by other collaborators.
- It includes a high number of critical activities [1–3] mobilizing crucial knowledge. In our context, a critical activity mobilizes different types of knowledge: (i) imperfect individual and collective knowledge (tacit and/or explicit) (i.e. missing, poorly mastered, incomplete, uncertain, etc.) which are necessary for solving critical determining problems; (ii) a great amount of heterogeneous knowledge recorded on multiple knowledge sources (dispersed and sometimes lacking accessibility); (iii) expertise and/or rare knowledge held by a very small number of experts (who carry out actions with high levels of experience, expertise, and creativity); flexible knowledge owned by experts; (iv) very important tacit organizational knowledge (like competences, abilities and practical experiences).
- Its execution involves a large number of business domains (in terms of internal and external organization unit/agents operating in the BP), having distinct experience and expertise levels.

- It include a high number of collaborative organizational activities (inter/intra organizational) that mobilize, exchange, share and generate new individual and collective knowledge that is created by dynamic conversion of existing ones in the BP. So, it depends on knowledge flows and transfer of data, information and knowledge objects between communicating process participants.
- It is typically unstructured or semi-structured. A flexible BP contains a very dynamic and unpredictable control-flow, comprising complex activities that may frequently change over time or at design-and run-time. Moreover, SBP participants may be influenced by or may have to comply with constraints and rules that drive organizational actions performance and decision making.
- It possesses a high degree of dynamism in the objectives' change associated to it, essentially, in decision making context. The change of organizational objective leads to a new organizational distal intention (which is necessary to control the SBP) and influences experts' decision making.
- Its contribution to reach strategic objectives of the organization is very important. Also, their realization duration are important and often their costs are very high.

Due to those characteristics, modeling and organizing the knowledge involved in SBP is relatively critical.

2.2 SBP Specification: A Multi-Perspective Meta-Model of BPs for Knowledge Identification Based on Core Ontologies

In order to enrich and improve the SBP modeling, we propose a new multi-dimensional generic meta-model of BP representation for knowledge identification, entitled « BPM4KI: Business Process Meta-model for Knowledge Identification ». This enriched meta-model serves two purposes: (i) to deepen the elements and dimensions defining an SBP, by offering a coherent conceptual specification for this BP type, and (ii) to develop a rich and expressive graphical representation of SBPs to improve the localization and identification of crucial knowledge mobilized and created by these processes. This new extended meta-model, which is a continuation of previous works [7–9], is well founded whose concepts (and the relationships between them) are semantically enriched by the «core» domain ontologies. Specifically, the SBP conceptualization is jointly supported, on the one hand, by the specialization of the *DOLCE* foundational ontology [6], to specify and define the invariant generic BP concepts, including SBPs, and on the other hand, by the conception of Ontological Conceptual Patterns (OCP) [4, 5] based on the reuse of ontological modules relating to the «core» domain ontologies [10–14] (which were based on top of DOLCE).

The current version of BPM4KI offers a referential of generic concepts and semantic relationships relevant to the BPM-KM domain, exploring the dynamic, the collaboration and the knowledge aspects of the SBP in greater detail. It is categorized in six perspectives (or dimensions) which are complementary and essential for a comprehensive and expressive characterization and representation of an SBP:

- *Functional Perspective*, represents the BP elements which are being performed. The main concept that reflects this dimension is `Action`[2]. It includes: `Individual Action`, `Action of Collective`, `Action of Organization`, `Organizational Action`, `Organizational Unit Action`, `Critical Organizational Activity`, `Collaborative Organizational Activity`, etc.
- *Organizational Perspective*, represents the different participants (the organizational resources) invoked in the execution of process elements as well as their affiliation. The basic element of this perspective is `Agentive Entity` and includes: `Collective`, `Organization`, `Organization Unit`, `Human`, `Expert`, `Internal Agent`, `External Agent`.
- *Behavioral perspective*, depicts the concepts required to demonstrate the flows of activities and information, coordination between different participants as well as concepts that effect, trigger or control flows of activities. The basic and generic concepts of this perspective are `Control Flow`, `Flow Node`, `Control Node`, `Connecting Object`, `message Flow`, `Conditional Control Flow`, `Non Conditional Control Flow`, etc.
- *Informational perspective*, describes the informational entities (such as input, output, physical or data objects, artifacts, messages) which are generated, consumed, or exchanged within a process/activity and between different participants. The following concepts are related to this dimension: `Resource`, `Material Resource`, `Immaterial Resource`, `Informational Resource`, `Event`, `Contingency`, `Information Object Input`, `Data Object Input`, `Information Object Output`, `Data Object Output`, `Physical Artifact`, `Communication`, `Informal Exchange`, `Message`, `Physical Knowledge Support`, etc.
- *Intentional perspective*, provides an overview perspective of the process and describes major BP characteristics. It captures important BP context information in order to ensure the BP flexibility. It comprises: `Intention`, `Collective Distal Intention`, `Organizational Distal Intention`, `Individual Objective`, `Collective Objective`, `Organizational Objective`, `Strategic Objective`, `Deliberate Action`, `Culminated Process`, `Client`, `Sensitive Business Process`, `Internal Process`, `External Process`, `Core Process`, `Strategic Process`, etc. (which are some process types), etc.
- *Knowledge perspective*, provides an overview perspective of the organizational and individual knowledge mobilized and created by an BP/organization as well as the knowledge flow proceeding within and between BPs/organizations. It addresses all relevant aspects related to KM. This vantage presents the generic concepts: `Knowledge`, `Capacity`, `Individual Knowledge`, `Collective Knowledge`, `Internal Knowledge`, Tacit Knowledge, `External Knowledge`, `Explicit Knowledge`, `Propositional Knowledge`, `procedural`

[2] With respect to our notation, the informal labels on BPM4KI concepts appear in the text in the Courrier new font with First Capital Letters for the concepts and a javaLikeNotation for relations.

`Knowledge`, `Strategic Knowledge`, `Expert`, `Physical Knowledge Support`, `CommunitativeInteraction`, `Knowledge Flow`, `Socialization`, `Externalization`, `Explicitation`, `Combination`, etc.

In the following, we describe our proposal of modeling of the « Perspective Knowledge » that we have extended in greater detail.

3 Knowledge Dimension in SBP Modeling

According to the literature review, the integration of KM into BPs was identified as the most pressing as well as the most promising practical and theoretical task in KM [15]. Therefore, several attempts have already been made to integrate the domain of KM and BPM, which consist in introducing the process dimension into KM [16–19] or knowledge dimension into BP models [20–27]. However, the integration of KM and BP orientation has not yet received sufficient attention. In fact, the knowledge dimension needed for BPM is not explicitly represented, integrated and implemented in BP models and BPM approaches/formalisms. To address this research gap, the current paper proposes an extended knowledge perspective as the most relevant BPM dimensions, which has not yet evolved into BP models, in order to enrich the graphical representation of SBPs and improve the localization and identification of crucial knowledge mobilized and created by these processes. This extended perspective highlights the concepts and relationships needed to completely and adequately address SBP essential characteristics.

Knowledge perspective in BPM focuses on the knowledge flow and the dynamics of acquisition, preservation, conversion, transfer, sharing, development, and (re) use of individual and organizational knowledge within and between organizations. However, this perspective considers the different types of knowledge (tacit/explicit dimension, declarative/procedural dimension, etc.) mobilized and created by each type of activity related to the organization's BPs, the different sources of knowledge, their localization and where they are usable or used, their nature, their mode of dissemination/organizational coverage (individual/collective dimension) (individual/collective dimension), the different opportunity of knowledge conversion and creation, etc.

Our proposal, presented in the form of an Ontological Conceptual Pattern (OCP) of knowledge (see Fig. 1) is based on the reuse and the specification of central generic concepts (and the relationships between them) defined in different ontological modules of the global and consistent ontology OntoSpec[3] [10]: *Capacity-OS, Action-OS, Action of Organization-OS, Agentive Entity-OS, Participation-role-OS, Organization-OS, Function & Artefact-OS Resource-OS, Communication-OS and Discourse_Message_Discoure_act-OS*. Furthermore, we reused some concepts proposed in the ontological modules *IE&C* (Inscription, Expression and Conceptualization) [11], *Know-How* and *Knowing-That* [14]. These different core domain ontologies are useful and sufficient to characterize the useful concepts for a rigorous specification and an enriched modeling of Knowledge dimension of SBPs. Figure 1 organizes and explicit

[3] http://home.mis.u-picardie.fr/∼site-ic/site/spip.php?article53.

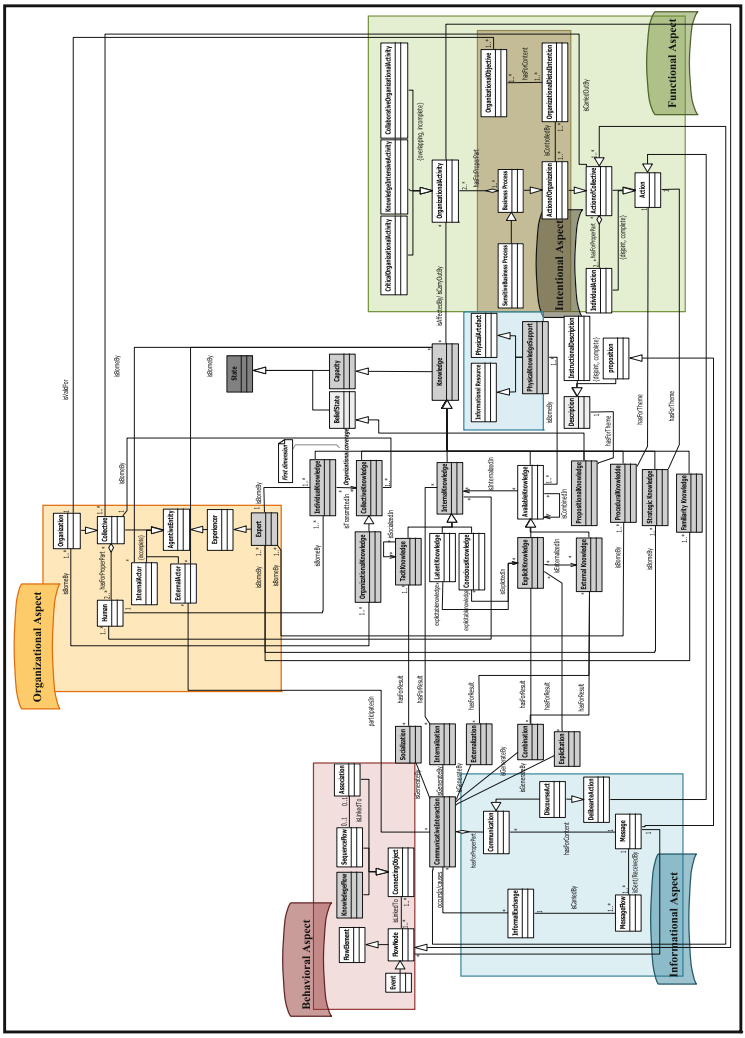


Fig. 1. Ontological conceptual pattern relating to the knowledge perspective/dimension of SBP

the central concepts of the knowledge perspective (marked in gray), in addition to inter-aspects relationships, giving a view of all relevant aspects of BPM4KI as a whole.

In the literature review, different connotations were associated with the term « Knowledge ». A more general characterization of this concept is given in [28–31]: « *A knowledge (or know-how) is the capacity (or the disposition) to perform an action aiming to achieve a goal"*. Meanwhile, *A Capacity is a state which enables to fulfill at least one Action Role* (i.e. play some role of participation in an Action) » [31]. Thus, we propose to characterize knowledge as follows: A `Knowledge` is the `Capacity` which `carriesOut` (and `affects`) a type of `Action` (i.e. an `Organizational Action`), which `isBornedBy` an `Agentive Entity` (which can be a `Human`, a `Collective`, an `Expert` or an `Organization`).

Several features of Knowledge have been further described in the literature and this leads us to classify it according to some dimensions (see Fig. 2). In this paper, we retain the following dimensions: (a) the *source of knowledge,* (b) *nature (content) of knowledge*, (c) *organizational value of knowledge*, and (d) *organizational coverage of knowledge*, which seem useful and relevant to the context of our research work, *the localization and identification of knowledge*.

(a) ***Source of knowledge***. According to this dimension, we propose to make the distinction between three types of knowledge: `Internal Knowledge`, `Explicit Knowledge` and `External Knowledge`. `Internal Knowledge` (or implicit knowledge) `isBornBy` one or more `Persons` (`Humans`). It is an implicit knowledge related to the human being, who exists in the head of its holder, and includes innate or acquired competences and skills, his/her beliefs and aspirations. An `Internal Knowledge` could be either `Conscious knowledge`, `Latent knowledge` or `Tacit Knowledge`. `Conscious knowledge` is a conscious and intentional knowledge, is cognitively available

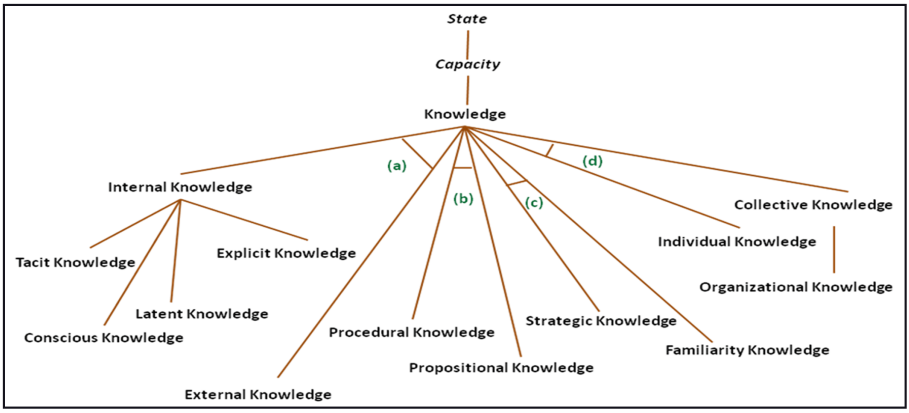


Fig. 2. Knowledge is classified according to several dimensions: (a) source of knowledge, (b) nature of knowledge, (c) organizational value of knowledge, (d) organizational coverage of knowledge. A descending edge between two concepts represents a subsumption link. A horizontal line between edges from a father concept indicates that the sibling concepts are incompatible.

and may be made explicit easily. Latent knowledge has been typically learning as a by-product and is not available consciously. It may be made explicit, for example in situations, which are similar to the original learning situation, however. Tacit knowledge is highly personal, developed and acquired through experiences and (cultural) socialization situations. It is a non explainable knowledge, and is specific in its context and based on intuition and perception. Explicit Knowledge concerns knowledge that is « made explicit » (explained) to the outside world, e.g., through spoken language, but is still bound to the human being. External Knowledge is detached from the human being and may be codified, formalized, transferable and usable at the collective level of the organization, in the form of reusable storage media (e.g. documents) as part of the organizational memory and independently of the subject which specified it. Available Knowledge (either Explicit Knowledge or External Knowledge) isBornBy one or more Physical Knowledge Supports (e.g. documents, computer system, etc.) enabling their capitalization, dissemination and sharing among stakeholders of the organization. It is a way for formalizing and storing Explicit/External Knowledge. A Physical Knowledge Support is a Material Resource (informational resource), having source of knowledge information interpreted and mobilized by the agents (operating in the BP) during the execution of their activities (these supports transmit not only information, but also significance). From this viewpoint, knowledge is useful for interpreting information while information is useful for transferring knowledge.

(b) ***Nature of knowledge***. We mainly distinguish between two categories of knowledge: Propositional Knowledge and Procedural Knowledge. These types of knowledge are dependent on each other and have some points in common with the typology presented above. Propositional knowledge « Knowing-That /What » (also called declarative, factual or descriptive knowledge [32] corresponds to theoretical knowledge, « knowledge » memorization of facts, events, rules or principles of a given domain. They are usually made in a formal language. Procedural knowledge « Knowing-How » corresponds to the Experience, the practical knowledge, the knowledge on « how-to-do », the procedure of performing the Action, the « know-how ». These are dynamic knowledge that may develop only in a context of action. The authors [14] have defined two terms: « Know-How » and « Knowing-That » as defined by Ryle [28, 33], and distinguished between the disposition to perform an action, and the belief state concerning a description, which may be either factual or propositional. Moreover, the correspondence of these two concepts of the core ontology COOK [14] with Procedural Knowledge and Propositional knowledge concepts, leads us to propose the following characterization: A Propositional knowledge is a Capacity (or disposition) which hasForTheme (concerns) an Action while Propositional knowledge is a Belief State (which is a State) which hasForTheme a Description.

(c) ***Organizational value of knowledge***. We propose two classes of knowledge: Strategic Knowledge and Familiarity Knowledge. Strategic knowledge « Know-Why/When » (also called conditional or pragmatic

knowledge) is meta-cognitive knowledge on optimal strategies for structuring a problem-solving approach or for carry out an Action. They are also dynamic knowledge, which correspond to the why and how of the action and, also, allow to determine the time and context in which it is appropriate to use such a procedure or such knowledge. This type of knowledge is one of the characteristics of expertise. Thus, a Strategic Knowledge hasForTheme an Action. Finally, the Familiarity Knowledge (called environmental or deep knowledge) is acquaintance with certain situations and environments; it also resembles aspects of situational knowledge, i.e., knowledge about situations, which typically appear in particular domains.

(d) ***Organizational coverage of knowledge***. We distinguish between two main classes: Individual Knowledge and Collective Knowledge. An Individual Knowledge isBorneBy an Individual (a Human). This knowledge includes, besides the intellectual capacity, the vision that every individual holds of the organization to which he/she is affiliated, as well as the explicit knowledge in the form of personal notes. As opposed to the individual, a Collective Knowledge isBornBy a Collective. It's a knowledge sharable by several people. Their coverage can be global, or partial which bound to a structure. An Organizational Knowledge is a Collective Knowledge, which isBornBy an Organization. Thus, the different types of knowledge, presented above, can be held individually or collectively, internally or externally. In the first case, Knowledge areBornBy an Internal Actor, which is an Agentive Entity (an Individual, a Collective, an Organizational Unit, or an Organization) internal to an Organisation (affiliatedTo the same Organization). While in the second case, Knowledge areBornBy an External Actor, which is an Agentive Entity external to an Organisation (which is not affiliatedTo Organization). Individual Latent Knowledge, Individual Tacit Knowledge, Individual Procedural Knowledge, Individual Strategic Knowledge and Individual Familiarity Knowledge areBornBy one Expert. It is linked to the mental models, talents, innate or acquired experiences and skills, abilities, individual tricks, trades secrets, etc. While, Collective Latent Knowledge, Collective Tacit Knowledge Collective Procedural Knowledge, Collective Strategic Knowledge and Collective Familiarity Knowledge areBornBy at least two Experts (which constitute a Collective) and mobilized by a Collective Action. It manifests itself in routines which may be shared and exchanged through direct communication with others. Un Expert is an Experiencier (which is an Agentive Entity) who bears a Capacity To Perform a type of Action with high levels of experience, expertise, performance, creativity and innovation. Furthermore, Agentive Entity carriesOut one or more Communicative Interactions, which hasForProperPart Communications. Within this type of action, different types of Agentive Entity interact and exchange Knowledge through Messages typically in an informal way (Informal Exchanges). Therefore, new Knowledge is developed, which is the result of a sequence of operations which,

starting from the private and non formalized individual dimension of knowledge, determines the collective dimension (formalized and disseminated) of knowledge. Therefore, we relied on the theory of knowledge creation [34], and we propose to distinguish between five main modes/processes of conversion and transmission of different types of knowledge[4], namely: Socialization, Explicitation, Externalization, Internalization and combination. Indeed, Individual Tacit Knowledge isTransmittedIn Collective Tacit Knowledge through practice, sharing of experiences, conscious or unconscious observation, constructive discussions, learning, etc. (in the context of a Communicative Interaction). This is a matter of Socialisation. Socialisation is a Deliberate Action, which hasForAgent a Collective (which hasForProperPart at least two Humans), which isGeneratedBy Communication, and hasForResult new TacitKnowledge (mental models). Socialisation may involve the participation of External Actors. Knowledge Flows are needed to model the dynamics of Knowledge, i.e., all of organization, acquisition, conversion, transfer, sharing, development and usage of Knowledge among the different sources of knowledge, participants and among BP activities.

Once modeled, the SBPs can be graphically represented, using the most popular standard for BPM, BPMN 2.0 [35], in order to localize the knowledge that is mobilized and created by these processes. We selected BPMN as the most suitable BPM notations for representing SBPs, because it incorporates requirements for SBP modeling better than other BPM formalisms [9]. Nevertheless, despite its strength representation, BPMN 2.0 does not yet provide support for SBP modeling. Some of its concepts should be adapted and extended to be convenient for a rich and expressive representation of SBPs. In fact, this notation does not explicitly support the key concepts of the Knowledge perspective of BPM4KI (as Critical Organizational Activity, Explicit Knowledge, Individual Tacit Knowledge, Collective Tacit Knowledge, Expert, Distal Intention, etc.). So, to overcoming the shortcomings of BPMN 2.0, this extension must take into consideration, on the one hand, the knowledge dimension, and on the other hand, integrate the new concepts of BPM4KI to represent issues relevant at the intersection of KM and BPM with a sufficient level of details.

[4] Socialization converts tacit knowledge of a person into tacit knowledge of another person. This may succeed by exchange of experience or in a learning-by-doing situation. Explicitation is the internal process of a person, to make internal knowledge of the latent or conscious type explicit, e.g., by articulation and formulation (in the conscious case) or by using metaphors, analogies and models (in the latent case). Externalization converts from explicit knowledge to external knowledge or information and leads to detached knowledge as seen from the perspective of the human being, which can be kept in organizational memory systems. Internalization converts either external or explicit knowledge into internal knowledge of the conscious or latent types. It leads to an integration of experiences and competences in your own mental model. Finally, combination combines existing explicit or external knowledge in new forms.

4 Case Study

We intend to apply the concepts proposed by the Knowledge perspective of BPM4KI meta-model on a real case in the medical domain in the context of the Association of Protection of the Motor-disabled of Sfax-Tunisia (ASHMS). We aim to evaluate their potential to build an understandable, adequate and expressive representation of SBP, in order to improve the localization and the identification of knowledge. Particularly, we are interested in the early care of the disabled children with cerebral palsy (CP). An depth analysis of this care has been made by [36]. Our main objective consists in providing better localization and identification of relevant and pragmatic medical knowledge necessary to the conduct of the medical care process of the disabled children with Cerebral Palsy, which is a SBP. In this study, we take into consideration the results of experimentation of the methodology SOPIM proposed by Turki et al. [3] in the context of ASHMS. We have opted for the «Process related to the neuro-pediatric care of a child with CP» to illustrate the contributions of our enriched Knowledge perspective of BPM4KI. Figure 3 illustrates a SBP model extract of the neuropediatric consultation process using BPMN 2.0 [35], enriched with the knowledge dimension. As stated above, this notation does not, however, provide primitives to explicitly represent all relevant aspects related to knowledge dimension in BP models. To remedy for the shortcomings, we tried to extend this notation and started by integrating some specific graphical icons in the form of some BPMN modeling elements relating to several new BPM4KI concepts (see Fig. 3). During our experimentation, we have identified different types of medical knowledge mobilized and created by each critical activity related to the SBP of neuro-pediatric care.

For instance, the knowledge A_2K_{p1} « Synthesis of neurological abnormalities related to motor, somatic and sensory development of the young child at risk and the different clinical signs » is produced by the critical activity A_2 « Clinical neurological examination ». Note that this materialized/externalized knowledge is created as a result of the activity execution by the Neuro-Pediatrician, during which he interacts with information (i.e. source of knowledge information) related to the child with CP (based on his tacit knowledge) to generate and communicate his own knowledge. A_2K_{p1} is stored in the following physical media: the neurological assessment sheet, neuropsychological assessment, the sensitive assessment sheet and the neuro-motor assessment. These physical media of knowledge are located internally within the Neonatology service in the University Hospital Hedi Chaker, precisely in the various archives drawers or patients' directories. A_2K_{p1} is of a scientific, technical and measure nature which is related to patients. It represents a collective explicit knowledge, part of which can be represented in the form of an individual explicit knowledge recorded on the care data collection sheet of the Neuropediatrician. This knowledge is imperfect (general, incomplete and uncertain). A_2K_{p1} is mobilized by the activity A_3 « Evaluation of intellectual functioning of young child with CP ».

It is important to mention that not all BPM4KI concepts are applicable and must be instantiated in every SBP scenario. Precisely, relevant tacit aspects could not be represented explicitly, such as: the tacit knowledge embedded in the Neuropediatric's mind, the knowledge conversion and the knowledge flows exchanged between communicating process participants and among activities, and the distal intentions which

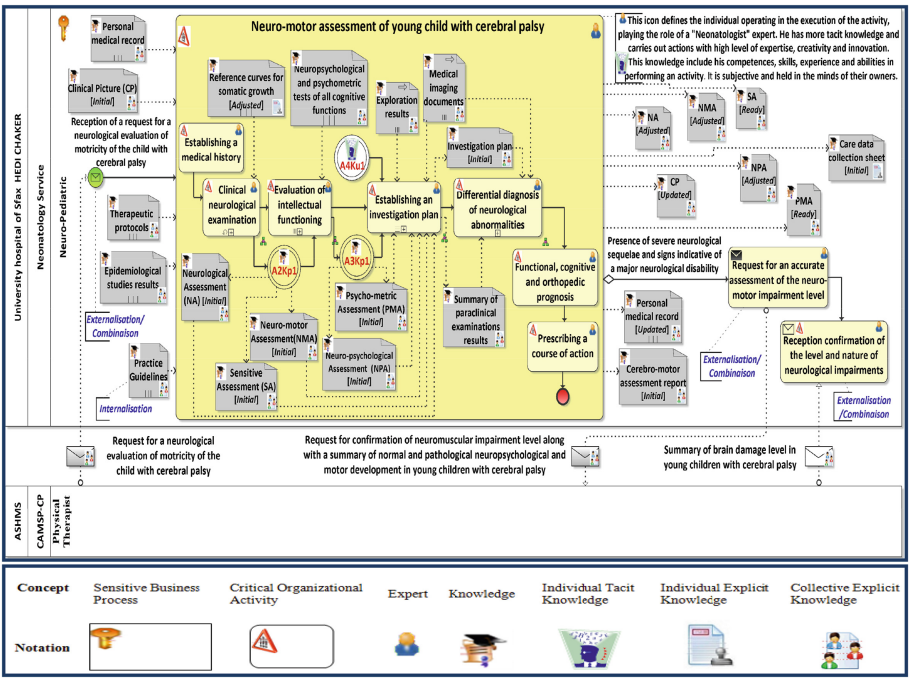


Fig. 3. Fragment of SBP model in BPMN related to the neuro-pediatric consultation of a child with CP

are responsible for making Neuro-Pediatrician to perform any action and achieve an organizational objective.

Therefore, the relevance of extending BP models with the knowledge dimension is manifold:

1. Enhance the localization and identification of the crucial knowledge mobilized and produced by the critical activities: (i) Illustrating the knowledge and its sources that are necessary for the execution of BP activities and are generated, created and/or modified as a results of activities. (ii) Illustrating the knowledge localization (where knowledge can be obtained and clearly stated) as well as experts who hold the (tacit) knowledge. (iii) Illustrating transfers of knowledge between sources, and among activities as well as the different opportunities of knowledge conversion. (iv) Defining the knowledge that is being captured or obtained from specific sources. (v) Giving an opportunity to improve understanding about the knowledge usefulness, validity, and relevance for particular activities in a SBP. (vi) Possibility to evaluate the amount of lost knowledge if a person-owner of knowledge- leaves the organization.
2. A deeply characterizing of the identified knowledge to determine which ones are more crucial to be exploited: (i) Illustrating the nature and degree of formalization of knowledge. (ii) Illustrating the organizational coverage of knowledge, their quality, etc.

5 Conclusion and Perspectives

In this paper, we have addressed the problematic of localization and identification of crucial knowledge. Therefore, we proposed a new multi-perspective meta-model representation of BPs, called « BPM4KI » allowing a rigorous specification and rich and expressive graphical representation of SBPs to improve the identification of knowledge mobilized and created by these processes. We focused, specifically, on the description of the « Knowledge Perspective » needed for performing SBP. This dimension modeled in the form of ontological conceptual pattern, is semantically rich and well-based on «core» domain ontologies. Furthermore, we illustrated the application of some offered concepts on a model of medical care process, using the BPMN 2.0 standard. There are several open issues in this paper that we plan to address in the future to deepen the so-called problematic of identification of crucial knowledge that is mobilized by SBPs. Further work is underway to deepen the characterization of concepts relating to the notion of sensitive process business and propose semantically rich and consensual definitions of concepts and dimensions necessary for the characterization and modeling of SBPs. In the medium term, we plan to offer a rigorous scientific approach to help implement an extension of BPMN 2.0 notation to integrate core BPM4KI concepts, and have a rich and expressive representation of SBPs.

References

1. Grundstein, M.: From capitalizing on company knowledge to knowledge management. In: Morey, D., Maybury, M. (eds.): Knowledge Management, Classic and Contemporary Works, pp. 261–287. The MIT Press, Cambridge (2000)
2. Saad, I., Grundstein, M., Sabroux, C.: Une méthode d'aide à l'identification des connaissances cruciales pour l'entreprise. Revue SIM **14**(3), 43–79 (2009)
3. Turki, M., Saad, I., Gargouri, F., Kassel, G.: A business process evaluation methodology for knowledge management based on multi-criteria decision making approach. In: Saad, I., Sabroux, C.R., Gargouri, F. (eds.) Information Systems for Knowledge Management. Wiley-ISTE, Chichester (2014). ISBN 978-1-84821-664-8
4. Gangemi, A., Borgo S. (Eds.) Proceedings of the 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2004), Workshop on Core Ontologies in Ontology Engineering, Northamptonshire (UK), vol 118 (2004). http://ceur-ws.org
5. Gangemi, A.: Ontology Design Patterns: A primer, with applications and perspectives. Tutorial on ODP, Laboratory for Applied Ontology Institute of Cognitive Sciences and Technology CNR, Rome, Italy (2006)
6. Masolo, C., Vieu, L., Bottazzi, E., Catenacci, C., Ferrario, R., Gangemi, A., Guarino, N.: Social roles and their descriptions. In: Dubois D., Welty C., Williams M.-A. (eds.) Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning, pp. 267–277 (2004)
7. Hassen, M.B., Turki, M., Gargouri, F.: Sensitive business process modeling for knowledge management. In: Chen, Q., Hameurlain, A., Toumani, F., Wagner, R., Decker, H. (eds.) DEXA 2015. LNCS, vol. 9262, pp. 36–46. Springer, Heidelberg (2015). doi:10.1007/978-3-319-22852-5_4

8. Hassen, M.B., Turki, M., Gargouri, F.: Towards business process modeling for knowledge management. In: Shishkov B. (eds.) International Symposium on Business Modeling and Software Design BMSD 2015. Selected best Papers. LNBIP, vol. 257, pp. 37–61 (2016)
9. Hassen, M.B., Turki, M., Gargouri, F.: Choosing a sensitive business process modeling formalism for knowledge identification. In: International Conference on ENTERprise Information Systems (CENTERIS 2016), Porto, Portugal, pp. 1002–1015 (2016)
10. Kassel, G.: Integration of the DOLCE top-level ontology into the OntoSpec methodology (2005)
11. Fortier, J.-Y., Kassel, G.: Managing knowledge at the information level: an ontological approach. In: Proceedings of the ECAI 2004 Workshop on Knowledge Management and Organizational Memories, Valencia, pp. 39–45 (2004)
12. Turki, M., Kassel, G., Saad, I., Gargouri, F.: A core ontology of business processes based on DOLCE. J. Data Semant. **5**(3), 165–177 (2016)
13. Kassel, G., Turki, M., Saad, I., Gargouri, F.: From collective actions to actions of organizations: an ontological analysis. In: Symposium Understanding and Modelling Collective Phenomena (UMoCop). University of Birmingham, Birmingham, England (2012)
14. Ghrab, S., Saad, I., Kassel, G., Gargouri, F.: An ontological framework for improving the model of contribution degree of knowledge. In: International Conference on Knowledge Management, Information and Knowledge Systems (KMIKS 2015), Tunisia, pp. 45–58 (2015)
15. Scholl, W., Konig, C., Meyer, B., Heisig, P.: The future of knowledge management. an international delphi study. J. Knowl. Manage. **8**(2), 19–35 (2004)
16. Suyeon, K., Hyunseok, H., Euiho, S.: A process-based approach to knowledge flow analysis: a case study of a manufacturing firm. Knowl. Process Manage. **10**(4), 260–276 (2003)
17. Gronau, N., Korf, R., Müller, C.: KMDL-capturing, analyzing and improving knowledge-intensive business processes. J. Univ. Comput. Sci. **11**, 452–472 (2005)
18. Heisig, P.: The GPO-WM® method for the integration of knowledge management into business processes. In: International Conference on Knowledge Management, Graz, Austria, pp. 331–337 (2006)
19. Zhaoli, Z., Zongkai, Y.: Modeling knowledge flow using petri net. In: IEEE International Symposium on Knowledge Acquisition and Modeling Workshop, China, pp. 142–146 (2008)
20. Woitsch, R., Karagiannis, D.: Process oriented knowledge management: a service based approach. J. Univ. Comput. Sci. **11**(4), 565–588 (2005)
21. Weidong, Z., Weihui, D.: Integrated modeling of business processes and knowledge flow based on RAD. In: IEEE International Symposium on Knowledge Acquisition and Modeling, China, pp. 49–53 (2008)
22. Supulniece, I., Businska, L., Kirikova, M.: Towards extending BPMN with the knowledge dimension. In: Bider, I., Halpin, T., Krogstie, J., Nurcan, S., Proper, E., Schmidt, R., Ukor, R. (eds.) BPMDS 2010 and EMMSAD 2010. LNBIP, vol. 50, pp. 69–81. Springer, Heidelberg (2010)
23. Businska, L., Kirikova, M.: Knowledge dimension in business process modeling. In: Nurcan, S. (ed.) CAiSE Forum 2011. LNBIP, vol. 107, pp. 186–201. Springer, Heidelberg (2012). doi:10.1007/978-3-642-29749-6_13
24. Businska, L., Supulniece, I., Kirikova, M.: On data, information, and knowledge representation in business process models. In: The 20th International Conference on Information Systems Development (ISD 2011), pp. 24–26. Springer, Edinburgh (2011)

25. Sultanow, E., Zhou, X., Gronau, N.: Modeling of processes, systems and knowledge: a multi-dimensional comparison of 13 chosen methods. Int. Rev. Comput. Softw. **7**(6), 3309–3319 (2012)
26. Liu, D.R., Lai, D.R., Liu, C.H., Chih-Wei, L.: Modeling the knowledge-flow view for collaborative knowledge support. J. Knowl. Based Syst. **31**, 41–54 (2012)
27. Netto, J.M, Franca, J.B.S., Baião, F.A., Santoro, F.M.: A notation for knowledge-intensive processes. In: IEEE 17th International Conference on Computer Supported Cooperative Work in Design, vol. 1, 1–6 (2013)
28. Ryle, G.: The Concept of Mind. The University of Chicago Press, Chicago (1949)
29. Bachimont, B.: Engagement sémantique et engagement ontologique: conception et réalisation d'ontologies en Ingénierie des connaissances». In: Charlet, J., Zacklad, M., Kassel, G., Bourigault, D. (eds.) Ingénierie des connaissances, évolutions récentes et nouveaux défis. Eyrolles, Paris (2000)
30. Maier, J.: Abilities. In: Zalta, E.N. (ed.) Stanford(2011)
31. Kassel, G.: A formal ontology of artefacts. Appl. Ontology **5**(3–4), 223–246 (2010)
32. Bonnet, C., Ghiglione, R., Richard, J-F.: Traité de Psychologie Cognitive, Tome 1: Perception, action, langage. Dunod, vol. 3. Paris, 280 p. ISBN 2100078445 (2003)
33. Fantl, J.: Knowledge how. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2012). http://plato.stanford.edu/archives/win2012/entries/knowledge-how/
34. Nonaka, I., Takeuchi, H.: Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, New York (1995)
35. OMG (2011). Unified modeling language (UML). Version 2.0. http://www.uml.org/
36. Turki, M., Saad, I., Gargouri, F., Kassel, G.: Towards identifying sensitive processes for knowledge localization. In: International Conference on Collaboration Technologies and Systems (CTS 2011), pp. 224–232 (2011)

Using Intelligent Systems to Improve Case Flow in Court Systems

Ana Lúcia Martins(✉)

Instituto Universitário de Lisboa (ISCTE-IUL),
Business Research Unit (BRU-IUL), Lisbon, Portugal
almartins@iscte.pt

Abstract. Courts of Law often take so long to produce final decisions. The reasons why court cases do not flow faster have already been approached, but literature in scarce in exploring how to enhance that flow. This paper aims at identifying ways to make legal cases flow faster through the legal process and increase courts' value proposition from a technological perspective. Research is based on a flow and value approach. Two procedures are compared, a national Civil Declarative Procedure and the European Small Claims Procedure. Court procedures are analyzed from the perspective of its process' activities and human intervention. Findings point that delays of court cases occurs mainly due to direct human intervention, particularly from the judge the case is assigned to, and that the activities involved can be performed based on pre-defined rules. Recommendations are produced on how to improve court procedures' customer value using intelligent systems.

Keywords: Process management · Court procedures · Value · Intelligent systems

1 Introduction

Courts of law as known to take longer than expected to produce results. This situation results from accumulation of cases in specific points in the court process [1, 2]. These critical points are located at activities that depend heavily on human interference [3].

Courts of law and their procedures are often analyzed from a legal perspective and this perspective is based on how the law is applied, on how the legal procedures are fulfilled and on the quality of the decisions produced. From a management perspective research is based on the enhancement of case flow and reduction of delays [1, 2, 4] and on streaming procedures and processes [5]. The management perspective is still scarce in terms of providing capacity improvements in bottleneck points in ways other than the increase of the number of human resources.

From the intelligent systems perspective it is possible to find positions defending that technological solutions can lead to better balanced resolution of disputes [6]. Nonetheless there are arguments highlighting underlying problems that need to be addressed before these technological solutions can fully replace human analytical and critical thinking capabilities, such as the complexity of the knowledge that needs to be

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_101

considered and the issue of the responsibility for the decision [7]. These contributions provide a hint on the possibility to add capacity to the problematic legal procedural points, but are focused on the technological perspective and lack the resource and process management impact.

Interface solutions between management and intelligent systems might provide some clues on how to enhance case flow without the immediate approach of increasing human resources. Can procedures be adjusted to allow more standardization in terms of the analysis of the cases? Can standardization provide robust solutions and value to customers? Can intelligent systems deal with specific requests from legal procedures?

The debate concerning the use of intelligent systems in the legal decision process making is still active and the impact of such systems on the management of flows in the courts has not yet been approached. The purpose of this research is to overcome this gap in literature by analyzing to what extent intelligent systems can be used in the context of legal procedures and to assess the managerial impacts of such use. As a consequence this research has three main objectives. The first objective is to perceive the premised and rhetorical perspective of the use of intelligent systems in the legal arena to create human-like decisions to legal disputes and improve court's value proposition. The second objective is to identify the adjustments required in the legal procedures to accommodate the use of intelligent systems. The last objective is to assess the impacts of intelligent systems on the flow of cases and the delays in the legal processes.

To illustrate the debate a national declarative procedure will be used as well as the European Small Claims Procedure. As a conceptual paper it starts with an analysis of the fundamental conceptual arguments of flow management in service processes and on the ability of the court system to create value. The conceptual advantages of the use of intelligent systems in the legal context are also approached. A brief description of both legal procedures is provided as ground for the following discussion. Finally, suggestions are provided on how intelligent systems can contribute to enhance case flow in legal processes as well as challenges that still need to be addressed in parallel from the managerial and the technological perspectives.

2 Theoretical Background

2.1 Flows in the Service Process

According to the Council of Supply Chain Management Professionals [8], logistics management is the part of the supply chain process that plans, implements, and controls the efficient, effective forward and reverse flow and storage of goods, services, and related information between the point-of-origin and the point-of-consumption in order to meet customer's requirements.

Services are provided using more or less pre-defined process and its activities are supported on resources. Services provide their output to its customers based on the capacity installed at each of the pools of resources (always limited) in each of the activities in the process and on how the flow is managed between the different activities. Services are characterized by having a level of demand that is not stable and each unit of demand may require different service times from the different resources. The

consequence is a flow of service requests that is uneven. Requests may accumulate between the activities, which allows improved efficiency in the use of resources but leads to poorer effectiveness of the overall service process.

All processes aim at creating value to its customers [9]. From a logistics perspective, of flow and process management, value is generated when customer satisfaction is achieved using the lowest total cost for the set of specific requests, which is to say that service benefits should be provided with cost minimization for the desired service level [10]. Value and logistics competitive positioning can be generated either by a service perspective or by a cost perspective [11]. Harrison and van Hoek [12] developed a more detailed approach to value from a logistics perspective by arguing that the hard goals of logistics are quality (supplying what the customer wants, which is translated as the quality of the service provided), time (a measurement of how long the customer has to wait to receive the service), and cost (the cost supported by the process to supply the requested service in a time frame that is acceptable by the end costumer). Soft logistics goals also need to be considered [12], which vary from service to service and are more difficult to measure, such as confidence (efficiency of the process or requests answered promptly) and security (confidentiality in customer's information treatment).

Service capacity is a perishable commodity [13], which leads to the need to balance the level of capacity installed with the demand that service experiences. In order to reduce service provision costs resources need to be used to the maximum level possible. At the same time if a high level of resource utilization and variation in demand is experienced long lead times are expected as the resources needed to supply the service are disputed by several requests simultaneously.

Process synchronization is a way to overcome variation in waiting times and a means to, simultaneously, allow faster flows and reduced costs [14]. Streamlined flexible flows are suggested to this purpose, matching supply with demand, and reducing variability as means to achieve reduction in variation [14]. Additionally, employee involvement and a continuous improvement approach are identified as fundamental for the success of the endeavor.

2.2 Value in the Court System

The value to be provided by a process has to be defined by its customer [15]. The court system, i.e., the overall aggregation of processes and operations through the different entities in the legal supply chain that come together to provide a solution to disputes, aims at servicing more than one customer. The parties involved in a dispute are the basic level of customers of the court system, but as this system is financially supported by government funding, the overall population are also customers. The overall population aims not only at processes that are efficient but also at processes that are effective. This effectiveness can be read both in terms of the correctness of the procedure and the quality of the decisions produced [5] and in terms of the use of the resources available to solve the disputes. Nonetheless, the speed of the cases through the process could also be pointed out as a relevant goal, as research show the interest of some legal stakeholders in this benefit [3, 16, 17]. From the perspective of each of the parties involved value is the maximization of its individual position, which leads to conflicting perspectives on

the outcome of the legal procedure (not only in terms of the contents of the decision but also in terms of how long the legal process takes to produce results). Each entity will act upon the process attempting to influence the timing and the content of the final decision in his favor. Overall, the value produced by the legal system is mostly based on the quality of the decisions and the assurance of justice, but also on the time needed to reach a final decision [5].

The court system is based on an overall operation consisting of several processes (here defined as a set of activities linked together for the purpose of supplying a service) with resources available at each node and flows that flow between them [18]. In order to increase capacity at each node costs will rise but a faster service could be provided.

In order to create value to its customers the court has to balance the flows of cases though the process focusing both on supplying the fastest possible service and minimizing the costs it has to support.

In legal procedures, the overall service process is pre-defined. Nonetheless there is always discretionary assessment by the service provider (in this case, the judge), which can influence the value produced [19]. This discretionary assessment limits the universal applications of the procedures at the same time that it allows dealing with the complexity of specific legal cases, i.e. on the one hand it promotes value by being flexible enough to deal with specific situations but on the other hand it limits value by reducing standardization and requiring more time to produce results. Dealing with this complexity requires additional time in some of the legal process activities. Nonetheless, if management practices are improved, such as scheduling practices, delays could be reduced [3, 17] leading to an improved value proposition by the courts.

2.3 Intelligent Systems in the Legal Process Decision Making

The introduction of intelligent system in the legal systems started to be discussed many decades ago. Although it is theoretically possible, in practice, due to the natural ambiguities of the law, it is not yet possible to be used massively in the court system.

There are reports of several projects that used intelligent systems to produce/facilitate/support decision making in the legal area [6]. These can be from Decision Support Systems that help judges in the decision making process by analyzing the inputs against the applicable law and past decisions; to online dispute resolution systems; to the application of game theory in maximizing the outcome for the several parties involved in a dispute (assuming that all parties aim at maximization of the overall outcome). All these applications are specific to very narrow areas and the outcome has to be validated, i.e., not only the outcome does not enforce the participants as it is voluntary; they act as facilitators for decision making [7]. These intelligent systems have to be feed with predefined explicit rules. The systems can even be supplied with knowledge modules to be able to deal with uncertainty, but such as the law can be interpreted in different ways and the same set of norms may end up being able to produce different outcomes, complexity and intentions would have to be considered, and the current intelligent systems are not yet able to accommodate such situations.

It is clear that overloaded courts take longer to produce results. Resources are always limited and are used for a multitude of different requests, many of them very repetitive and with little impact of value creation. The use of intelligent systems has been attempted in the legal area but always with limited scope. There is not yet research aiming at a more broad use of intelligent systems in courts of law. Aiming at filling this gap, the main research questions pursued in this research are:

RQ1: To what extent can the court system benefit from the use of intelligent systems?
RQ2: What are the expected impacts in the value propositions of court systems by the use of intelligent systems?

3 The Portuguese Civil Declarative Procedure

The Portuguese court system is continental style and is based on a set of predefined laws that are applied by the courts and its judges to solve disputes. This specific court system is divided in three levels, each acting as the appeal arena for the decisions from the previous level. The court system is an independent entity but works along with other entities to collect information on the accused or asking for technical advice/support.

Each court consists of several departments, each with a judge and its support assistants, hereby called "judge staff". Judge staff receives all requests to and from the judge, links the judge to the entities outside the court (parties involved or other entities), and schedules the cases that are sent to the judge on a daily basis.

The procedure is pre-defined and every activity in the process has an owner (the resource that supports processing in that specific activity). The declarative civil procedure is composed out of three sub-processes, each needing to be completed before the following one can be initiated: arguments (in which parties involved present their reasoning, which can be conflicting, and the outcome is a set of the specific arguments each party presented); definition (in which the judge assesses the arguments and evidence provided and defines what is in fact the nature of the dispute); verdict or conclusion (where a final decision is produced by the judge). Most of the decision making in terms of pathway in this procedure is of IF-THEN nature. Final decisions are of a different nature as many other factors have to be considered and do not fit that kind of approach (such as intention or motivation).

The judge staff, the judge and the lack of courtrooms (problems in terms of scheduling) are the causes of delays in this procedure [2], as well as difficulties in communication between entities in the legal service supply chain, the continuous participation of the judge in every decision concerning the case, "even trivial ones", the courtrooms availability as consequence of poor scheduling skills, and to judge staff batch mentality and shortage of understanding the case flow perspective [16]. On a broader level delays also emerge from overloaded judge staff and their lack of support from technology; from the pressure placed on the judges to make more cases flow faster and the consequence of delaying the most complex/time consuming ones; and from the insufficient management skills for case scheduling issues in both the use of courtrooms and case flow [3].

Although there are several reasons why this legal process takes so long to produce results, these can be summarized in lack of management skills and overload of human

resources which leads judges to focus on the most repetitive decisions and to delay/sacrifice the more complex ones in order to achieve a better overall flow of cases.

4 The European Small Claims Procedure

In order to overcome the time consuming and expensive national procedures used to solve civil disputes between parties from different European Union (EU) countries, the European Parliament along with the European Council developed the European Small Claims Procedure (ESCP) (set by Regulation (EC) No 861/2007 [20] and currently under revision). This can be seen as an alternative procedure to the Portuguese civil declarative one, in the case of disputes that include parties from different European Union countries when that dispute emerged in the Portuguese jurisdiction.

Aiming at disputes involving no more than 2000 euros between parties from different EU countries, this procedure has an online exchange platform based on standardized forms. All decisions are still based on a judge, but decisions making is less complex due to the use of the standardized forms and online document exchange. This allows a faster flow of cases through the process.

As the Portuguese civil procedure, the ESCP also has three sequential phases: commencement, conduct, and conclusion, each with similar aim as the Portuguese overall civil procedure. Systematized diagrams with the flow of the procedure [21, 22] show that the process has little possible alternative paths, all of them pre-defined. Although not yet very much used in EU as the awareness of its existence is not very high, it is possible to perceive that there are three critical points in this procedure that may lead to longer lead times as the complexity of the activities is higher: deciding if there is a counterclaim in the conduct phase; deciding if there is enough information to conclude; producing the final decision. Excluding these activities, and helped by the fact that the process is based on forms, all other activities are routinely decided based on pre-defined guidelines, which can be approached from an IF-THEN perspective.

The high level of standardization of the procedure allows stating that it is aimed to reduce waste [21] but at the same time some flexibility is also required for the complex activities and deciding, for instance, on the alternative to include more information to the conclusion phase as at that point in the procedure non-documental evidence might be required.

5 Discussion

5.1 Comparison of the Two Procedures

The civil national declarative procedure is an overall guideline with specific milestones that have to be fulfilled. Some parts of the procedure are common to all cases and there is discretionarily of the judge to allow specific cases to jump to stages further down in the procedure as long as specific conditions are fulfilled. There are activities in the procedure that involve high levels of complexity, but these are only a few (such as assessing the arguments from the parties and the sufficiency of evidence in the arguments

phase, the hearing, as well as issuing judgment and issuing the verdict). Most of the remaining activities (under the responsibility of the judge or the judge staff) are very standardized. The absence of standardized forms or even structure of documents increase the difficulty of analysis, and the complexity of the contents of some documents, which is most times introduced on purpose by the lawyers, can require much more additional time of analysis by the human resources at the court.

The ESCP is a procedure that is more standardized than the civil national procedure, involving medium complexity at assessing the correctness of the scope of the claim, and with higher level of complexity in analyzing the sufficiency of evidence, clarifying if a counterclaim actually exists, and issuing the judgment. Much of the complexity of the traditional national procedures was removed with the introduction of forms and with the online platform that allows immediate communication with the several parties involved.

In both cases the delays in the legal processes limit the flow of cases and reduce the value propositions of the courts where they are.

5.2 Alternatives Ways to Divert Demand from the Courts

The civil national declarative procedure used to be flooded with very simple cases (for instance traffic tickets, unpaid service bills) which delayed all cases. All of these involved the attention of a judge at least once. The use of this scarce resource on such routine and standardized decisions limited the time available for more complex cases, leading to longer waiting times in the legal process. As a consequence these simpler cases were channeled to an alternative path, outside the court (for instance the traffic tickets are automatically processed based on information from the traffic teams), reducing the work load on the court resources.

Additionally alternative dispute resolutions have already been rehearsed [16]. These considered solutions such as specialized courts (which allows more economies of scale and specialization) and mediation. The goal of these solutions is to reduce demand in the courts of law, allowing faster total processing times and more quality in the decisions produced.

The change in the way the process is fulfilled to comply with the legal procedure has to take into consideration those who relate to it (internally and externally). Standardization of forms and new ways to contact the court can present itself as a radical change in the way processes are conducted, which requires training to judge staff and the judge and time to allow cultural adjustment (for instance to deal with the dematerialization of cases). Outside the court adjustment can be easier as the introduction of automatized systems only relates to them in terms of the interfaces they need to use. At the same time, these interfaces might limit the access to justice (or force the use of lawyers) in situations that legally do not require a lawyer and less educated defendants and/or claimants are involved.

5.3 Impact of Intelligent Systems on the Procedures' Value Proposition

The use of technological solutions to deal with sub-processes requires their standardization or at least the standardization of some of its activities. The use of forms can help

feeding decisions support systems and automatize several parts of the procedures, but there are complex activities that cannot be overcome by the use of intelligent systems. These cannot interpret norms [6] nor decide upon intensions, which prevents them to be used in all process activities.

Nonetheless, the use of these systems in specific parts of the process or even in specific activities would allow the flow of cases to move faster in the administrative area (activities that are of judge staff responsibility) and even in some less complex decision making activities (that are responsibility of the judge). In parallel, the waiting times of the cases in the complex activities would be reduced as there would be more capacity available (measured in terms of the number of judge hours available) to analyze those cases.

The use of Decision Support Systems (DSS) shows itself as one of the most adequate solutions to support both judges and judge staff in decision making activities. Although the ESCP is already based on pre-defined forms it is possible to standardize the filling of those forms even more to adjust it, for the most current cases, in the choosing of pre-defined options that can be analyzed by the system to automatically produce suggestions of decisions (for instance, in terms of deciding the country to which the case is assigned to; to assess if the case is within scope of the procedure; to identify if a counterclaim exists and if the case is still within the scope of the procedure). The national declarative procedure is still very open in terms of the information inputs it accepts. Making these inputs more standardized (for instance by the development of pre-defined forms with specific information fields and on-line submission) would allow feeding DSS more easily and automatize some of its activities (for instance, assessing the scope of the call; calling the accused to the case; production of at least some parts of both the preliminary decision and the final decision). This solution is not intended to replace judges of staff, only to produce suggestions on the documents to be produced. Nonetheless, the time required per case would be less than before.

As each lawyer tends to use his specific way to report information concerning the legal cases, there might be resistance in the adoption of the more standardized forms, but as long as legal rules are issued resistance fades ways. As a consequence, the development of intelligent interfaces [6] would allow direct feed of DSS.

There are cases that, for its complexity, might require specific submission and/or analysis. These cases can be identified in advance and eventually obtain special permit to use alternative submission procedures, but this should be used only in exceptional situations pre-defined by law.

Although there are always limitations in the applicability of these solutions, they have been applied successfully in specific areas and/or sub-processes [6], which allow reducing the workload of the court resources (especially judges and other decision making resources) and therefore enhance case flow in the legal processes. The impact on the delays the cases experience through the procedure is expected to be positive as capacity is added to the bottleneck points by reducing average service time per case.

The value proposition of the court systems is a direct consequence of the performance of the court system and its processes. By using intelligent system in the court processes the time required to produce results would then be shorter and justice would be applied in a more standardized way to all situations. These intelligent systems would involve

cost, but the investment not only would be diluted through time but can also be considered as an alternative to the traditional solution of adding capacity to these systems by hiring more staff and/or judges. Although the value proposition of the court system would improve in terms of assurance of application of the law if these systems are introduced, the fact that there are conflicting values from the different parties in the dispute [5] makes the system less appealing to those attempting to use its lack of efficiency in their favor (reduce demand placed on courts). As a consequence, a prevention effect would emerge and reinforce justice.

By being able to reduce the time cases remain in the court system, intelligent systems contribute to a virtuous cycle of effectiveness at the level of the courts and of prevention of disputes.

6 Conclusion

From the previous discussion it is possible to conclude that it is possible to use intelligent systems in court processes, but its application is limited due to the fact the currently these systems lack the ability to deal with intentionality and different interpretation of the same set of legal norms. As a consequence, although it is not yet possible to use intelligent systems to produce human-like decisions, it is possible to use them to support decision making and enhance the value propositions of the different courts and their procedures. To do so standardization of procedures and adjusting the way the different actors of the court system relate to the court and the procedure has to be introduced to fit the specific requests of the intelligent systems.

References

1. Pekkanen, P., Pekkanen, J., Pirttila, T.: Justice system process as a supply chain – analysis of throughput-times, queues and capacity bottlenecks. In: 14th International Symposium on Inventories. Hungary, Budapest (2007)
2. Martins, A.L., Carvalho, J.C.: Creating value in a services supply chain. In: 10th International Symposium on Logistics, pp. 708–713 (2005)
3. Martins, A.L., Carvalho, J.C.: Sliming lead times in courts of law – a case study. Int. J. Ind. Eng. Man. **4**(3), 123–130 (2013)
4. Seppala, P., Kerkkanen, A., Pekkanen, P., Pirttila, T.: Applying the operations management – approach to reduce process delays in justice courts. Int. J. Bus. Ex. **6**(2), 131–147 (2013)
5. Martins, A.L.: Visiting legal systems using S-D logic lens: the impact on operations. In: 23rd EurOMA Conference (2016)
6. Carneiro, D., Novais, P., Andrade, F., Zeleznikow, J., Nevez, J.: Online dispute resolution: an artificial intelligence perspective. Artif. Intell. Rev. **41**, 211–240 (2014)
7. Kalinauskas, M.: Problematic aspects of the use of expert systems in law. Socialinés Technologijos **1**(1), 81–90 (2011)
8. CSCMP. http://cscmp.org. Accessed 16 Sept 2016
9. Davenport, T.H.: Process Innovation: Reengineering Work Through Information Technology. Harvard Business School Press, Boston (1993)
10. Bowersox, D.J., Closs, D.J., Cooper, M.B., Bowersox, J.C.: Supply Chain Logistics Management, 4th edn. McGraw-Hill International Edition, Boston (2012)

11. Rushton, A., Croucher, P., Baker, P.: The Handbook of Logistics & Distribution Management – Understanding the Supply Chain, 4th edn. Kogan Page, London (2014)
12. Harrison, A., van Hoek, R., Skipworth, H.: Logistics Management & Strategy – Competing Through the Supply Chain, 5th edn. Pearson, London (2014)
13. Fitzsimmons, J., Fitzsimmons, M., Bordoloi, S.: Service Management: Operations, Strategy, Information Technology, 8th edn. McGraw-Hill International Edition, Hardcover (2014)
14. Slack, N., Chambers, S., Johnston, R.: Operations Management, 6th edn. Prentice Hall/ Financial Times, Upper (2010)
15. Womack, J., Jones, D.: Lean Thinking, 2nd edn. Free Press Business, London (2003)
16. Hines, P., Martins, A.L., Beale, J.: Testing the Boundaries of Lean Thinking: Observations from the Legal Public Sector. Pub. Mon. Man. **28**(1), 35–40 (2008)
17. Pekkanen, P., Karppinen, H., Pirttilä, T.: What causes prolonged lead-times in courts of law? In: Reiner, G. (ed.) Rapid Modelling for Increasing Competitiveness - Tools and Mindsets, pp. 221–231. Springer-Verlag, London (2009)
18. Martins, A.L., Carvalho, J.C.: The court system and its dynamics – a case study. In: 9th International Symposium on Logistics, pp. 29–33 (2004)
19. Gronross, C.: Service logic revisited: Who creates value? And who co-creates? Eur. Bus. Rev. **20**(5), 298–314 (2008)
20. Regulation (EC) no 861/2007, Official Journal of the European Union, L 199, p. 1 of 31.7.2007
21. Storsjo, I., Martins, A.L., Zanoni, S.: Leanness and agility in the judicial supply chain. In: 22nd EurOMA Conference (2015)
22. Martins, A.L., Storsjo, I., Zanoni, S.: Lean justice. In: Netland, T., Powell, D. (eds.) The Routledge Companion to Lean Management, Chap. 31. Routledge, London (2016)

Similarity and Trust Metrics Used in Recommender Systems: A Survey

Maryam Jallouli[1(✉)], Sonia Lajmi[1,2], and Ikram Amous[1]

[1] MIRACL Laboratory, Technopole of Sfax, University of Sfax, P.O. Box 242, 3031 Sfax, Tunisia
Jallouli.maryam@gmail.com, slajmi@bu.edu.sa, ikram.amous@isecs.rnu.tn
[2] Al Baha University, Al Bahah, Saudi Arabia

Abstract. Recommender systems suggest the most appropriate items to users in order to help customers to find the most relevant items and facilitate sales. Collaborative filtering recommendation algorithm is the most successful technique for recommendation. In view of the fact that collaborative filtering systems depend on neighbors as the source of information, the recommendation quality of this approach depends on the neighbor's selection. However, selecting neighbors can either stem from similarity or trust metrics. In this paper, we analyze these two types of neighbor's selection metrics used in the field of recommendation in the literature. For each type, we first define it and then review different proposed metrics.

Keywords: Recommender system · Neighbor's selection · Similarity metric · Trust metric

1 Introduction

With the growth development of social networks such as Facebook and Twitter and video streaming web sites like Youtube and Dailymotion. Companies are more and more interested to publish their products in social networks to increase their profits. Their product, named item, can be of different sorts such as friends, movies, music, books, news, images, etc. In the other hand, users using social network continue to increase. For example, and according to a statistic[1], number of users in Facebook is 1,13 billion, where 1,03 billion of them log on every day. Consequently, recommender systems have then emerged as a way to offer for a specific user the most appropriate item.

In general, the process of recommendation consists of an input, a recommender system algorithm, and an output. These steps are shown in Fig. 1. Input can be represented by two steps. First, we obtain users history interactions towards items. How to represent it depends on the type of RS used, and it can be a vector, matrix or tensor. The

[1] http://www.journaldunet.com/ebusiness/le-net/1125265-nombre-d-utilisateurs-de-facebook-dans-le-monde/ (Juillet 2016).

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_102

typical input of most RS algorithms is represented as a matrix of ratings. It consists of an $m * n$ table, where each row represents a user, each column represents an item, and the intersection of a row and a column represents the user's rating value given to an item [1]. The absence of a rating value at this intersection indicates that user has not yet rated the item.

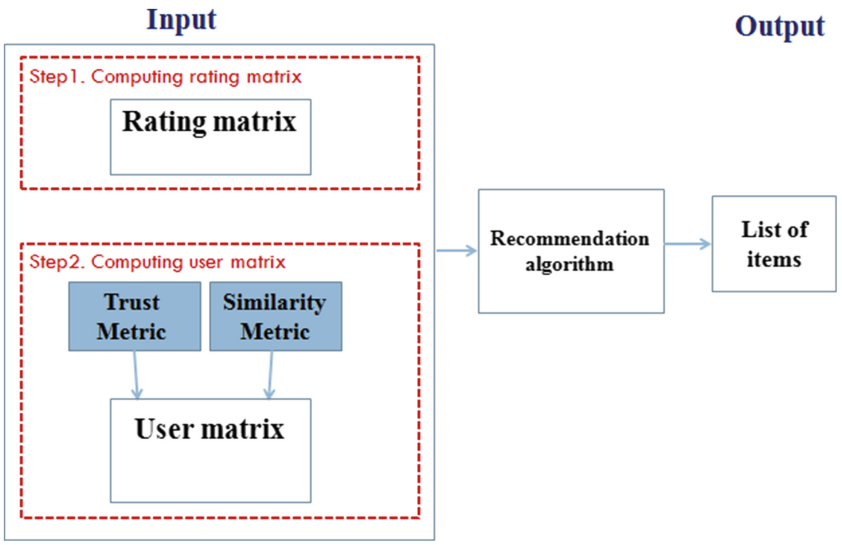


Fig. 1. General structure of a recommender system.

The second step is to calculate the distance between the target user and other users to find their nearest neighbors. The output of this step is an $m * m$ matrix, where m is the number of users and content of cell u, v presents the trust intensity between users u and v. In traditional approaches neighbors are chosen based on similarity, however latest approaches are based in trust.

Concerning the recommender system algorithms, the objective is to calculate the missing ratings in the rating matrix. In the last decade, several approaches have been proposed as recommender systems. These approaches can be divided into three distinct classes: there are content based, collaborative filtering and hybrid based approaches. Collaborative filtering algorithm (CF) is the most widely used technique to build recommender systems (RS) [1]. It consists of, for a target user u, to find his nearest neighbors from the user matrix, and according to his neighbor's interests, predict his interest for an item not yet appreciated. The rating is computed by a weighted average of the ratings by the neighbors. The CF success depends, however, critically upon the similarity metric used to find the most similar or trusted users (neighbors).

From the above process, we can find that the second step in the input, that is the most important step and our focus in this paper, is to obtain $m * m$ user matrix. According to Fig. 1, this matrix can be obtained by either similarity or trust metrics. In the next two sections, these two types of metrics are detailed. For simplicity, a number of notations

are introduced. $R \in \mathbb{R}^{m \times n}$ to represent the 2D rating-matrix, with user-item dimensions. Each value in this matrix, denoted by r_{uj}, represent the rate of user u, about item j. m and n denotes respectively the number of users and items. $T \in \mathbb{R}^{m \times m}$ is a 2-D trust matrix with user-user dimension. Each case in this matrix, denoted by t_{uv}, represents trust value between users u and v.

Our survey is organized as follow. First, we expose some similarity metrics used in this field. Secondly, we detail specific metric related to trust. For this aim, we define the trust notion in recommender system. Summarize, next, the four aspects of trust. Then, we review three approaches in this area categorized as: trust metrics based on (i) trust propagation, (ii) user's interactions and (iii) ratings vector.

Three main research contributions of this paper are summarized as follow: (1) we clarify the majority of terms used in this field and thus unveil ambiguities (2) propose a new categorization for trust metric, which can be based on trust propagation, user's interaction or user's rating. And Finally, (3) present a survey that presents different similarity and trust metrics unlike other surveys which present only one kind of approach.

2 Similarity Metrics

2.1 Definitions

In the field of RS, similarity values between two users are measured by observing all the items who have been rated to both users. There are a number of different mathematical formulations that can be used to calculate the similarity between two users. As can be seen in the formula below, each formula includes terms summed over the set of common users U. The cosine similarity and Pearson correlation coefficient are the most widely used metrics. Considering the ineffectiveness of these two measures, Bayesian similarity [2] is an alternative metric that is proposed for better similarity measure. In the following formulas, $sim(u, v)$ denotes the similarity value between users u and v.

2.2 Metrics Based on Vectors of Ratings

Cosine Similarity (COS). The earliest work using cosine similarity metric [3] for user-user CF is [4]. This similarity metric [3] represents each user by a vector which contains ratings of appreciated items, and defines the similarity between two users as the cosine value of the angle between these vectors. That's why it is also known as vector-based similarity. The cosine-based approach defines the cosine-similarity between two users u and v as demonstrated in Eq 1:

$$sim(u, v) = \cos\left(\vec{u}, \vec{v}\right) = \frac{\vec{u} \cdot \vec{v}}{\left\|\vec{u}\right\| \left\|\vec{v}\right\|}. \tag{1}$$

where "·" denotes the dot-product of the two vectors.

Pearson Correlation Coefficient (PCC). This formula [3] is based on how much the rating by common users for a pair of items deviate from average ratings for those items. The PCC of two users u and v is defined in Eq. 2:

$$sim(u, v) = \frac{\overrightarrow{u}.\overrightarrow{v}}{||\vec{u}||^2.||\vec{v}||^2}. \tag{2}$$

Bayesian Similarity (BS). This approach is designed by [2] taking into consideration both the direction and length of rating vectors. Therefore, the rating similarity between a pair of users is measured by excluding the chance correlation and user bias from the overall similarity as in Eq. 3:

$$sim(u, v) = \max\left(s'_{(u,v)} - s''_{(u,v)} - \delta\right). \tag{3}$$

Where $s'_{(u,v)}$ denotes the 'raw' similarity between two users u and v, $s''_{(u,v)}$ denotes the chance correlation and δ is a constant representing the general user bias.

The similarity metrics presented in this section are general. The next section is more specific. It detail the trust metrics used in the literature to improve recommender systems.

3 Trust Metrics

3.1 Definitions

There have been some proposed definitions of trust in the context of Recommender Systems [5, 6]. But, these definitions fall into various evidences, and a solid definition for it, in many cases can be quite elusive. In fact, we propose to categorise trust into trust propagation, user's interactions and ratings vector. According to each approach, definition of trust differs.

In RS, trust can be broadly classified into two types: explicit and implicit. Explicit trust refers to the situations where trust information is explicitly specified by users themselves. For example, users in FilmTrust [7] can declare others users as trusted neighbours. However, this information is usually not available due to the concern of privacy or because the users don't like to make much effort to fill information. Hence, the implicit methods are preferred. Efforts have been made to infer trust using computational methods rather than specified by users.

In the following formulas, $trust(u, v)$ denotes the trust value from user u to user v.

3.2 Properties

Four distinct properties that can be attributed to trust are contended in [8].

Asymmetry. Trust depends from user to another. It is personal and subjective. In fact, towards a same user, each user, according to his understanding to the target user, may

hold various opinions. For example, considering two users u and v, user u trusts user v does not mean that user v trusts user u.

Transitivity. In real life, people tend to trust friend of a friend rather than a stranger. Based on this assumption, transitivity means that if user u trusts v, and v trusts p, it can be concluded that u trusts p to some degree. This information can be inferred by propagating trust in social networks, to identity trusted friends.

Dynamicity. With the change of evidences or experiences over time, trust changes too. That is, it can be increased with positive feedbacks and decreased with negative feedbacks. So, trust is not stable or fix. Also, trust between two users is difficult to establish but easy to crash. That's mean, to form a solid trust, we need more evidences, but in the other way, to decrease trust we need just few evidences.

Context Dependence. Trust depends with context which we are most interested in. In fact, a situation where a user has to trust another user depends with respect to one specific situation, but not necessarily to another. Situation here means the context, such as time and location. For example, a user u can trust user v in the context of electronic, but not in the context of movie.

3.3 Metrics Based on Trust Propagation

In trust propagation methods, the system infers trust intensity between two users u and v based on non-empty value in the trust matrix. First, trust intensity between users contained in this matrix is presented in a graph, named trust graph. Based on the trust matrix, a trust graph is constructed. This graph is two dimensional, where nodes present users and edges presents trust value between users. In general, trust matrix is incomplete, that's mean the graph is not complete too. Therefore, trust propagation is one of the most important processes that calculate trust intensity on missed edges.

Some of trust metrics based trust propagation has been proposed in the literature. In this section, we demonstrate how Advogato, Appleseed, TidalTrust and MoleTrust compute the indirect trust between pairs of users and how the trust propagates in the trust network.

Advogato. The advogato maximum flow trust metric has been proposed by Levin [9] in order to discover the most trustworthy users. They take as input a seeds of trustworthy users and a number of users defined as base users. This metric uses a breath-first search algorithm. It uses the distance from the seeds denoted by D_u and the out-degree of nodes as in Eq. 4.

$$trust(u, v) = D_u. \tag{4}$$

Appleseed. Appleseed has been proposed by Ziegler [10] in his PhD thesis. His approach is similar to Advogato. But on it, the basic intuition of Appleseed is motivated by spreading activation mechanism. In other words, Advogato models trust as energy and propagates it using spreading activation models.

TidalTrust. Golbeck proposed TidalTrust [11] to compute trust prediction. He performs a modified depth first search for finding the highest trust users. Hence, to compute the trust value between user u and v, this metric aggregates the trust value between u's direct neighbors and v weighted by the direct trust values of u and its direct neighbors as in Eq. 5:

$$trust(u, v) = \frac{\sum_{w \in N_u} trust(u, w).trust(w, u)}{\sum_{w \in N_u} trust(u, w)}. \tag{5}$$

Where N_u denotes the set of neighbors of u. We should note here that $trust(w, u)$ will be computed recursively if the trust value is an indirect trust.

MoleTrust. Authors of [12] introduce MoleTrust as a trust metric. The underlying idea is similar to TidalTrust in the way that it also uses explicit trust. But the difference lies in the method of propagation. MoleTrust used breadth-first search while TidalTrust used depth-first search. The trust value from u to v is the aggregation of trust values between u and users directly trusting v weighted by the direct trust values as in Eq. 6.

$$trust(u, v) = \frac{\sum_{x \in N_u^-} trust(u, x).trust(x, u)}{\sum_{x \in N_u^-} trust(u, x)}. \tag{6}$$

Where N_u^- is the set of in-links of v. That's mean, N_u^- is the set of users for whom v is a direct neighbor.

3.4 Metrics Based on Users 'Interactions'

Mayer et al. [5] defined trust as the "*willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party*". In fact, users' interactions can be used to infer implicit trust. For example, if user v wrote a positive comments about many items rated by user u, then, user u will consider user v as an expert and then consider it as a trustworthy user. In this context, three most used and effective trust metrics based on users 'interactions' are elaborated. Each of them takes into consideration a group of factors and uses it as a criterion of selection to select trust users. These factors are defined according to each approach as follow:

Expertise and Preference Based-Trust (EPT). Denotes the approach proposed by Kim and Phalak [5]. The key idea of this work is that a user u would trust an expert v according to an area of interest of this user. For example, we consider user u is interested in movie according to his history of activities in a social network, and a user v has a good reputation on movie and electronic products because he writes many helpful comments. In this case, we can expect that user u will trust user v to some extent in the categorie of movie not in the electronic product category. Based on this assumption; Kim and Phalak had taken into consideration expertise (reputation) and preference (interest) as principle factors of trust value. Thus, the degree of trust between two users u and v is a function

based on the expertise of a content provider and the preference of a content user, as follow:

$$trust(u, v) = \frac{\sum_{k=1}^{C} p_{uk}.u_{vk}^{e}}{\sum_{k=1}^{C} p_{uk}}. \tag{7}$$

Where p_{uk} $(0 \leq p_{uk} \leq 1)$: the preference level of user u on category k, u_{vk}^{e} $(0 \leq u_{vk}^{e} \leq 1)$: the expertise of content provider j(u_j) on category k and C: the total number of categories. Functions of computing a content provider's expertise on a category and a content user's preference for a category are defined in their paper [5].

Trust Antecedent Framework (TAF). In [13], authors assume that trust is calculated as the result of three main trustee factors [5], including ability, benevolence and integrity and one trustor factor know as trust propensity. Ability determines skills and competence to deliver desired outcome, benevolence is the willingness to do good with the trustor and integrity is to adhere to a set of good moral principles. In the other hand, trust propensity means how easy a trustor trusts someone. Given the above factors, trust metrics is calculated as in Eq. 8.

$$trust(u, v) = i_{uv}.\bar{s}_{uv}.b_{uv}.x_v.p_u \tag{8}$$

where i_{uv} is an interaction intensity from u to v and $\bar{s}_{uv}$ is average rating v received from u. i_{uv} and $\bar{s}_{uv}$ determines ability factor. b_{uv} determines benevolence feature. x_v measure integrity value and p_u is the normalized trust propensity factor of the trustor. Formula of each factor is well defined in their paper [5].

Extended Trust Antecedent Framework ETAF. In [15], author have used the four generic trust factors used by [13] but in a different way. However, the formalization of each trust factor in separated, according to [15] into two perspectives: local and global. Local trustworthiness means that the trust factors are implemented using the direct interactions between two users, as in [13], whereas global trustworthiness indicates that trust factors are modeled based on the interactions of all users, with the target trustor or trustee. Both kinds of trustworthiness are taken into consideration by [15] and the formalization of their approach is as follow:

$$trust(u, v) = \left(\alpha.lt_{vu}(u, v) + (1 - \alpha).gt(v)\right).p_u. \tag{9}$$

Where $lt_{vu}(u, v)$ and $gt(v)$ are the local and global trustworthiness of user v from the perspective of user v. Formulas of these metrics are presented in detail in [15].

3.5 Metrics Based on Vectors of Ratings

Since our focus here is to infer trust from user ratings, the most appropriate definition of trust in this case was proposed by Guo [6]: "*Trust is defined as one's belief towards*

the ability of others in providing valuable ratings". In this section we review different trust metrics based on vectors of rating used in the literature.

k-Nearest Recommenders (KNR). Lathia [16] compute a predicted trust value by considering ratings of other users. In particular, a user is considered more trustworthy when he provides even opposite ratings than the one who is not willing to share opinions. Here, trust is obtained by the average of provided values over all the rated items.

$$trust(u,v) = \frac{1}{|I_{uv}|}\sum_{i\in I_{uv}}(1 - \frac{|r_{ui} - r_{vi}|}{r_{max}}). \tag{10}$$

Where $I_{uv} = I_u \cap I_v$ defines the set of items rated by both users u and v, and r_{max} is the maximum rating scale.

Pearson Correlation Coefficient (PCC). In [17], Papagelis adopt the PCC metric to first compute user similarity as follow:

$$sim(u,v) = \frac{\sum_i \left(r_{ui} - \bar{r}_u\right)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_i \left(r_{ui} - \bar{r}_u\right)^2}\sqrt{\sum_i (r_{vi} - \bar{r}_v)^2}}. \tag{11}$$

In a second step, they define θ_s and θ_I as the threshold of the user similarity and the number of co-rated item, respectively. Thus, trust value between two users u and v is defined as:

$$trust(u,v) = \left\{ \begin{array}{c} sim(u,v), if\ sim(u,v) > \theta_s, |I_{uv}| > \theta_I \\ 0, otherwise \end{array} \right\}. \tag{12}$$

Trust-Based CF. In [18], Hwang and Chen predict a rating by using a simple version of Resnick's prediction formula, based only on a single user:

$$p_{ui} = \bar{r}_u + \left(r_{vi} - \bar{r}_v\right). \tag{13}$$

Where $\bar{r}_u$ and $\bar{r}_v$ refers to the mean ratings of users u and v, respectively. Then, the trust value is calculated by averaging the prediction error on co-rated item:

$$trust(u,v) = \frac{1}{|I_{uv}|}\sum_{i\in I_{uv}}(1 - \frac{|p_{ui} - r_{ui}|}{r_{max}}). \tag{14}$$

By analogy to their work, [19] compute prediction using Resnick's formula prediction, but compute trust based on mean squared distance (MSD):

$$trust(u,v) = \frac{|I_{uv}|}{|I_u \cap I_v|}\left(1 - \frac{1}{|I_{uv}|}\sum_{i\in I_{uv}}\left(\frac{p_{ui} - r_{ui}}{r_{max}}\right)^2\right). \tag{15}$$

In their approach, they consider that if $trust(u, v) > \theta$, where θ is a predefined threshold, then these two users are considered as trusted neighbors.

4 Conclusion

In this paper, we discussed on process of making recommendation in CF approaches. This process is categorized into three steps: Input, an algorithm of recommendation and output. Furthermore, the second step in the input concerns user matrix calculation. Two distinct metrics can be considered to obtain this matrix: similarity and trust metrics. This paper proposed an empirical study of these two types of metrics. To this end, similarity and trust metrics are first defined and then well reviewed in detail. For future work, we intend to present in detail the second step in the process of making recommendation, named recommendation algorithm.

References

1. Gediminas, A., Alexander, T.: Toward the next generation of recommender systems: a survey of the state of the art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**, 734–749 (2005)
2. Guibing, G., Jie, Z., Neil, Y.S.: A novel Bayesian similarity measure for recommender systems. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI), pp. 2619–2625 (2013)
3. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Addison Wesley, Boston (2006). Chap. 2
4. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 43–52 (1998)
5. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manag. Rev. **20**, 709–734 (1995)
6. Guo, G.: Integrating trust and similarity to ameliorate the data sparsity and cold start for recommender systems. In: Proceedings of the 7th ACM Conference on Recommender Systems (RecSys) (2013)
7. Jennifer, G., James, H.: Filmtrust: movie recommendations using trust in web-based social networks. In: Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC), vol. 1, pp. 282–286 (2006)
8. Christiano, C., Rino, F.: Trust Theory: A Socio-Cognitive and Computational Model. Wiley, Hoboken (2010)
9. Raph, L., Alexander, A.: Advogato's Trust Metric (2002). http://advogato.org/trust-metric.html
10. Ziegler, C.-N.: Towards decentralized recommender systems. Ph.D. thesis, University of Freiburg (2005)
11. Jennifer, A.G.: Computing and applying trust in web-base social networks. Ph.D. thesis (2005)
12. Paolo, M., Paolo, A.: Trust-aware recommender systems. In: Proceedings of the 2007 ACM Conference on Recommender Systems, pp. 17–24 (2007)
13. Roger, C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manage. Rev. **20**(3), 709–734 (1995)

14. Young, A.K., Rasik, P.: A trust prediction framework in rating-based experience sharing social networks without a web of trust. Inf. Sci. **191**, 128–145 (2012)
15. Nguyen, V.A., Lim, E.P., Jiang, J., Sun, A.: To trust or not to trust? Predicting online trusts using trust antecedent framework (2008)
16. Guinbing, G., Jie, Z., Daniel, T., Neil, Y.S.: ETAF: an extended trust antecedents framework for trust prediction. In: International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 540–547 (2014)
17. Guinbing, G.: Integrating trust and similarity to ameliorate the data sparsity and cold start for recommender systems. In: 7th ACM Conference on Recommender Systems (RecSys) (2013)
18. Lathia, N., Hailes, S., Capra, L.: Trust-based collaborative filtering. In: Karabulut, Y., Mitchell, J., Herrmann, P., Jensen, C.D. (eds.) IFIPTM 2008. ITIFIP, vol. 263, pp. 119–134. Springer, Heidelberg (2008). doi:10.1007/978-0-387-09428-1_8
19. Papagelis, M., Plexousakis, D., Kutsuras, T.: Alleviating the sparsity problem of collaborative filtering using trust inferences. In: Herrmann, P., Issarny, V., Shiu, S. (eds.) iTrust 2005. LNCS, vol. 3477, pp. 224–239. Springer, Heidelberg (2005). doi:10.1007/11429760_16
20. Hwang, C.-S., Chen, Y.-P.: Using trust in collaborative filtering recommendation. In: Okuno, H.G., Ali, M. (eds.) IEA/AIE 2007. LNCS (LNAI), vol. 4570, pp. 1052–1060. Springer, Heidelberg (2007). doi:10.1007/978-3-540-73325-6_105
21. Qusai, S., Jie, L.: A trust-semantic fusion-based recommendation approach for e-business applications. Decis. Support Syst. **54**(1), 768–780 (2012)

Analysing the Performance of a Tomographic Reconstructor with Different Neural Networks Frameworks

Sergio Luis Suárez Gómez[1], Carlos González Gutiérrez[1], Jesús Daniel Santos Rodríguez[2(✉)], María Luisa Sánchez Rodríguez[2], Fernando Sánchez Lasheras[3], and Francisco Javier de Cos Juez[1]

[1] Department of Exploitation and Exploration of Mines, University of Oviedo, Oviedo, Spain
{suarezsergio,gonzalezgcarlos,fjcos}@uniovi.es
[2] Department of Physics, University of Oviedo, Oviedo, Spain
{jdsantos,mlsr}@uniovi.es
[3] Department of Construction and Manufacturing Engineering, University of Oviedo, Gijón, Spain
sanchezfernando@uniovi.es

Abstract. Correction of atmospheric turbulences with the use of guide stars as reference, is one of the most relevant issues of adaptive optics (AO). This is addressed with tomographic techniques such as Multi-object adaptive optics (MOAO). Next generations of extremely large telescopes, will require improvements in computational capabilities of real time control systems. An improved version of CARMEN, a tomographic reconstructor based on machine learning, is presented here. The performing time of two dedicated neural network frameworks, Torch and Theano, is compared, with significant improvements on the training and execution times of the neural networks due to calculations on GPU. Also, the differences between both frameworks are discussed.

Keywords: Neural networks · Torch · Theano · Adaptive optics

1 Introduction

In recent times, Adaptive Optics has become mandatory for stellar observation when it is performed with grounded telescopes, since the atmosphere produces aberrations in the light that passes through it. A first approach is to measure the distortions on the incoming wave-front and then correcting it by calculating the appropriate shape of a system of deformable mirrors to compensate, with the highest precision, the aberrations in the wave-front. This process has to be done as fast as possible due to the rapid changing nature of the atmosphere [1]. With the aim of measuring the wave-front more accurately, guide stars are used as reference. Natural stars can be used when they are in the field of view of the object of interest, or even artificial stars that are created by laser scattering in the upper atmosphere. For correcting the error from the atmospheric turbulences, computer tomography techniques are considered to control deformable

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_103

mirrors with data coming from guide stars [2, 3] and obtaining high quality astronomical images.

Different algorithms have been created specifically for their use in AO, like the Least Squared (LS) type matrix vector multiplication [4], the Learn and Apply (L&A) method [5], and the Complex Atmospheric Reconstructor based on Machine lEarNing (CARMEN) [6], developed at the University of Oviedo. CARMEN is a reconstructor for Multi-object adaptive optics (MOAO). It was initially created using regression techniques as MARS [7], with promising results [2]. However, the use of machine learning techniques such as Supporting Vector Machines (SVM) [8, 9] or Artificial Neural Networks (ANN) [10–12], has been very successful in different fields, which led us to create a solution based on ANN. The neural network is trained off-line with a large range of possible turbulent layer positions and therefore does not require any input of the optical turbulence profile. CARMEN has shown also excellent results working on-sky [13].

The latest advances in optical systems motivate the creation of large grounded telescopes, like the future 39 m diameter European Extremely Large Telescope (E-ELT). One of the challenges of ELT is the increase of number of guide stars, wave-front sensors and deformable mirror size, that brings about an enormous increase of income data and therefore a huge computational capability is needed [14, 15]. ANN architecture allows its parallelism in neural processing. The use of Graphics Processor Units (GPUs) provides a solution to this problem, due to the parallelization of different calculations, and therefore, the speeding up of the processing times. We have previously presented an initial approximation to this solution, obtaining a significant speed-up in the execution and training of the neural network [16]. The performance of two different frameworks is analysed in this paper, with the aim of obtaining the fastest results.

The main purpose of this paper is to compare the implementation of CARMEN in two different neural networks frameworks based on GPU, and expose their training and execution times. Next section shows how AO systems work, and on the following one the architecture of CARMEN is shown as well as a small description of the frameworks used. The experiment is defined afterwards and different variables to compare both frameworks are proposed. Finally, a discussion on results is presented, along with conclusions and some insights in future research.

2 Adaptive Optics Systems

The Shack-Hartmann Wave-front Sensor (SH WFS) is commonly used in astronomy to characterize an incoming wave-front. It consists of an array of lenses with the same focal length (called lenslets) each focused on a photon sensor. The incoming wave-front is divided into discrete areas and the local tilt of each lenslet can be measured as the deviation of the focal spot of the sensor from the positions due to a plane wave-front as shown in Fig. 1, creating a matrix of tilts that characterize the wave-front aberration. With this data, the aberration can be approximated in terms of Zernike Polynomials.

CANARY is an adaptive optics (AO) on-sky demonstrator, principally intended for developing and testing AO concepts for the future 39 m European Extremely Large

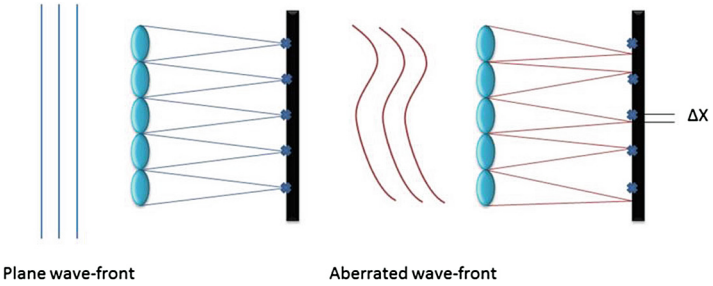


Fig. 1. Measurement of wave-front tilts

Telescope. It is operated on a Nasmyth platform of the 4.2 m William Herschel Telescope, one of the Isaac Newton Group of Telescopes (ING) of the Observatorio del Roque de los Muchachos (ORM), La Palma, Canary Islands, Spain.

There are severals configurations available:

- CANARY Phase B1 is designed to perform observations with 1 Laser Guide Star (LGS), and up to 4 Natural Guide Stars (NGS). It has a Shack Hartman Wave-front Sensor with 7 × 7 subapertures, although only 36 of them are functional.
- CANARY Phase C2 is designed for the study of Laser Tomography AO (LTAO) and Multi-Object AO (MOAO). There are 4 Rayleigh Laser Guide Stars, each with a 14 × 14 subaperture Shack Hartman Wave-front Sensor, where we only have 144 of them working.
- DRAGON aims to replicate CANARY concepts, to provide a single channel MOAO system with a woofer-tweeter DM configuration, 4 NGSs and 4 LGSs each with 30 × 30 subapertures. In this case, DRAGON is still a prototype, so we are going to use the worst scenario where all the subapertures are functional, which gives as a total of 900 subapertures per star.

3 CARMEN Architecture

CARMEN is a tomographic reconstructor based on artificial neural networks, whose architecture is a multi-layer perceptron, with a single hidden layer. It is composed by two fully-connected layers, where each neuron is connected to all the neurons in the previous layer. The output of each neuron follows (1), where w is the weight of each connection, x is the value of the neurons in the previous layer, b is a constant value called bias, and f is an activation function.

$$Y = f\left(\sum_{i=0}^{n} (w_i \cdot x_i) + b\right) \tag{1}$$

The number of input, hidden and output neurons, is directly related to the optical instrumentation used, since it changes with the number of subapertures of the device. It is not the purpose of this paper to analyse the net size or quality, since it has been

previously tested [17]. The number of neurons in the hidden layer will be equal to the number of neurons in the input layer. Also, the number of input and hidden neurons, depends on the number of stars (both natural and artificial). This means that the size of the network is highly variable, and should be adapted to each observation. However, and to make it easier to compare, we are going to use the three configurations explained in the previous section, and assume that we can observe both natural and laser guide stars.

These rules make that the number of input neurons matches the number of functional subapertures multiplied by 2 -due to the two dimensional input in the lenslet array-, and also multiplied by the number of reference stars. The corrected deviations of the lenslet array will appear in the output, which is the number of functional subapertures multiplied by 2. The final size of the network can be summarized in Fig. 2.

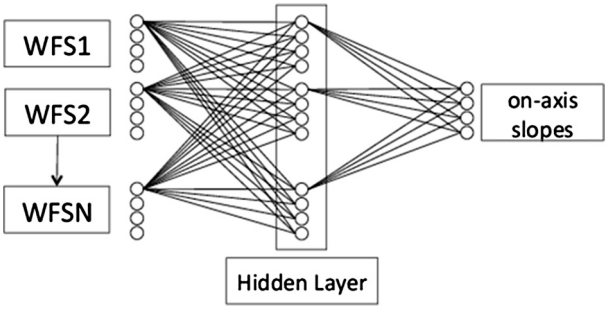


Fig. 2. CARMEN architecture

A network with 1 LGS and 2 NGS will be used in the case of CANARY Phase B1, that is, 216 input neurons. For CANARY Phase C2, with 4 Laser Guide Stars, there will be 1152 input neurons. In both cases the training data will be obtained from the CANARY simulator. Regarding DRAGON, the situation is different, due to the fact that it is still under development and there are still no simulated data, so the number of natural stars is unknown. In this case, laser guide stars will be used, and random data will be generated to train the network. This means there are 7200 input neurons for DRAGON. In Table 1 we can see a summary of the sizes of the different networks.

Table 1. Size of neural networks for adaptive optics systems

Name	Size	Number of training samples
CANARY-B1	216-216-72	350 k
CANARY-C2	1152-1152-288	1.5 M
DRAGON	7200-7200-1800	1 M

3.1 Neural Networks Frameworks

The growing popularity of Artificial Neural Networks in recent years, has caused the emergence of numerous frameworks, which help researchers in the use and development of more complex neural networks, without the need of hard programming efforts. There is a long list of existing neural networks frameworks[1], but as a first approach only Torch and Theano are considered.

Torch is a scientific computing framework with wide support for machine learning algorithms. It is written in Lua and C/CUDA to provide a fast execution and the possibility of importing modules to complete and accelerate the system.

Theano is a Python library developed to help researchers in their deep learning investigations [18]. It is developed at the University of Montreal, and uses some common python libraries such as numpy and scipy. It also provides GPU support to speed up training and execution.

4 Experiment Description

Following the procedure from previous work [16], two different measurements are considered. On one hand, training times will be compared for the frameworks here considered. Also, we take into consideration the execution time of the networks, which is crucial in the adaptive optics systems.

4.1 Training Benchmark

Despite the wide range of parameters to fit when training a neural network, only a few of them affect directly the training time. To simplify the comparison, we will only consider parameters that are especially relevant for the different systems.

For all the considered cases, Stochastic Gradient Descent (SGD) will be implemented in Torch and Theano, with mini-batches and momentum. Two parameters are needed for this method: the learning rate and the momentum. Although these variables are crucial regarding the quality of the resulting network, they have no influence in the training times, so it will be assumed that both are optimized to achieve the best result as possible.

Another critical parameter is the size of the training data. However, is easy to notice that in a training method based on mini-batches, the time grows proportionally to the number of samples employed. Keeping this in mind, it is easy to calculate how much time will take a network to be trained when changing the training dataset. This idea makes the choice of the dataset size irrelevant for benchmarking purposes, although is crucial for obtaining a good network. In this case, the size of the training data is specified in Table 1.

Three different networks will be used, one per optical system, as shown in Table 1. The main parameters to be changed will be the network size and the size of the

[1] http://deeplearning.net/software_links/.

mini-batch, in order to compare the times. The number of samples in the mini-batch will be 16, 32, 64, 128 and 256.

First the initialized weights and bias are copied to the GPU's VRAM[2], and all the training dataset is loaded in the main RAM, before starting the training. Then a timer starts and the program begins to copy the first mini-batch from RAM to VRAM, and then performs the calculation loop over the entire dataset. This operation is repeated for 20 epochs, while measuring its computation time individually. This procedure provides an average time for each epoch, and more reliable results, making it possible to see if there are significant time variations among different epochs.

4.2 Execution Benchmark

In this case, the same networks defined in Table 1 are used. Due to important differences with the training benchmark arise, it is needed to give a detailed explanation.

First, the net is fed with a single input, instead of using mini-batches, simulating what happens in a real telescope. As the execution program is expected to be included in the telescope management system, is a fair assumption that all the variables are already initialized, and the weight matrices copied into the VRAM. The loading of input data from the SSD will be taken into account in the time measurement, as well as the copy from VRAM to the system RAM, and the writing process to the disk. It was not possible to measure Theano execution when the input is already loaded in RAM and not in VRAM, so we have included loading and writing data from the main disk in the comparison. Each input is found in a separate file, in h5 format, and the output is written in a separate file. The system is fed with 10,000 inputs, one at a time that allows us to promediate the execution time. We will compare the average time of the different frameworks, looking for which one is the fastest.

4.3 Experiment Equipment

The experiments are performed on a computer running on Ubuntu LTS 14.04.3, with Intel Xeon CPU E5-1650 v3 @ 3.50 GHz, 128 Gb DDR4 memory, Nvidia GeForce GTX TitanX, and SSD hard drive. We used CUDA 7.5, cuDNN v3, Theano v0.7 and Torch with commit ID b33ac90.

5 Results and Discussion

The obtained results from training and execution process are shown in this section. Times will be split through different adaptive optics systems, and possible explanations for the different results will be analysed.

[2] VRAM: GPU Video RAM
RAM: CPU RAM.

5.1 Canary-B1

As explained above, the network size is 216-216-72, and 350 k samples are used for training. As it is shown in Fig. 3, increasing the size of the batch has an obvious improvement on training times. For small batch sizes, Torch has better performance than Theano. However, when the size increases, Theano becomes much faster, since it completes one epoch in about a second.

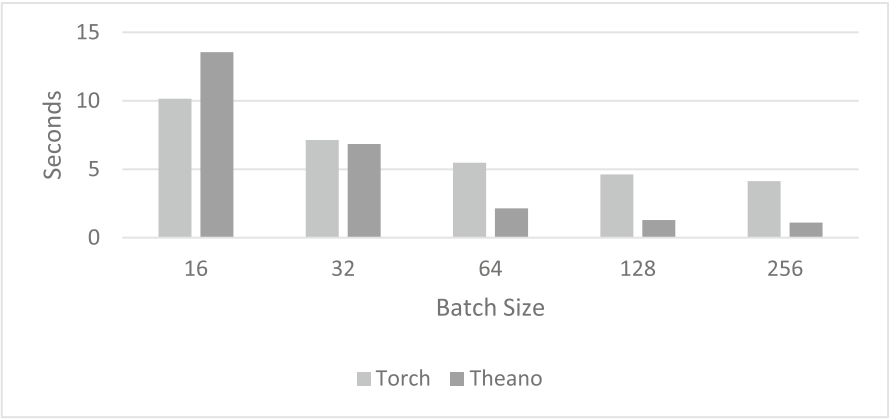


Fig. 3. CANARY-B1 seconds of training per epoch

5.2 Canary-C2

For CANARY-C2, the network size is 1152-1152-288, and use 1.5 M training samples. As Fig. 4 shows, increasing the batch size has more impact reducing the time, due to the bigger size of the network. For this optical system, Theano is always faster than Torch, although the difference for larger batch sizes, is smaller than in CANARY-B1.

5.3 Dragon

For DRAGON, the largest network is employed, with 7200-7200-1800 neurons and 1 M samples for training. In this case, Fig. 5 is similar to the one shown in CANARY-C2, where increasing the batch size has a great impact lowering training times. Also, Theano is always faster than Torch for all the situations, although the difference became smaller for larger sizes.

5.4 Execution Times

Regarding execution times shown in Table 2, we can see a different scenario than the one observed in the training process. In this case, Torch is always faster than Theano, around 10–20%, for the three optical systems.



Fig. 4. CANARY-C2 seconds of training per epoch

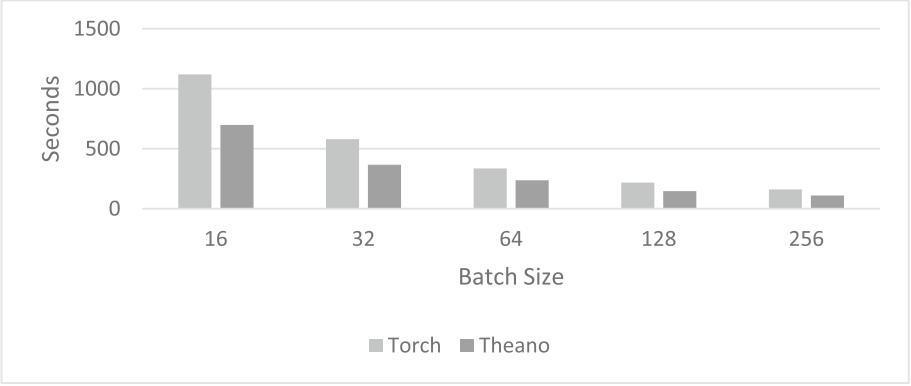


Fig. 5. DRAGON seconds of training per epoch

Table 2. Execution times

	CANARY-B1	CANARY-C2	DRAGON
Torch	0.7077 ms	0.9346 ms	3.1777 ms
Theano	0.8182 ms	1.1604 ms	3.5818 ms

5.5 Discussion

For most of the analysed scenarios, Theano provides a much faster training time. In particular, for CANARY-B1, the improvement over Torch for bigger batch sizes is huge, and it could be useful for on-sky training. However, it is interesting to notice how Theano has low performance for small batch sizes. This could be explained by the fact that small workloads are less optimized to work on GPU than in the case of Torch.

Regarding execution times, Torch is always faster than Theano. Although this behavior could seem odd compared with the results obtained before, it can be explained by a faster implementation of the execution process, but a much more slower execution of the backpropagation process.

6 Conclusions and Future Lines

Using a GPU based framework to train and execute the network, provide us a powerful tool to test and improve our reconstructor. However, as results show above, there is no ideal solution, and we have to keep in mind that depending on the situation, could be more appropriate the use of Torch or Theano. Also, the creation of a benchmark for our neural network is useful to repeat the experiment periodically, and for checking if the improvements in the different frameworks have a substantial impact in the results.

As this is a work in progress, we still have a lot of different challenges to solve. One of them is checking the performance of different frameworks, such as TensorFlow or Caffe, or even developing the code directly in C/CUDA, and compare the training and execution times for all the different adaptive optics systems.

The use of neural network frameworks, allow us to think about using convolutional neural networks or recurrent networks, which could provide some boost regarding the quality of the reconstruction. Although this could be done programming the code directly in any programming language, the use of Torch, Theano, or any of the existing frameworks could make much easier to find a new architecture that could help CARMEN.

In the near future, some challenges are expected, as huge increments of inputs and outputs during the next years, with new instruments as the E-ELT in development. This will make that the training process of the networks will require much more time. Also it will imply problems regarding the amount of GPU memory needed to deal with bigger weights matrices. A possible workaround for this is the training and execution of the network with the help of multi-GPU systems, where the computation and the memory could be split across different graphics cards.

References

1. Guzmán, D., Juez, F.J.D.C., Myers, R., Guesalaga, A., Lasheras, F.S.: Modeling a MEMS deformable mirror using non-parametric estimation techniques. Opt. Express **18**(20), 21356–21369 (2010)
2. Guzmán, D., de Cos Juez, F.J., Lasheras, F.S., Myers, R., Young, L.: Deformable mirror model for open-loop adaptive optics using multivariate adaptive regression splines. Opt. Express **18**(7), 6492–6505 (2010)
3. de Cos Juez, F.J., Sánchez Lasheras, F., Roqueñí, N., Osborn, J.: An ANN-based smart tomographic reconstructor in a dynamic environment. Sensors (Basel) **12**(7), 8895–8911 (2012)

4. Ellerbroek, B.L.: First-order performance evaluation of adaptive-optics systems for atmospheric-turbulence compensation in extended-field-of-view astronomical telescopes. J. Opt. Soc. Am. A **11**(2), 783 (1994)
5. Vidal, F., Gendron, E., Rousset, G.: Tomography approach for multi-object adaptive optics. J. Opt. Soc. Am. A. Opt. Image Sci. Vis. **27**(11), 253–264 (2010)
6. Osborn, J., et al.: Using artificial neural networks for open-loop tomography. Opt. Express **20**(3), 2420–2434 (2012)
7. Nieto García, P.J., Alonso Fernández, J.R., Sánchez Lasheras, F., de Cos Juez, F.J., Díaz Muñiz, C.: A new improved study of cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain) using the MARS technique. Sci. Total Environ. **430**, 88–92 (2012)
8. de Andrés, J., Sánchez-lasheras, F., Lorca, P., Javier, F., de Cos Juez, F.J.: A hybrid device of Self Organizing Maps (SOM) and Multivariate Adaptive Regression Splines (MARS) for the forecasting of firms' bankruptcy. Account. Manag. Inf. Syst. **10**(3), 351–374 (2011)
9. Suárez Sánchez, A., Riesgo Fernández, P., Sánchez Lasheras, F., de Cos Juez, F.J., García Nieto, P.J.: Prediction of work-related accidents according to working conditions using support vector machines. Appl. Math. Comput. **218**(7), 3539–3552 (2011)
10. VilánVilán, J.A., Alonso Fernández, J.R., García Nieto, P.J., Sánchez Lasheras, F., de Cos Juez, F.J., Díaz Muñiz, C.: Support vector machines and multilayer perceptron networks used to evaluate the cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain). Water Resour. Manag. **27**(9), 3457–3476 (2013)
11. Álvarez Menéndez, L., de Cos Juez, F.J., Sánchez Lasheras, F., Álvarez Riesgo, J.A.: Artificial neural networks applied to cancer detection in a breast screening programme. Math. Comput. Model. **52**(7), 983–991 (2010)
12. Lasheras, F.S., De Cos Juez, F.J., Suárez Sánchez, A., Krzemień, A., Fernández, P.R.: Forecasting the COMEX copper spot price by means of neural networks and ARIMA models. Resour. Policy **45**, 37–43 (2015)
13. Osborn, J., et al.: Open-loop tomography with artificial neural networks on CANARY: on-sky results. Mon. Not. R. Astron. Soc. **441**(3), 2508–2514 (2014)
14. Ramsay, S.K., Casali, M.M., González, J.C., Hubin, N.: The E-ELT instrument roadmap: a status report, p. 91471Z (2014)
15. Basden, A.G., et al.: Experience with wavefront sensor and deformable mirror interfaces for wide-field adaptive optics systems. Mon. Not. R. Astron. Soc. **459**(2), 1350–1359 (2016)
16. González-Gutiérrez, C., Santos-Rodríguez, J.D., Díaz, R.Á.F., Rolle, J.L.C., Gutiérrez, N.R., Cos Juez, F.J.: Using GPUs to speed up a tomographic reconstructor based on machine learning. In: Graña, M., López-Guede, J.M., Etxaniz, O., Herrero, Á., Quintián, H., Corchado, E. (eds.) ICEUTE/SOCO/CISIS-2016. AISC, vol. 527, pp. 279–289. Springer, Heidelberg (2017). doi:10.1007/978-3-319-47364-2_27
17. Osborn, J., et al.: Open-loop tomography using artificial nueral networks. Adapt. Opt. Extrem. Large Telesc. II (2011)
18. Al-Rfou, R., et al.: Theano: a Python framework for fast computation of mathematical expressions, p. 19 (2016)

Robot Swarms Theory Applicable to Seek and Rescue Operation

Jose León[1(✉)], Gustavo A. Cardona[3], Andres Botello[2], and Juan M. Calderón[1,2]

[1] Department of Electronic Engineering, Universidad Santo Tomás, Bogota, Colombia
joseleonl@usantotomas.edu.co
[2] Department of Electrical Engineering, University of South Florida, Tampa, FL, USA
{abotello,juancalderon}@mail.usf.edu
[3] Department of Electrical and Electronics Engineering, Universidad Nacional de Colombia, Bogota, Colombia
gacardonac@unal.edu.co

Abstract. An important application of cooperative robotics is search and rescue of victims in disaster zones. The cooperation between robots requires multiple factors that need to be taken into consideration such as communication between agents, distributed control, power autonomy, cooperation, navigation strategy, locomotion, among others. This work focuses on navigation strategy with obstacles avoidance and victims localization. The strategy used for navigation is based on swarm theory where each robot is an swarm agent. The calculation of attraction and repulsion forces used to keep the swarm compact is used to avoid obstacles and attract the swarm to the victim zones. Additionally, an agent separation behavior is added, so the swarm can leave behind the agents, who found victims so these can support the victims by transmitting their location to a rescue team. Several experiments were performed to test navigation, obstacle avoidance and victims search. The results show how the swarm theory meets the requirements of navigation and search operations of cooperative robots.

1 Introduction

Unfortunately, Throughout the history of mankind, we have experienced uncountable natural disasters and terrorist attacks, which decimate cities and towns. Evidently, these events end with the lives of many people, and that is the main reason why numerous approaches exist to deal with the before, during and aftermath of a disaster.

When a disaster occurs, the most critical steps that need to be followed in order to save lives are exploration, search, localization and rescue of survivors. Each step is accompanied by a set of challenges for rescue teams, because

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_104

affected areas are difficult to access and are extremely dangerous, for example they might have uneven grounds, which are prone to landslides, or collapsed buildings, among other possible scenarios. Fortunately, we have seen significant progress in the interaction between rescue workers and robotic platforms used in rescue operations, as shown in the work presented by Casper in [1], which deals with the events presented in the attacks of the World Trade Center. This type of work has allowed the public to see the advantages, disadvantages and applicability of robotics in disaster zones.

One of the recent theories applied to rescue robotics are bio-inspired systems, which are based on animals, specifically insects like ants, termites and bees [2,3]. These animals show applicable behaviors, such as localization, recognition and are also able to explore large areas, when they are searching for an specific target. Ethology says these animals possess cooperative dynamics, which can be explained in mathematical form. In fact, the implementation and theoretical development of the behavior of swarm agents can open a new path to multi-robots systems, which can provide different kinds of solutions with simple structures to complex challenges. One of the possible approaches of this bio-inspired system, which is the one being covered by this paper, is swarm theory applied to the simulation of robots for rescue applications. These robots can keep a formation, while searching for position targets and can also avoid obstacles.

The use of swarms of robots in exploration of disaster zones facilitates and streamlines rescue tasks, thus making the recognition and mapping of affected areas faster and safer for rescue workers, while providing more information for search-and-rescue of possible survivors.

The other sections that complete this paper are Sect. 2 - Related Work, Sect. 3 - Robotic Swarms, Sect. 4 - Experiment and Results, Sect. 5 - Conclusions.

2 Related Work

Proper search and rescue planning should take into account all of the factors, that can hamper decision-making and lower the performance of duties, such as high stress levels and disorientation that normally affect rescue workers during disaster events. This is critical to safeguard lives and facilitate prompt assistance to those who require it. In disaster events, it is important to have access to a solid contingency procedure, as explained in the work presented by Huder in [4], which discusses the first hours and critical initial days that take place during disaster responses. It is in the first hours and initial days when the robot action can be crucial to save lives. It mentions the precautions and planning that individuals must carry out beforehand and also gives different perspectives on how to manage critical situations presented in diverse types of disasters.

As mentioned in previous section, a possible addition to current search and rescue procedures is the implementation of robotic platforms to locate survivors. As previously mentioned, a benchmark for these type of robotic platforms was displayed during the 911 attacks, where Casper and Murphy in [1] presents the advantages and disadvantages associated with the implementation of this

technology and the performance of each robot that deployed. It is also worth mentioning that the robots used by Casper were not conventional, since they were structurally designed to withstand though disaster conditions. In the work presented by Ventura and Lima [5], the need to use heterogeneous robotics teams, such as aerial robots that perform exploration and recognition of the disaster zone, was discussed. In addition to that the author mentioned the need to include powerful robots able to remove debris and rubble, and small agile robots capable of reaching people trapped under ruins.

Regardless of the robotic platform selected, it is clear that they must have a certain level of adjustable autonomy, basic learning capabilities, and object handling abilities, in addition to sufficient human interface options for rescue teams. In this paper, we have outlined the achievements of several search and rescue robots from around the world, in order to shine more light on how these platforms work and to show the increasing level of autonomy that we have witnessed in the last decade. In the contribution made by Seljanko in [6], They proposed a low cost and low maintenance add-on search and rescue robotic system with high reliability and robustness, which includes electronic devices such as microphones, speakers, GPS and integrated camera, thus resembling a smart-phone.

Swarm robotics theory can be best implemented in search and localization of victims in disaster zones, due to the teamwork or cooperation displayed by the swarm, as presented by Tan in [7], who said that the main advantages of using swarms robots is the exploration and recognition of larger areas in less time, specially when compared to the single robot case, thus allowing greater flexibility which is ideal for our application. It is also advantageous for rescue teams to have a swarm of robots, because if some of the agents of the swarm were to fail, the absence of these agents does not seriously affect the overall success of the swarm, which clearly demonstrates the greater robustness of the decentralized control system.

To work with robotic swarms, we must take into consideration factors such as communication, information sharing between agents, the distributed control algorithm, cooperation and navigation, and that is why will explore them in this paper.

Several algorithms designed for the navigation of swarm robots have been tested in the past, one of the more commonly used ones is the swarm intelligence, which was presented by Couceiro et al. in [8]. This paper presents a comparison between the Particle Swarm Optimization (PSO) and an algorithm derived from the PSO called Darwin Particle Swarm Optimization, which is based on the theory of evolution. The results obtained from the PSO algorithm are generally good, but sometimes fail to be trapped in the local optima and can not find the global optimum, on the other hand, the DPSO is a variation, which can perform better, because it divides the swarm into sub-swarms. Dividing the swarm into several sub-swarms is beneficial, because the sub-swarms can share the best solution with the other sub-swarms and if this new team solution is better than the possible solution derived by the swarm as a whole, the sub-swarms are allowed by the complete swarm to go to the new optimal solution, thus avoiding being trapped in the local optima.

3 Robot Swarms

In nature, there are several species that display swarm like behavior, such as bacteria, insects, birds, fish and horses among others. As can be expected, these groups are composed of individuals, who possess specific distinct capabilities and behaviors, but when they all get together as one group, and operate as such, they will behave differently in order to work as a group. One of the more notable examples of this group behavior is a swarm of honey bees choosing a place to build a new honeycomb, this phenomenon was presented by Seeley and Buhrman in [9], whose paper showed that this swarm behavior, which takes place during the spring and summer, is best exemplified when the colony outgrows its honeycomb, and then proceeds to divide itself to find new grounds and then gets back together as a swarm once it finds the correct location to build a bigger hive. To be more specific, The new site selection begins with several hundred scout bees, who leave the colony in search of potential new sites, once the scouts find an appropriate candidate, they return to the hive in order to report with a wiggle dance where the potential new place is located. Once the scouts select the correct location from the pre-selected sites, they will steer the rest of the swarm to the new location by chemical stimulation. After the selection step is completed, the division process of the hive begins, and the old queen with nearly half of the colony leaves the hive to build the new one, while a younger queen stays with the other half of the colony in the old honeycomb.

It is also important to highlight that the ability of the swarm to navigate can be affected by a set of critical factors, such as collisions between themselves, large obstacles in route, pheromone communication errors, erroneous scout indications and even poor sense of direction.

We can represent the swarm of bees with the help of graph theory with the application of the concepts worked by Mesbahi in [10]. One can describe the system by defining N agents representing each member of the bee swarm, or in the case of our application a robot, and a connection representing the communication and transmission of information between the agents, either bees or robots. As mentioned by Mesbahi, we first need to define the agents of the system (V, E), where V is the set of nodes in the system and E stands for the topology of the system, we also need to define $N(i)$ as the neighborhood or adjacent nodes to node i. Regarding the links between the agents, we can be describe them as a dynamic system.

It is advantageous to apply graph theory to control swarm responses, because it allows us to apply concepts such as finding the Laplacian, which represents the system, and assuming that the whole system is a graph fully connected. In addition to that, we can assume that the system has no noise, links are bidirectional, or even digraphs with different hierarchies.

The Algorithm used to simulate the swarm behavior is based on the one proposed by Passino in [11]. This algorithm is explained as follows; We also want to specify to the reader that this algorithm is not focused on a spacial kind of animal swarm, and therefore we are just referring to N number of agents.

Agents: Each agent is described as a point in the space with position and velocity. Initially, the position and velocity are defined randomly. We assumed that each agent can detect the position and velocity of the rest of the swarm agents. The swarm agents interaction is represented by graphs (G, A) where $G = \{1, 2, ..., N\}$ is a set of nodes and $A = \{(i, j) : i, j \in G, i \neq j\}$ represents the communication and sensing topology between the i^{th} agent and each of the other j^{th} swarm agents. We will now define the terminology and elements used in our algorithm.

Interaction: The interaction of the swarm agents is defined by the attraction and repulsion strategy. This strategy allows the agents to keep a comfortable distance over the other swarm agents. The attraction parameter tries to maintain in close proximity every element of the swarm and thus gives the group a mechanism to remain grouped together. The attraction parameter is defined by (1)

$$- k_a(x^i - x^j) \tag{1}$$

where k_a is the attraction force, this mechanism can be local (having a restriction by the sensing range) or global (Agents can move other agents from the group regardless of how far they are).

The repulsion parameter allows the agent to keep distance from the other members of the swarm. This avoids collision between agents when the swarm is moving. The repulsion mechanism can respond in two different ways. First one, when a comfortable distance is reached and its parameter is expressed by (2)

$$[-k(\|x^i - x^j\| - d)](x^i - x^j) \tag{2}$$

where $\|x^i - x^j\| = \sqrt[2]{(x^i - x^j)^T(x^i - x^j)}$, $k > 0$ is the repulsion magnitude and d is the comfortable distance between i^{th} agent and the j^{th} agent. On the other hand, if two agents are two close to each other, the parameter is represented by (3).

$$k_r exp(\frac{-\frac{1}{2}\|x^i - x^j\|^2}{r_s^2})(x^i - x^j) \tag{3}$$

where $k_r > 0$, represents the repulsion magnitude and $r_s > 0$, is the repulsion range.

Environment: This is the space where the agents can move and it is composed of good and bad areas. These good and bad areas are analogous to places where agents can find food and to areas that have predators respectively. For this present work, the good areas are defined by the presence of survivors, who need to be rescued and the bad areas are full of ruins, collapsed structures and obstacles that impair the ability of the swarm to navigate freely.

The environment is defined by $J(x)$, where $x \in \Re^n$. Its is assumed that $J(x)$ is continuous and that it has a finite slope in every direction. The agents navigate to the negative gradient (4) of the $J(x)$

$$- \nabla J(x) = -\frac{\partial J}{\partial x} \tag{4}$$

As mentioned above, the obstacles and victims that must be found by the agents are part of the environment. For the current work, both elements generate forces of repulsion and attraction over the agents. In this way, the interaction between agents with victims and obstacles is modeled using equation (3) with some changes in the meaning of the variables.

Interaction Between Agents and Obstacles: The obstacles produce a repulsion force over the agents. This repulsion is modeled using (3) where $k_r > 0$, and r_s is related to the size of the obstacle.

Interaction Between Agents and Victims: The victims exert an attraction force, where $k_r < 0$ and r_s is proportional to the victims density in the near area.

Left Behind: Every agent is exposed to different attraction and repulsion forces, which help to keep the swarm cohesion and at same time pushes the swarm towards the goal. The victims generate an attraction force over close agents and this force can disturb the normal behavior of the swarm and reduce the velocity of the agents in proximity to the victims. The k_r magnitude was put in place so that it can stop at least one of the closest agents near the victims. When agents are navigating close to the victims, at least one agent must stop near the victims. Once one or more agents stop beside to the victims, those agents are left behind by the swarm. They are separated from the swarm, breaking the communication and attraction forces that keep them into the swarm. The agents left behind create a new smaller swarm close to the victims, which changes its status to stationary with the victims and the starts to transmit the localization of the newly found victims.

4 Experiment and Results

In order to perform experiments for swarm navigation, the model previously explained was implemented using Matlab. Four different types of experiments were developed to show and analyze the swarm behavior in search and rescue operations.

Obstacle Avoidance: The first version of this experiment is performed using 10 agents and small obstacle as shown in Fig. (1). This part of the experiment depicts how the swarm goes around the obstacle in order to avoid it. This is possible, because the obstacle is relatively small and the agents are able to tolerate the obstacle between the attraction forces. The second version uses a bigger obstacle as depicted in Fig. (2). In the second case, the agents avoid the obstacle by taking a side path, this occurs because there is not enough space between the attraction forces to allow broad obstacles to stay in between the agents.

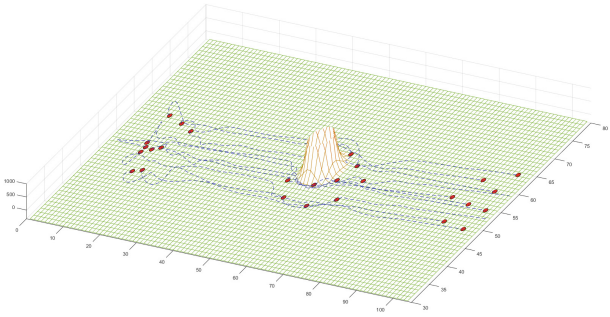


Fig. 1. Navigation with a small obstacle

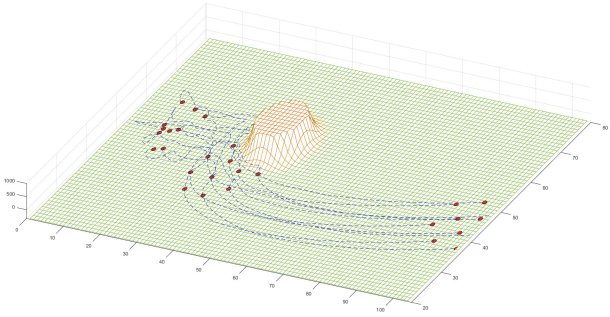


Fig. 2. Navigation with a big obstacle

Multiple Obstacles: This part of the experiment depicts how the swarm goes around obstacles in order to avoid them. This is possible, because the obstacles are relatively small and the agents are able to tolerate the obstacles between the attraction forces, as depicted by Fig. 3. This case uses nine obstacles distributed throughout the area between the start point and the goal point. It forces the swarm to navigate through the obstacles. while avoiding them and moving towards to goal point. Figure 3 shows the agents navigating and exploring the zone. At same time, the path described by every agent is drawn and it depicts the explored area.

Victim Localization: This test uses a flat terrain to show how the swarm localizes victims. This process is accomplished with the use of the "left behind" process. Once an agent has localized a victim, it stops near the potential victim. At the same time, the swarm stops, because of the attraction force between the agents. In that moment the process called "Left behind" detaches the agents found victims. The detached agents create a new swarm surrounding the victims. The original swarm restarts to move towards the target point and leaves behind the swarm agents that are in charge of the victims. Figure (4) shows how the agents stop near the victims and surround them, while the swarm leaves them behind.

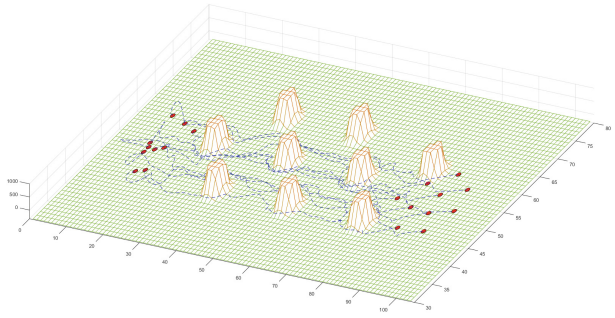


Fig. 3. Navigation with several obstacles

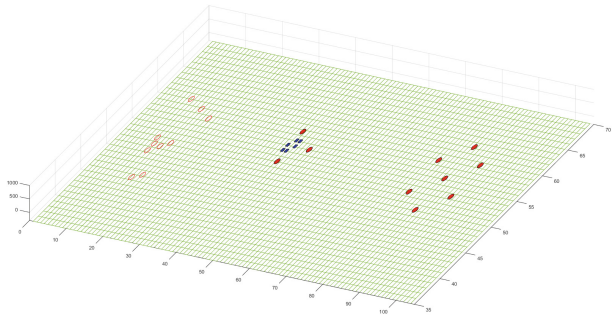


Fig. 4. Search of victims area

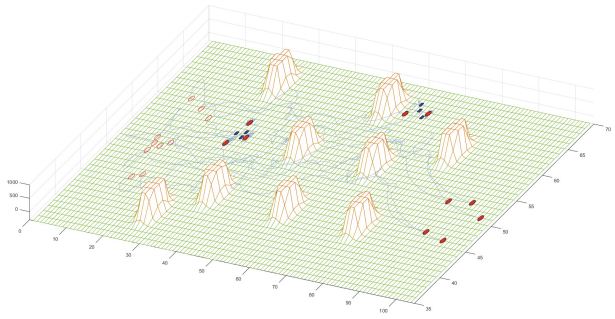


Fig. 5. Victims zone with multiples obstacles

Navigation and Victim Localization: This is a complete case where the swarm navigates through the area full of obstacles and some places with potential victims. This test is divided in two experiments. The first one uses 10 agents and 2 victim places as shown by Fig. 5. The second one has 40 agents and 5 victim places (Fig. 6). Both cases depict how the swarm covered the area navigating through obstacles and localizing victims at the same time. The performance

differences between these two cases are the covered area and the number of agents surrounding the victims in independent clusters. With large numbers of agents, the covered area is bigger, because the repulsion force pushes the agents harder to keep the distance between them. These agents use more area, to find an equilibrium point of repulsion and attraction forces. The number of agents around victims is also bigger, because there are more agents in the swarm and therefore rises the probability of finding victims for each agent.

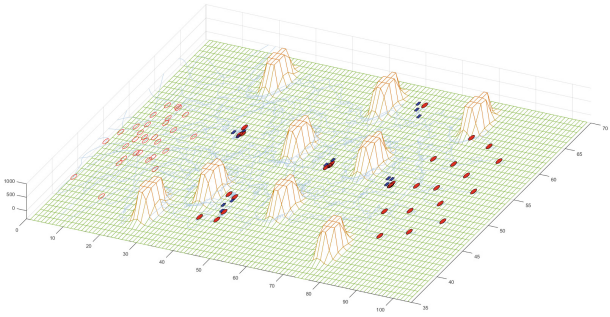


Fig. 6. Several victim areas with a big swarm

5 Conclusions

An algorithm for search and rescue operations has been proposed using swarm theory. The algorithm is based on concepts of attraction and repulsion forces. These concepts keep the swarm as compact as physically possible and fix a minimum distance between them. It is possible to represent the swarm with graph theory, where the agents are nodes, the attraction and repulsion forces are links between the nodes. The repulsion force is used to avoid obstacles and to keep a minimum distance between the agents. The attraction forces allow the swarm to stay compact and to be attracted by the victims. Additionally, a concept called "left behind" is introduced with the aim of making possible the opportune separation of some agents from the swarm. Agents who find victims are separated from the swarm and they are in charge of the global localization of the victims so they can be rescued. Four different experiments were developed with the objective of showing the performance of the swarm in several cases of navigation. The experiments depicted show how the swarm navigates in areas with multiple obstacles and simultaneously searches for victims. The algorithm shows creation of new small swarms around victims, which can be used to support and facilitate the localization of victims for rescue teams. The swarm algorithm shows redundancy and robustness characteristics when it loses agents, but it regroups with the ones that are left and keeps the navigation process. Future work includes the use of algorithms based on graph theory to improve the relation

of the sub swarms and the global swarm. Additionally, the algorithm proposed in the current work will be performed in a robotics simulator with the purpose to check this ideas on aerial and ground robots.

References

1. Casper, J., Murphy, R.R.: Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. IEEE Trans. Syst. Man Cybern. Part B Cybern. **33**(3), 367–385 (2003)
2. Quijano, N., Passino, K.M.: Honey bee social foraging algorithms for resource allocation, part I: algorithm and theory. In: American Control Conference, pp. 3383–3388. IEEE, July 2007
3. Quijano, N., Passino, K.M.: Honey bee social foraging algorithms for resource allocation, part II: application. In: American Control Conference, pp. 3389–3394. IEEE, July 2007
4. Huder, R.C.: Disaster Operations and Decision Making (1). Wiley, Hoboken (2013)
5. Ventura, R., Lima, P.U.: Search, rescue robots: the civil protection teams of the future. In: Third International Conference on Emerging Security Technologies (EST), pp. 12–19 (2012)
6. Seljanko, F.: Low-cost electronic equipment architecture proposal for urban search and rescue robot. In: IEEE International Conference on Mechatronics and Automation (ICMA), pp. 1245–1250 (2013)
7. Tan, Y., Zheng, Z.: Research advance in swarm robotics. Def. Technol. **9**(1), 18–39 (2013)
8. Couceiro, M.S., Rocha, R.P., Ferreira, N.M.F.: A novel multi-robot exploration approach based on particle swarm optimization algorithms. In: 2011 IEEE International Symposium on Safety, Security, and Rescue Robotics, pp. 327–332 (2011)
9. Seeley, T.D., Buhrman, S.C.: Group decision making in swarms of honey bees. Behav. Ecol. Sociobiol. **45**(1), 19–31 (1999)
10. Mesbahi, M., Egerstedt, M.: Graph Theoretic Methods in Multiagent Networks. Princeton University Press, Princeton (2010)
11. Passino, K.M.: Biomimicry for Optimization, Control, and Automation. Springer Science & Business Media, Berlin (2004)

Linguistic Representation by Fuzzy Formal Concept and Interval Type-2 Feature Selection

Sahar Cherif[1(✉)], Nesrine Baklouti[1], Adel M. Alimi[1], and Vaclav Snasel[2]

[1] REGIM-Lab.: REsearch Groups in Intelligent Machines, National Engineering School of Sfax, Sfax, Tunisia
{sahar.cherif.tn,nesrine.baklouti,adel.alimi}@ieee.org
[2] Departement of Computer Science Faculty of Electrical Engineering and Computer Science, Ostrava, Czech Republic
vaclav.snasel@vsb.cz

Abstract. Natural language is always seen as a source of uncertainty and vagueness. Fuzzy logic (FL) is a powerful tool for representing and treating perceptions which are the inputs and outputs of a linguistic model. In fact, a linguistic representation is a methodology that moves from crisp measures to uncertain words or fuzzy concepts. This theory uses fuzzy sets to encode and represent linguistic concepts. In this paper, an interval type-2 fuzzy formal concept IT-2FFC is presented as a new approach for extracting knowledge in a linguistic model. The method represents a combination of two techniques: fuzzy formal concept (FFC) for visualizing data and interval type-2 fuzzy sets (IT-2FSs) for feature selection. The obtained results demonstrate that the method applied can help human to make subjective judgments and make decision in a knowledge model.

Keywords: Fuzzy formal concept · IT-2 FSs · Feature selection · Concept lattice

1 Introduction

In general, human interpret their actions in a natural language by using words which can have different meanings. This can lead to misunderstandings or misconceptions. Words or concepts lexicalized in natural language can then be vague and uncertain. Recall that a linguistic representation is a readable machine in which the objects of computation are words and propositions drawn from a natural language as cited by Zadeh [1] in 1996. To make subjective judgments, psychologists and scientists aim to develop concept models for representing human natural language. Many theories were proposed in concept's psychology such as the classical, the prototype theory and the knowledge views of concepts [2]. In the classical representation of concepts, concepts have a definitional form. Concepts can be well defined and have clear boundaries. Recall that a concept

© Springer International Publishing AG 2017
A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_105

of y (example: a concept of temperature) is a knowledge source of y (example: temperature). The classical view represents some problems, for example, an instance can belong to a category or to another. The prototype theory was then advanced in 1970 by Rosch [2]. In this view, every concept is represented by a central exemplar also called the typical member or the prototype. This element is the most central member of a concept which contains the maximum number of features. In the knowledge view [3], concepts are learned as parts of the human actions or experiences in interaction with the world. Many studies were done to explain advantages and limitations of the three views cited previously. In the psychology of concepts, many researchers opted on the use of FL to study concepts properties and concepts aims. They considered that FL can be a powerful representational tool for concepts. Rosch in [2] showed that type-1 fuzzy sets can better represent concepts than crisp sets. In 1972, Lackoff in [4] has confirmed Rosch s idea and showed that FL is important in handling linguistic information. In 1982, Wille in [6] advanced the formal concept analysis (FCA) for dealing with concepts views. The FCA is used for dealing with concepts and concept classes. It is a mathematical tool for interpreting data in the psychology of concepts. Because human interaction is natural, vague or uncertain, the fuzzy formal concept (FFC) was introduced to deal with imprecise concepts. The FFC is an extension of the FCA when the data contains fuzzy attributes. The remaining of the paper is organized as follows: In Sect. 2, we present first an overview of the FCA and its extension the FFC. Secondly, in Sect. 3, we describe our approach in which we use the FFC based on an interval type-2 feature selection as a new linguistic representation. Some experimental results will be interpreted to show how the FFC can be used to make subjective judgments. Finally, conclusions and some future works are presented in Sect. 4.

2 Fuzzy Formal Concept

2.1 Formal Analysis Concept: An Overview

The FCA is a technique which analyzes data in terms of concepts. Each concept is represented by objects, attributes and the relation between them. Gajdos and Snasel [5] have presented a state of arts of the FCA from 1967. Wille in [6] presented concepts as a collection of objects and attributes (the attributes are related by those objects). Ganter and Wille in [7] demonstrate that a concept is reported by its extension and its intention. The extent set of a concept constitutes all the objects while the intent set is composed by all the attributes in relation with the objects in the extent set. In mathematical term, FCA is composed by 2 sets A and B and a relation C (A is the objects sets and B is the attributes sets). C is a binary connection between A and B composed of 0s and 1s. Hence, for each $a \in$A and each $b \in$B, C = (a,b) is the relation between a and b. C means that the object a is related to the attribute b. The triplet (A, B, C) is called a formal context which is represented in general by a table (see Table 1 as an example of the formal context: x1, x2, x3, x4 and x5 constitute the objects where y1, y2 and y3 represent the attributes). When a table entry contains a 1

Table 1. A formal context example

	y1	y2	y3
x1	1	0	0
x2	0	1	1
x3	0	1	1
x4	0	1	0
x5	0	0	1

value, it means that there are a connection between the object and the attribute (the object possess the attribute). When the entry table is 0 means that there is no association between the object and the attribute. The set of all the formal contexts (A, B, C) constitutes a formal concept, it corresponds to largest regions in the table that are composed of 1s as illustrated in Fig. 1.

The 6 formal concepts obtained can be illustrated in Table 2. The column extent indicates all the objects involved in the formal context while the column intent is about all the attributes which are in relation with objects. The set of all the formal contexts is generally presented graphically by a concept Lattice which can be very useful for mining concepts. The concept Lattice can be seen via a line diagram called also Hass diagram which consists of nodes connected by lines. The nodes are assigned with numbers illustrating the formal contexts. Figure 2 presents the line diagram of the concept lattice extract from Table 2. In this figure, numbers from 1 to 6 represents the obtained 6 formal contexts. The main problem of the concept lattice consists on the fact that it can not be generated when dealing with important number of attributes or instances [2].

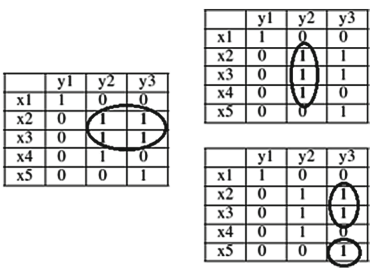


Fig. 1. Formal contexts by maximal rectangles

2.2 Fuzzy Formal Concept

As cited above, since 1970, some researchers [2,4] have demonstrated that FL is a powerful tool for dealing with concepts. FL aids users to make subjective judgments from uncertain, imprecise or noisy data (e.g. natural language). T-1 FSs and T-2 FSs were used to resolve some problems of linguistic uncertainty in human interactions. In the psychology of concepts, the FFC is a form of

Table 2. List of all concepts

Index	Extent	Intent
1	x1 x2 x3 x4	Ø
2	x2 x3	y2 y3
3	x1	y1
4	x2 x3 x4	y2
5	x2 x3 x5	y3
6	Ø	y1 y2 y3

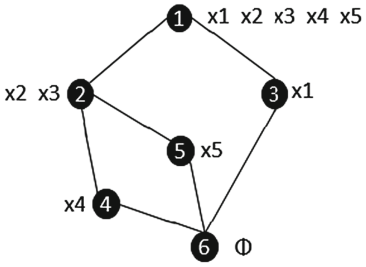


Fig. 2. Line diagram of a concept Lattice

knowledge representation used when the FCA has fuzzy attributes. In fact, the connection between objects and attributes is no longer binary. For a an object and b an attribute, the relation C(a,b) is, in fact, the extent to which the object a has the attribute b (between 0 and 1). As in FCA representation, data are illustrated in a table with entries varying from 0 to 1. Table 3 illustrate the fuzzy relation between the objects x1, x2, x3, x4, x5 and the attributes y1, y2, y3, y4, y5, y6, y7, y8, y9 and y10. Some works like [3] uses shades of gray to represent fuzzy contexts which correspond to darker rectangles in the fuzzy table. Figure 3 shows an example of a FFC table where the shades of gray are, the truth of degrees of relations between objects and attributes illustrated by Table 3. More the shade of gray is dark more the degree of truth is close to 1, it can be deduced that each object has the correspondent attribute with a fixed degree of truth. Other works like [8–10] use Type-1 memberships values to indicate a degree of belonging describing relations between objects and attributes. Their approaches are in general applied to specific applications in data mining, medical or hotel fields. As in FCA, the relation is illustrated via a fuzzy concept lattice for representing objects and attributes connections. In general, a threshold operation is applied on Table 3 in order to eliminate irrelevant membership values (values that are less than 0.5). The result of such operation is illustrated in Table 4.

The concept lattice presented in Fig. 4 and generated from Table 4 is illustrating all the fuzzy relations existing between the objects and the features. For a specific application, the context visualization given by the lattice can offer a

Table 3. A fuzzy formal context example

	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10
x1	0.25	0.75	0.8	0.32	0.8	1	0	0	1	1
x2	0	0	0	0.25	0.75	0.5	0.65	1	1	1
x3	0	0	0	0	0.5	0.36	0.1	1	1	1
x4	1	0.25	0.25	0.8	0.8	1	1	0	0.23	1
x5	1	1	1	0.75	0.5	0.9	0	0	0	0.32



Fig. 3. Shades of gray corresponding to Table 3

better interpretation and can aid people to make a good interpretation and of course make a subjective judgment. In order to enhance such judgment or decision, some concepts similarities were done after the implementing the FFC or the FCA approaches. An overview of many similarity methods were done in [8].

Table 4. The new incidence Table obtained after the threshold operation

	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10
x1		0.75	0.8		0.8	1			1	1
x2					0.75	0.5	0.65	1	1	1
x3					0.5			1	1	1
x4	1			0.8	0.8	1	1			1
x5	1	1	1		0.5	0.9				

3 IT-2 FFC: Interval Type-2 Fuzzy Formal Concept

3.1 Proposed Method

Many applications such as decision making, robots control, biomedecine, etc. deals with concepts lexicalized in natural language. Consequently, using a machine to convert human actions must be a hard task [11]. Almost all the methods implementing FFC, use Type-1 FL for representing relations between objects and attributes. When dealing with natural language, many linguistic approaches have adopted the Yager [12] and the Mendel models [14] and have shown that Type-1 FL present some limits comparing to type-2 FL. An exposition of all these methods was done in [13]. The common point of these methodologies is the use of T-2 FSs as models for representing concepts and handling

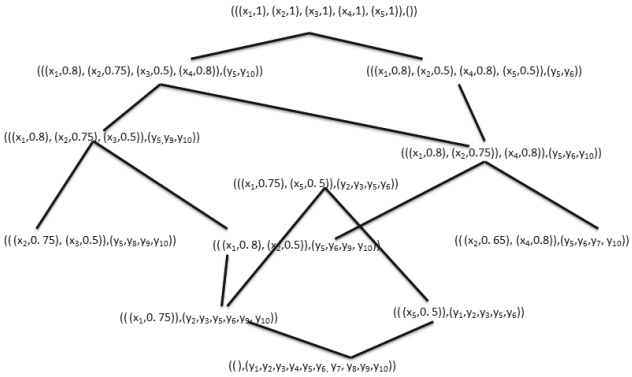


Fig. 4. Fuzzy formal Lattice corresponding to Table 4

the categories uncertainties thanks to the footprint of uncertainty (FOU). As described in Sect. 2, the FFC is used for extracting information and visualizing fuzzy data. When the size of inputs objects and attributes is important, the computation time and the complexity will increase. The concept lattice cannot then be generated. In order to resolve the incertitude and the computation problems, the following idea was advanced: a new IT-2 FFC based on an IT-2 feature selection (IT-2 FS) and describing a T-2 fuzzy relation between objects and attributes. In fact, when manipulating data in such domain, only relevant features are kept. The other attributes are not significant and can lead to an important computation time. The idea is mainly a clustering approach based on a feature selection by eliminating non significant concepts or noisy categories. Based on the fact that a concept can be equalized to an IT-2 FS as described in [15], our attribute/feature selection is done by computing the centroid of all the concepts FOUs. The concept which have important centroid value will be kept, the other concepts are irrelevant. This can be interpreted by the following statement: an IT-2 FS having a considerable centroid value means that a large group of subjects approved this concept, they are more interested with this concept than the other attributes available in the data. After this feature selection, the data size is much reduced, the Fuzzy lattice can hence be generated based on a new incidence table. The proposed method process is described in Fig. 5. The first step in our approach is the data construction. IT-2 fuzzy data are obtained by asking people (which are the objects) to give reasonable values (left and right values) for words (which are the attributes). The IT-2 feature selection step seems to be fundamental When dealing with important data in order to generate the concept lattice.

3.2 Experiments and Results

Our approach will be clarified through an example which demonstrates the changes from an FFC of important data size to an IT-2 FFC based on an

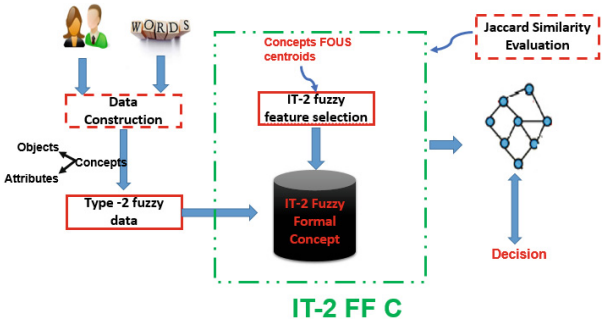


Fig. 5. The proposed Method

attribute selection and an IT-2 fuzzy relation. Let's take the Mendel database [16] which is collected from 28 instances for a vocabulary of 32 attributes. These features were randomized, and for all words, each subject was asked the question, "On a scale of 0 and 10, what are the endpoints of an interval that you associate with the word W?" Each subject has given an interval composed of two values: left and right values. When creating the fuzzy lattice from the initial data, We remark that the lattice generated present a huge number of connections between attributes which are so complicated and the lattice interpretation is hence so difficult. For each attribute, the FOU is generated; It can be interior, left shoulder or right Shoulder. The obtained FOUs are shown in Fig. 6. An IT-2 feature selection is done in order to reduce input data by computing all the FOU centroids of all the concepts as shown in Table 5. Only concepts with centroids higher than a fixed threshold will be kept (threshold = 7). From 32 words, only 13 words will be kept. The new incidence matrice corresponding to Mendel database will contain 28 subjects and 13 attributes. In order to confirm the results obtained after the IT-2 fuzzy feature selection, an evaluation of the 13 attributes kept must be done regardless the concept lattice. The verification step consists on computing the similarity between the remaining 13 attributes. The Jaccard Similarity [14] is used to calculate the similarity between the memberships (lower and upper) of two IT-2 FSs. In fact, the Jaccard measure fulfills the conditions of reflexivity, symmetry, transitivity and overlapping as mentioned in [14]. For the attributes remained from the initial data, the Jaccard similarity computed between all the attributes is given in Table 6. As we remark, almost all the similarities computed are less than 0.6. Recall that when similarity is close to 0, the attributes are dissimilar and when the similarity is close to 1, the attributes are similar and can be considered as the same attribute. The similarities computed in Table 6 confirm the obtained results and indicate that the new incidence table will effectively be composed of 13 attributes and 28 subjects.

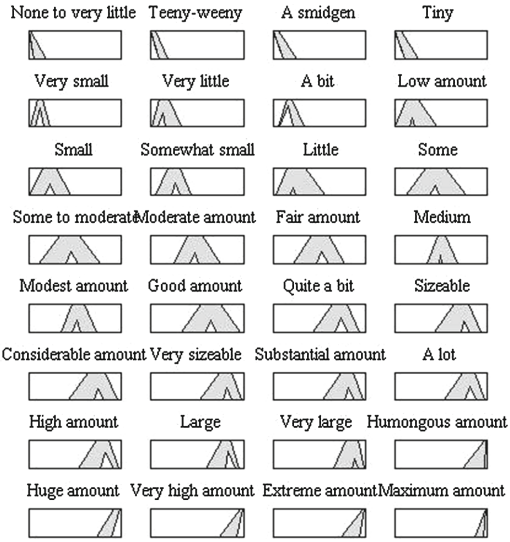


Fig. 6. The obtained FOUs for the 32 Mendel attributes

Table 5. Centroids values computed for all the concepts FOUs

Attribute	Centroid value	Attribute	Centroid value
Little	2.16	Medium	5.1838
Sizeable	3.5	Good amount	6.5
Large	8.1398	Extreme amount	9.555
Quit A bit	7.4794	Teeny-weeny	0.3511
Low amount	1.9912	Considerable amount	7.5895
Somewhat small	2.3933	A lot	8.2322
A smidgen	0.4211	A bit	1.7505
None to very little	0.4450	Very sizeable	8.2549
Humongous amount	9.4956	Some to moderate	4.823
Huge amount	9.4487	High amount	8.9146
Very small	1.25	Small	2.3121
Very large	8.8679	Maximum amount	9.6834
Fair amount	5.5	Some	3.5
Modest amount	5.2222	Tiny	0.4855
Very little	0.5369	Very high amount	9.3672
Moderate amount	5.0345	Substantial amount	8.092

Table 6. The Jaccard similarity between the 13 attributes kept

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Large		0.5103	0.1216	0.1274	0.3198	0.2068	0.0563	0.5935	0.7172	0.6799	0.428	0.0494	0.1909
2. Quite a bit	0.5103		0.078	0.0783	0.1607	0.3509	0.024	0.8418	0.042	0.3923	0.2632	0.0272	0.1319
3. Humongous amount	0.1216	0.078		0.6464	0.3435	0.0058	0.4771	0.1052	0.1838	0.1909	0.3268	0.6641	0.6760
4. Huge amount	0.1274	0.0783	0.6464		0.3496	0.0058	0.6359	0.1061	0.1845	0.1903	0.507	0.493	0.7588
5. Very large	0.3198	0.1607	0.3435	0.3496		0.021	0.1846	0.2147	0.4104	0.4265	0.4214	0.1879	0.421
6. Extreme amount	0.0563	0.024	0.4771	0.6359	0.1846		0.0402	0.0821	0.0827	0.3854	0.2486	0.5152	0.0731
7. Considerable amount	0.5935	0.8418	0.1052	0.1061	0.2147	0.0402		0.5114	0.4743	0.3063	0.0455	0.1636	0.6136
8. A lot	0.7172	0.04281	0.1838	0.1845	0.4104	0.0821	0.5114		0.9211	0.452	0.0942	0.2609	0.8462
9. Very sizeable	0.6799	0.3923	0.1909	0.1903	0.4265	0.0827	0.4743	0.9211		0.4558	0.0968	0.2711	0.8047
10. High amount	0.428	0.2632	0.3268	0.507	0.4214	0.3854	0.3063	0.452	0.4558		0.2485	0.4842	0.4434
11. Maximum amount	0.0494	0.0272	0.6641	0.493	0.1879	0.6484	0.0455	0.0942	0.45558		0.4842	0.4434	
12. Very high amount	0.1909	0.1319	0.6760	0.7588	0.421	0.4959	0.1636	0.2609	0.2711	0.4842	0.2711		0.2348
13. Substantial amount	0.7975	0.5287	0.1634	O.1667	0.3636	0.0731	0.6170	0.8462	0.8047	0.4434	0.0828	0.2348	

4 Conclusion

In the psychology of concepts, FFC was advanced for extracting and visualizing data through the formal lattice which describe the type-1 fuzzy relation between objects and attributes. When the data size and the incertitude rate are important, the lattice generated cannot easily be interpreted and the knowledge cannot be well presented. Our new IT-2 FFC seems to be a good solution fOR knowledge representation by merging two tools: the IT-2 feature selection for eliminating irrelevant features and the FFC for visualizing data. As future works, we intend to enhance the proposed approach by exploring other similarity measures and testing the IT-2 FFC on big data by selecting the relevant features.

Acknowledgment. The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program.

References

1. Zadeh, L.A.: Fuzzy logic = computing with words. Trans. Fuzzy Syst. **4**, 103–111 (1996). IEEE Press
2. Belohlavek, R., Klir, G.J.: Concepts and Fuzzy Logic. The Mit Press, Cambridge (2011)
3. Belohlavek, R., Klir, G.J., Lewis III, H.W., Way, E.C.: Concepts and fuzzy sets: misunderstandings, misconceptions, and oversights. Int. J. Approx. Reasoning **51**, 23–34 (2009). Elsevier Science Inc.
4. Lackoff, G.: Hedges: a study in meaning criteria and the logic of fuzzy concepts. J. Philos. Log. **2**, 458–508 (1972)
5. Gajdos, P., Snasel, V.: A new FCA algorithm enabling analyzing of complex and dynamic data sets. Soft Comput. **18**, 683–694 (2014)
6. Wille, R.: Restructuring Lattice theory: an approach based on hierarchies of concepts. In: Ferré, S., Rudolph, S. (eds.) ICFCA 2009. LNCS (LNAI), vol. 5548, pp. 314–339. Springer, Heidelberg (2009). doi:10.1007/978-3-642-01815-2_23
7. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, New York (1997)
8. Formica, A.: Ontology-based concept similarity in formal concept analysis. Inf. Sci. **176**, 2624–2641 (2006). Elsevier Science Inc.
9. Zheng, S., Zhou, Y., Martin, T.: A new method for fuzzy formal concept analysis. In: Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, vol. 03, pp. 405–408. IEEE Computer Society (2009)
10. Li, S.-T., Tsai, F.-C.: A fuzzy conceptualization model for text mining with application in opinion polarity classification. Knowl.-Based Syst. **39**, 23–33 (2013)
11. Zhou, C., Ruan, D.: Fuzzy control rules extraction from perception-based information using computing with words. Inf. Sci. **142**, 275–275 (2002). Elsevier Science Inc.
12. Kacprzyk, J., Yager, R.R.: Linguistic summaries of data using fuzzy logic. Int. J. Gen. Syst. **30**, 133–154 (2001)

13. Cherif, S., Baklouti, N., Alimi, A.: The encoding part. In: 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), pp. 471–476 (2014)
14. Wu, D., Mendel, J.M.: A comparative study of ranking methods, similarity measures and uncertainty measures for interval type-2 fuzzy sets. Inf. Sci. **179**, 1169–1192 (2009). Elsevier Science Inc.
15. Cherif, S., Baklouti, N., Alimi, A.M., Snasel, V.: A type-2 fuzzy concepts lexicalized representation by perceptual reasoning and linguistic weighted average: a comparative study. In: Abraham, A., Han, S.Y., Al-Sharhan, S.A., Liu, H. (eds.) Hybrid Intelligent Systems. AISC, vol. 420, pp. 77–86. Springer, Heidelberg (2016). doi:10.1007/978-3-319-27221-4_7
16. Liu, F., Mendel, J.: Encoding words into interval type-2 fuzzy sets using an interval approach. IEEE Trans. Fuzzy Syst. **16**, 1503–1521 (2008)

Author Index

A
Abbes, Hanen, 612
Abdullah, Muhamad Yazid Che, 57
Abraham, Ajith, 417, 488, 559
Adel, Mahfoudhi, 116, 622
Ahiod, Belaïd, 135
Akid, Hajer, 965
Alam, Shafiq, 45
Alamir, Mazen, 154
Al-Behadili, Mohanad, 498
Albuquerque, Isabela Maria Carneiro, 362
Alimi, Adel M., 206, 289, 330, 417, 478, 488, 559, 863, 1071
Amin, Maalej Mohamed, 116, 622
Amous, Ikram, 549, 677, 781, 1041
Analide, Cesar, 984
Andrade, Marina A.P., 649
Antunes, Nuno, 582
Araújo, A., 728
Arroyo, José Elas C., 955
Arroyo, José Elias C., 889
Artabaz, Saliha, 260
Ayachi Ghannouchi, Sonia, 760, 800
Azaouzi, Mehdi, 250
Azmi, Mohd Sanusi, 77

B
Baati, Karim, 417
Bajo, Javier, 592
Baklouti, Nesrine, 1071
Bali, Olfa, 559
Barbudo, Rafael, 437
Belghith, Aymen, 749
Ben Abdellafou, Khaoula, 384
Ben Ammar, Lassaad, 879
Ben Ayed, Mounir, 602, 708, 841, 863, 965
Ben Boussada, Elhoucine, 863
Ben Hassen, Mariam, 1015
Ben Romdhane, Lotfi, 11, 250
Ben Said, Lamjed, 457
Benatchba, Karima, 260
Benlic, Una, 791
Bezine, Hala, 289
Bhiri, Mohamed Tahar, 821
Borges, Ana, 649
Botello, Andres, 1061
Bouaziz, Rafik, 900
Bouaziz, Senda, 467
Boubaker, Houcine, 1005
Bouhlel, Mohamed Salim, 239
Boujelben, Abir, 781
Boukhris Trabelsi, Randa, 811
Brahmi, Zaki, 372

C
Caldas, Rafael, 197
Calderón, Juan M., 1061
Camargos, Rodrigo C., 35, 67
Campos, Alba A., 889
Campos, Saulo C., 955
Cardona, Gustavo A., 1061
Carneiro, Davide, 592
Casado, L.G., 144
Cerqueira, Fabio R., 519
Chaari, Tarak, 781, 900
Chabchoub, Habib, 488
Chagas, Jonatas B.C., 921
Chakroun, Rania, 659
Charroux, Benoit, 667
Chen, Liang, 427
Cherif, Sahar, 1071
Chiha, Rim, 708
Choo, Yun-Huoy, 77
Costa e Silva, Eliana, 649
Covas, Ricardo, 649
Cunha, Bruno, 771, 911

© Springer International Publishing AG 2017

A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*,
Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0

D
Damak Masmoudi, Alima, 811
Damak, Wafa, 811
Dănciulescu, Daniela, 1
de Albuquerque, I.M.C., 352
de Carvalho, André C.P.L.F., 268
de Cos Juez, Francisco Javier, 1051
de Lima Neto, Fernando Buarque, 197, 394
de Moura Oliveira, P.B., 448
de Souza Junior, A.H., 529
de Souza, Wesley Willy Oliveira, 309
Dellys, Hachemi Nabil, 260
Desdouits, Chloé, 154
Dias, Madson L. Dantas, 341
do Carmo Nicoletti, Maria, 25, 35, 67
Dobbie, Gillian, 45
dos Santos, Edimilson Batista, 279
Duarte, Maísa Cristina, 279
Durães, Dalila, 592

E
El Benani, Bouazza, 135
El Krari, Mehdi, 135
Elleuch, Wiam, 478
Ellili, Walid, 227
Elloumi, Walid, 559
Elouni, Jihed, 841

F
Feki, Jamel, 995
Fermé, Eduardo, 630, 640
Fernandes, Massilon Lourenço, 279
Ferraz, Filipa, 984
Ferreira, Filipe Vieira Silva, 362
Ferreira, João C., 582, 630, 640
Filho, João Batista Monteiro, 352, 362
Fourati, Farah, 821
Freire, Ananda L., 88
Frikha, Mondher, 659

G
Gaied, Imen, 319
Gargouri, Faïez, 1015
Gargouri, Faiez, 467, 612, 935
Gaskó, Noémi, 871
Ghorbel, Leila, 549
Ghozzi, Faiza, 935
Giroudeau, Rodolphe, 154
Gomes, João Paulo P., 108
Gomes, Marco, 984
Gómez, Sergio Luis Suárez, 1051
Gonçalves, Frederica, 630, 640
Gonçalves, João N.C., 974
Gonçalves, Sérgio, 592
González-Ruiz, Vicente, 144
Gupta, Daya, 216
Gupta, Vishal, 216
Gutiérrez, Carlos González, 1051

H
Haas, Lukas, 165
Haataja, Keijo, 187
Hadda, Hatem, 384
Hamdani, Tarek M., 417
Haron, Habibollah, 57
Hassan, Doaa, 99
Hentati, Marwa, 879
Horváth, Tomáš, 268
Hruschka Júnior, Estevam Rafael, 279, 309
Hu, Yabing, 197

I
Ikonen, Tiia, 187
Isola, Jorma, 187

J
Jallouli, Maryam, 1041
Jamoussi, Salma, 677
Jemal, Hanen, 602
Jemili, Farah, 319
Jones, Dylan, 498

K
Kadir, M. Fadzil A., 406
Kahloun, Fouzia, 800
Kammoun, Wafa, 831, 852
Kamoun, Lotfi, 749
Kchaou, Hamdi, 330
Kechaou, Zied, 330, 602
Khairuddin, Alif Ridzuan, 57
Khdhir, Radhia, 749
Kherallah, Monji, 1005
Khrouf, Kaïs, 995
Khrouf, Omar, 995
Kitajima, Toshihiro, 175
Korbaa, Ouajdi, 319, 384
Koudil, Mouloud, 260
Krichene, Emna, 488
Kuroda, Yoshihiro, 175

L
Labidi, Noura, 900
Lachtar, Abdelfetteh, 227
Lage, Serge, 582
Lajmi, Sonia, 1041
Lasheras, Fernando Sánchez, 1051
Lazzez, Onsa, 206
Le Pape, Claude, 154

Lemlouma, Tayeb, 697, 738
León, Jose, 1061
Lima Santanelli, João, 394
Liouane, Zaineb, 697, 738
Liu, Hongbo, 427
Loukil, Habiba, 239
Ltaief, Mahmoud, 289
Ltifi, Hela, 841
Luna, J.M., 508
Lung, Rodica Ioana, 871

M
Maddar, Hela, 831, 852
Madureira, Ana M., 718, 728, 771, 911
Mahfoudhi, Adel, 879
Mallek, Hana, 935
Manoj, K., 728
Mantovani, Rafael G., 268
Manupati, V.K., 728
Marinho, Leandro B., 529, 539
Markert, Bernd, 197
Martinho, Ricardo, 760
Martins, Ana Lúcia, 630, 1031
Masmoudi, Mohamed, 841
Masmoudi, Youssef, 488
Medina-López, Cristóbal, 144
Mejri, Asma, 760
Mesquita, Diego P. Paiva, 108
Messaoud, Hassani, 697, 738
Mnif, Kais, 749
Mohamed, Abid, 116, 622
Monteiro, M. Teresa T., 974
Muda, Azah Kamilah, 77
Muda, Noor Azilah, 77
Murakami, Edwardo Arata Y., 175

N
Nabli, Ahlem, 467
Nait-Ali, Amine, 811
Narloch, Pedro Henrique, 125
Natale, Leandro Pupo, 945
Neto, Fernando Buarque Lima, 352, 362
Nietto, Paulo Rogerio, 25, 35
Nóbrega, Raul Victor M., 539
Noura, Aboudi, 11
Novais, Paulo, 592, 984

O
Omar, Nizam, 945
Oshiro, Osamu, 175
Ouarda, Wael, 206
Ouelhadj, Djamila, 498

P
Padillo, F., 508
Parpinelli, Rafael Stubs, 125
Pereira, Ivo, 771, 911
Pinto, Iola, 582
Pinto-Orellana, Marco A., 519
Pop, Petrică C., 1
Pour, Shahrzad M., 791
Putnik, G.D., 728

R
Rahali, Mourad, 239
Rahman, M. Nordin A., 406
Ramírez, Aurora, 437
Rebouças Filho, Pedro Pedrosa, 529, 539
Rim, Rebai, 116, 622
Robbana, Riadh, 821
Rocha Neto, Ajalmar R., 88, 341
Rodrigues, Helena Sofia, 974
Rodrigues, Murillo B., 539
Rodríguez, Jesús Daniel Santos, 1051
Rodríguez, María Luisa Sánchez, 1051
Romero, José Raúl, 437, 687
Roose, Philippe, 697, 738
Ruiz, Victoria, 299

S
Saany, Syarilla Iryani A., 406
Sabo, Cosmin, 1
Saidu, Abdullahi Nababa, 406
Salado-Cid, Rubén, 687
Samet, Mounir, 227
Sánchez, Ángel, 299
Santos, André G., 921
Santos, André S., 718
Sèdes, Florence, 549
Sellami, Dorra, 811
Serhrouchni, Ahmed, 667
Shakirov, Ilshat, 144
Shamsudin, Syadiah Nor, 406
Shili, Hechmi, 11
Silva, A.M.O., 352
Silva, António, 984
Silva, Fábio, 984
Sliman, Layth, 260, 667
Snasel, Vaclav, 1071
Solteiro Pires, E.J., 448
Soto, Daniel, 570
Soto, Wilson, 570
Souiden, Imen, 372
Souissi, Samih, 667
Souza, João Wellington M., 539
Steinbuch, Rolf, 165
Suciu, Mihai Alexandru, 871

Sueiras, Jorge, 299
Sun, XiaoBin, 45

T
Talib, Mohamad Shukor, 57
Tang, Han-Yang, 77
Tavares, Ricardo G., 955
Tenreiro Machado, J.A., 448
Teodoro, M. Filomena, 649
Teste, Olivier, 935
Toivanen, Pekka, 187
Tolonen, Teemu, 187
Toumi, Hajer, 372
Trabelsi, Abdelwaheb, 879
Troudi, Abir, 677
Turki, Mohamed, 1015

V
Vălean, Honoriu, 1
Varela, M.L.R., 728
Varela, Maria R., 718
Vélez, José F., 299
Ventura, Sebastián, 437, 508

W
Wali, Ali, 478
Weis, Fréderic, 697, 738

Y
Yoshimoto, Shunsuke, 175
Youssef, Habib, 831, 852

Z
Zayani, Corinne Amel, 549, 677
Zhang, Bo, 427
Zhang, Weishi, 427
Zouari, Leila Beltaïfa, 659
Zouari, Ramzi, 1005
Zouhaier, Houda, 457