Advances in Intelligent Systems and Computing 311

Ana Madevska Bogdanova
Dejan Gjorgjevikj *Editors*

ICT Innovations 2014

World of Data

Springer

Advances in Intelligent Systems and Computing

Volume 311

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.

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

Ana Madevska Bogdanova · Dejan Gjorgjevikj
Editors

ICT Innovations 2014

World of Data

Springer

Editors
Ana Madevska Bogdanova
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University
Skopje
Macedonia

Dejan Gjorgjevikj
Faculty of Computer Science and Engineering
Ss Cyril and Methodius University
Skopje
Macedonia

ISSN 2194-5357 ISSN 2194-5365 (electronic)
ISBN 978-3-319-09878-4 ISBN 978-3-319-09879-1 (eBook)
DOI 10.1007/978-3-319-09879-1

Library of Congress Control Number: 2014945766

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The ICT Innovations conference is a framework where academics, professionals, and practitioners interact and share their latest results and interests related to basic and applied research in ICT. The organizer of the conference is the Association for Information and Communication Technologies (ICT-ACT) that serve its mission to support the development of information and communication technologies in Macedonia, the Balkan region and beyond, especially in the area of research and application of innovative technologies.

The 6th ICT Innovations 2014 conference gathered 244 authors from 26 countries reporting their scientific work and novel solutions in data processing. Only 32 papers were selected for this edition by the international program committee consisting of 203 members from 47 countries, chosen for their scientific excellence in their specific fields.

ICT Innovations 2014 was held in Ohrid, at the Faculty of Tourism and Hospitality, September 9-12, 2014. The special conference topic was "World of Data". The conference focused on variety of ICT fields: Data Mining and Information Retrieval, Bioinformatics and Biomedical Engineering. Artificial Intelligence, Pattern Recognition, Big Data, Internet, Web Applications, Database and Information Systems, Wireless Communication and Mobile Computing, Digital Signal and Image Processing, Social Networking, Software Engineering.

We would like to express sincere gratitude to the authors for submitting their contributions to this conference and to the reviewers for sharing their experience in the selection process. Special thanks to Monika Simjanoska and Emil Stankov for their technical support in the preparation of the conference proceedings.

Ohrid
September 2014

Ana Madevska Bogdanova
Dejan Gjorgjevikj
Editors

Organization

ICT Innovations 2014 was organized by the Macedonian Society of Information and Communication Technologies (ICT-ACT).

Conference and Program Chairs

Ana Madevska Bogdanova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia
Dejan Gjorgjevikj	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia

Program Committee

Senthil Kumar A.V.	Bharathiar University, India
Rocío Abascal-Mena	Universidad Autonoma Metropolitana - Cuajimalpa, Mexico
Jugoslav Achkoski	Military Academy "General Mihailo Apostolski", Goce Delchev University, Macedonia
Nevena Ackovska	Ss. Cyril and Methodius University, Macedonia
Syed Ahsan	Technische Universität Graz, Austria
Zahid Akhtar	University of Cagliari, Italy
Abbas M. Al-Bakry	University of Babylon, Iraq
Azir Aliu	South East European University, Macedonia
Giner Alor Hernandez	Instituto Tecnologico de Orizaba, Mexico
Adel Alti	University of Setif, Algeria
Luis Alvarez Sabucedo	Universidade de Vigo. Depto. of Telematics, Spain
Hani Alzaid	Queensland University of Technology, Australia
Ljupcho Antovski	Ss. Cyril and Methodius University, Macedonia
Ezendu Ariwa	University of Bedfordshire, United Kingdom

Goce Armenski	Ss. Cyril and Methodius University, Macedonia
Hrachya Astsatryan	Institute for Informatics and Automation Problems, National Academy of Sciences of Armenia, Armenia
Tsonka Baicheva	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Verica Bakeva	Ss. Cyril and Methodius University, Macedonia
Valentina Emilia Balas	Aurel Vlaicu University of Arad, Romania
Antun Balaz	Institute of Physics Belgrade, Serbia
Lasko Basnarkov	Ss. Cyril and Methodius University, Macedonia
Ildar Batyrshin	Mexican Petroleum Institute, Mexico
Marta Beltran	Rey Juan Carlos University, Spain
Ljerka Beus-Dukic	University of Westminster, United Kingdom
Gennaro Boggia	DEI - Politecnico di Bari, Italy
Slobodan Bojanic	Universidad Politecnica de Madrid, Spain
Dragan Bosnacki	Eindhoven University of Technology, Netherlands
Zaki Brahmi	University of Manouba, Tunisia
Robert Burduk	Wroclaw University of Technology, Poland
Francesc Burrull	Universidad Politecnica de Cartagena, Spain
Kalinka Regina Castelo Branco	USP, Brasil
Nick Cavalcanti	UFPE, United Kingdom
Ruay-Shiung Chang	National Dong Hwa University, Taiwan
Somchai Chatvichienchai	University of Nagasaki, Japan
Jenhui Chen	Chang Gung University, Taiwan
Ivan Chorbev	Ss. Cyril and Methodius University, Macedonia
Ping-Tsai Chung	Long Island University, New York, USA
Betim Cico	South East European University, Macedonia
Boguslaw Cyganek	AGH University of Science and Technology, Poland
Ireneusz Czarnowski	Gdynia Maritime University, Poland
Ashok Kumar Das	International Institute of Information Technology, India
Danco Davcev	Ss. Cyril and Methodius University, Macedonia
Antonio De Nicola	ENEA, Italy
Zamir Dika	South East European University, Macedonia
Vesna Dimitrova	Ss. Cyril and Methodius University, Macedonia
Ivica Dimitrovski	Ss. Cyril and Methodius University, Macedonia
Ciprian Dobre	University Politehnica of Bucharest, Romania
Martin Drlik	Constantine the Philosopher University in Nitra, Slovakia
Suliman Mohamed Fati	Universiti Sains Malaysia, Malaysia
Victor Felea	"Al.I.Cuza" University of IASI, Romania
Sonja Filiposka	Ss. Cyril and Methodius University, Macedonia
Simon Fong	University of Macau, Hong Kong
Neki Frasheri	Polytechnic University of Tirana, Albania

Kaori Fujinami	Tokyo University of Agriculture and Technology, Japan
Slavko Gajin	University of Belgrade, Serbia
Joao Gama	University of Porto, Portugal
Salvador García	Universidad de Jaen, Spain
Andrey Gavrilov	Novosibirsk State Technical University, Russia
Amjad Gawanmeh	Khalifa University, United Arab Emirates
Sonja Gievska	The George Washington University, USA
Danilo Gligoroski	Norwegian University of Science and Technology, Norway
Abel Gomes	Univeristy of Beira Interior, Portugal
Manuel Grana	Universidad del Pais Vasco, Spain
Arkadiusz Grzybowski	Wroclaw University of Technology, Poland
David Guralnick	International E-Learning Association, New York, USA
Marjan Gusev	Ss. Cyril and Methodius University, Macedonia
Tianyong Hao	Columbia University, New York, USA
Nataša Hoić-Božić	University of Rijeka, Croatia
Uwe Hoppe	Bildungswerk der Sächsischen Wirtschaft gGmbH, Germany
Fu-Shiung Hsieh	Chaoyang University of Technology, Taiwan
Yin-Fu Huang	Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Taiwan
Chi-Chun Huang	Department of Information Management, National Kaohsiung Marine University, Taiwan
Ladislav Huraj	University of SS. Cyril and Methodius in Trnava, Slovakia
Albert M. Hutapea	Indonesian Adventist University, Indonesia
Hieu Trung Huynh	Department of Computer Engineering, Chonnam National University, Korea
Min-Shiang Hwang	Asia University, Taiwan
Barna Laszlo Iantovics	Petru Maior University of Tg. Mures, Romania
Sergio Ilarri	University of Zaragoza, Spain
Mirjana Ivanovic	Faculty of Science, Department of Mathematics and Informatics, Serbia
Konrad Jackowski	Wroclaw University of Technology, Poland
Boro Jakimovski	Ss. Cyril and Methodius University, Macedonia
Smilka Janeska-Sarkanjac	Ss. Cyril and Methodius University, Macedonia
Valentina Janev	The Mihajlo Pupin Institute, Serbia
Yichuan Jiang	Southeast University, China
Fr. Biju John	Amal Jyothi College of Engineering, India
Slobodan Kalajdziski	Ss. Cyril and Methodius University, Macedonia
Kalinka Kaloyanova	University of Sofia - FMI, Bulgaria
Damir Kalpic	University of Zagreb, Croatia

Aneta Karaivanova	Bulgarian Academy of Sciences, Bulgaria
Gong Ke	ChongQing Jiao Tong University, China
Ljupco Kocarev	Ss. Cyril and Methodius University, Macedonia
Margita Kon-Popovska	Ss. Cyril and Methodius University, Macedonia
Ivan Kraljevski	VoiceINTERconnect GmbH, Germany
Bartosz Krawczyk	Wroclaw University of Technology, Poland
Andrea Kulakov	Ss. Cyril and Methodius University, Macedonia
Siddhivinayak Kulkarni	University of Ballarat, Ballarat, VIC, Australia
Anirban Kundu	Kuang-Chi Institute of Advanced Technology, China
Minoru Kuribayashi	Kobe University, Japan
Eugenijus Kurilovas	Vilnius University Institute of Mathematics and Informatics, Lithuania
Arianit Kurti	Linnaeus University, Sweden
Sanja Lazarova-Molnar	United Arab Emirates University, United Arab Emirates
Nhien An Le Khac	University College, Dublin, Ireland
Shin-Jye Lee	University of Manchester, United Kingdom
Rita Yi Man Li	Hong Kong Shue Yan University, Hong Kong
Hwee-San Lim	School of Physics, Universiti Sains Malaysia, Malaysia
Suzana Loshkovska	Ss. Cyril and Methodius University, Macedonia
Gjorgji Madjarov	Ss. Cyril and Methodius University, Macedonia
Augostino Marengo	University of Bari, Italy
Smile Markovski	Ss. Cyril and Methodius University, Macedonia
Jasen Markovski	Eindhoven University of Technology, Netherlands
Cveta Martinovska	Goce Delchev University, Macedonia
Choras Michal	ITTI, Poland
Marcin Michalak	Silesian University of Technology, Poland
Marija Mihova	Ss. Cyril and Methodius University, Macedonia
Aleksandra Mileva	Goce Delchev University, Macedonia
Anastas Mishev	Ss. Cyril and Methodius University, Macedonia
Igor Mishkovski	Ss. Cyril and Methodius University, Macedonia
Kosta Mitreski	Ss. Cyril and Methodius University, Macedonia
Pece Mitrevski	St. Kliment Ohridski University, Macedonia
Irina Mocanu	Politehnica University, Romania
Ammar Mohammed	Koblenz University, Germany
Yves Moreau	K.U.Leuven, ESAT-SCD, Belgium
Radouane Mrabet	Mohammed V - Souissi University, Morocco
Viorel Nicolau	“Dunarea de Jos” University of Galati, Romania
Alexandru Nicolin	IFIN-HH, Romania
Novica Nosovic	University of Sarajevo, Bosnia and Herzegovina
Florian Nuta	Danubius University of Galati, Romania
Viji Pai	Department of Computer Applications, PSG College of Technology, India

Eleonora Pantano	University of Calabria, Italy
Joao Paulo Papa	Universidade Estadual Paulista, Brasil
Jehan-Francois Paris	Department of Computer Science, University of Houston, USA
Peter Parycek	Danube-University Krems, Austria
Shushma Patel	London South Bank University, United Kingdom
Antonio Pinheiro	Universidade da Beira Interior, Portugal
Matus Pleva	Technical University of Košice, Slovakia
Florin Pop	University Politehnica of Bucharest, Romania
Zaneta Popeska	Ss. Cyril and Methodius University, Macedonia
Hector Quintian	USAL, Spain
Dejan Rančić	University of Niš, Serbia
Ustijana Rechkoska Shikoska	St. Paul the Apostle University, Macedonia
Manjeet Rege	Rochester Institute of Technology, USA
Andreas Riener	Johannes Kepler University Linz, Institute for Pervasive Computing, Austria
Sashko Ristov	Ss. Cyril and Methodius University, Macedonia
Jatinderkumar Saini	Narmada College of Computer Application, India
Elena Serova	St. Petersburg State Economic University, Russia
Vladimír Siládi	Matej Bel University, Slovakia
Manuel Silva	ISEP, Portugal
Dragan Simic	University of Novi Sad, Serbia
Dr. Dharm Singh	MP University of Agri and Tech Udaipur, India
Brajesh Kumar Singh	Motilal Nehru National Institute of Technology, Allahabad, India
Ana Sokolova	University of Salzburg, Austria
Michael Sonntag	Johannes Kepler University Linz, Austria
Dejan Spasov	Ss. Cyril and Methodius University, Macedonia
Georgi Stojanov	The American University of Paris, France
Igor Stojanovic	Goce Delchev University, Macedonia
Toni Stojanovski	Evolve-IS, Australia
Stanimir Stoyanov	University of Plovdiv “Paisii Hilendarski”, Bulgaria
Chandrasekaran Subramaniam	Kumaraguru College of Technology, Coimbatore, India
Chang-Ai Sun	Beijing Jiaotong University, China
Kenji Suzuki	The University of Chicago, USA
Irfan Syamsuddin	State Polytechnic of Ujung Pandang, Indonesia
Jurij Tasic	University of Ljubljana, Slovenia
Ousmane Thiare	Gaston Berger University, Senegal
Dimitar Trajanov	Ss. Cyril and Methodius University, Macedonia
Ljiljana Trajkovic	Simon Fraser University, Canada
Vladimir Trajkovik	Ss. Cyril and Methodius University, Macedonia
Igor Trajkovski	Ss. Cyril and Methodius University, Macedonia
Bogdan Trawinski	Wroclaw University of Technology, Poland
Chidentree Treesatayapun	Cinvestav-Saltillo, Mexico

Yuh-Min Tseng	Department of Mathematics, National Changhua University of Education, Taiwan
Grigorios Tsoumakas	Aristotle University, Greece
Ilkay Ulusoy	Middle East Technical University, Turkey
Ventzeslav Valev	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Goran Velinov	Ss. Cyril and Methodius University, Macedonia
Nguyen Quoc Bao Vo	Posts and Telecommunications Institute of Technology, Vietnam
Boris Vrdoljak	University of Zagreb, Croatia
Wan Adilah Wan Adnan	UiTM, Malaysia
Santoso Wibowo	CQUniversity, Australia
Michal Wozniak	Wroclaw University of Technology, Poland
Lai Xu	Bournemouth University, United Kingdom
Shuxiang Xu	University of Tasmania, Australia
Tolga Yalcin	St. Paul the Apostle University, Macedonia
Wuyi Yue	Konan University, Japan
George Z. Chen	Heriot-Watt University, United Kingdom
Zoran Zdravev	Goce Delchev University, Macedonia
Katerina Zdravkova	Ss. Cyril and Methodius University, Macedonia
Xiangyan Zeng	Fort Valley State University, USA
Qingtian Zeng	Shandong University of Science and Technology, China
Defu Zhang	Xiamen University, China
Dawid Zydek	Idaho State University, USA

Organizing Committee

Vesna Dimitrova, PhD	Ss. Cyril and Methodius University, Macedonia
Gjorgji Madjarov, PhD	Ss. Cyril and Methodius University, Macedonia
Blagoja Ristevski, PhD	St. Kliment Ohridski University, Macedonia
Jugoslav Achkoski, PhD	Military Academy "General Mihailo Apostolski", Goce Delchev University, Macedonia
Saso Koceski, PhD	Goce Delchev University, Macedonia

Technical Committee

Tomche Delev, M.Sc.	Ss. Cyril and Methodius University, Macedonia
Emil Stankov, M.Sc.	Ss. Cyril and Methodius University, Macedonia
Monika Simjanoska, M.Sc.	Ss. Cyril and Methodius University, Macedonia

Contents

Invited Keynote Paper

Proceeding Papers

Challenges in Learning from Streaming Data Extended Abstract

João Gama[1,2]

[1] LIAAD-INESC TEC, University of Porto
[2] Faculty of Economics, University Porto
jgama@fep.up.pt

1 Introduction

Machine learning studies automatic methods for acquisition of domain knowledge with the goal of improving systems performance as the result of experience. In the past two decades, machine learning research and practice has focused on batch learning usually with small data sets. The rationale behind this practice is that examples are generated at random accordingly to some stationary probability distribution. Most learners use a greedy, hill-climbing search in the space of models. They are prone to overfitting, local maximas, etc. Data are scarce and statistic estimates have high variance. A paradigmatic example is the TDIT algorithm to learn decision trees [14]. As the tree grows, less and fewer examples are available to compute the sufficient statistics, variance increase leading to model instability Moreover, the growing process re-uses the same data, exacerbating the overfitting problem. Regularization and pruning mechanisms are mandatory.

The developments of information and communication technologies dramatically change the data collection and processing methods. What distinguish current data sets from earlier ones are automatic data feeds. We do not just have people entering information into a computer. We have computers entering data into each other [7]. Moreover, advances in miniaturization and sensor technology lead to sensor networks, collecting high-detailed spatio-temporal data about the environment.

These technical developments pose new challenges and research oportunities to the data mining community:

- Find the decision structure in the current window;
- What changed in the decision structure last week?
- Which patterns disappeared/appeared last week?
- Which patterns are growing/shrinking this month?
- Mine the evolution of decision structures.

In this paper we review some of the challenges in learning from continuous flow of data.

2 Algorithm Issues in Learning from Data Streams

The challenge problem for data mining is the ability to permanently maintain an accurate decision model. This issue requires learning algorithms that can modify the current

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_1

model whenever new data is available at the rate of data arrival. Moreover, they should forget older information when data is out-dated. In this context, the assumption that examples are generated at random according to a stationary probability distribution does not hold, at least in complex systems and for large periods of time. In the presence of a non-stationary distribution, the learning system must incorporate some form of forgetting past and outdated information. Learning from data streams require incremental learning algorithms that take into account concept drift. Solutions to these problems require new sampling and randomization techniques, and new approximate, incremental and decremental algorithms. [9] identify desirable properties of learning systems that are able to mine continuous, high-volume, open-ended data streams as they arrive. Learning systems should be able to process examples and answering queries at the rate they arrive. Some desirable properties for learning in data streams include: incrementality, online learning, constant time to process each example, single scan over the training set, and taking drift into account.

Incremental learning is one fundamental aspect for the process of continuously adaptation of the decision model. The ability to update the decision model whenever new information is available is an important property, but it is not enough, it also require operators with the ability to *forget* past information [13]. Some data stream models allow delete and update operators. Sliding windows models require forgetting old information. In all these situations the incremental property is not enough. Learning algorithms need forgetting operators that reverse learning: decremental unlearning [3].

The incremental and decremental issues requires a permanent maintenance and updating of the decision model as new data is available. Of course, there is a trade-off between the cost of update and the gain in performance we may obtain. Learning algorithms exhibit different profiles. Algorithms with strong variance management are quite efficient for small training sets. Very simple models, using few free-parameters, can be quite efficient in variance management, and effective in incremental and decremental operations being a natural choice in the sliding windows framework. The main problem with simple representation languages is the boundary in generalization performance they can achieve, since they are limited by high bias while large volumes of data require efficient bias management. Complex tasks requiring more complex models increase the search space and the cost for structural updating. These models, require efficient control strategies for the trade-off between the gain in performance and the cost of updating. A step in this direction is the so called *algorithm output granularity* presented by [5]. Algorithm output granularity monitors the amount of mining results that fits in main memory before any incremental integration. [6] illustrate the application of the *algorithm output granularity* strategy to build efficient clustering, frequent items and classification techniques.

In most applications, we are interested in maintaining a decision model consistent with the current status of the nature. This lead us to the sliding window models where data is continuously inserted and deleted from a window. Learning algorithms must have operators for incremental learning and forgetting. Incremental learning and forgetting are well defined in the context of predictive learning. The meaning or the semantics in other learning paradigms (like clustering) are not so well understood, very few works address this issue.

When data flows over time, and at least for large periods of time, it is highly unprovable the assumption that the examples are generated at random according to a stationary probability distribution. At least in complex systems and for large time periods, we should expect changes in the distribution of the examples. A natural approach for these *incremental tasks* are *adaptive learning algorithms*, incremental learning algorithms that take into account concept drift. Concept drift means that the concept related to the data being collected may shift from time to time, each time after some minimum permanence. Changes occur over time. The evidence for changes in a concept are reflected in some way in the training examples. Old observations, that reflect the past behavior of the nature, become irrelevant to the current state of the phenomena under observation and the learning agent must forget that information. The nature of change is diverse. It might occur, in the context of learning, due to changes in hidden variables, or changes in the characteristic properties of the observed variables. Most learning algorithms use blind methods that adapt the decision model at regular intervals without considering whether changes have really occurred. Much more interesting is explicit change detection mechanisms. The advantage is that they can provide meaningful description (indicating change-points or small time-windows where the change occurs) and quantification of the changes. The main research issue is how to incorporate change detection mechanisms in the learning algorithm, embedding change detection methods in the learning algorithm is a requirement in the context of continuous flow of data. The level of *granularity* of decision models is a relevant property, because if can allow partial, fast and efficient updates in the decision model instead of rebuilding a complete new model whenever a change is detected. The ability to recognize seasonal and re-occurring patterns is an open issue.

Novelty detection refers to learning algorithms being able to identify and learn new concepts. Intelligent agents that act in dynamic environments must be able to learn conceptual representations of such environments. Those conceptual descriptions of the world are always incomplete, they correspond to what it is *known* about the world. This is the *open* world assumption as opposed to the traditional *closed* world assumption, where what is to be learnt is defined in advance. In open worlds, learning systems should be able to extend their representation by learning new concepts from the observations that do not match the current representation of the world. This is a difficult task. It requires to identify the *unknown*, that is, the limits of the current model. In that sense, the *unknown* corresponds to an *emerging pattern* that is different from *noise*, or *drift* in previously known concepts.

Data streams are distributed in nature. Learning from distributed data, we need efficient methods in minimizing the communication overheads between nodes [15]. The strong limitations of centralized solutions is discussed in depth in [10,11]. The authors point out *a mismatch between the architecture of most off-the-shelf data mining algorithms and the needs of mining systems for distributed applications*. Such mismatch may cause a bottleneck in many emerging applications, namely hardware limitations related to the limited bandwidth channels. Most important, in applications like monitoring, centralized solutions introduce delays in event detection and reaction, that can make mining systems useless. Another direction, for distributed processing, explore multiple models [4,12]. [12] propose a method that offer an effective way to construct a

redundancy-free, accurate, and meaningful representation of large decision-tree ensembles often created by popular techniques such as Bagging, Boosting, Random Forests and many distributed and data stream mining algorithms.

In some challenging applications of Data Mining, data are better described by sequences (for example DNA data), trees (XML documents), and graphs (chemical components). Tree mining in particular is an important field of research [1,2]. XML patterns are tree patterns, and XML is becoming a standard for information representation and exchange over the Internet; the amount of XML data is growing, and it will soon constitute one of the largest collections of human knowledge.

In the static case, similar data can be described with different schemata. In the case of dynamic streams, the schema of the stream can also change. For example, in monitoring sensor networks, and social network analysis, new nodes might appear and others might diappear. We need algorithms that can deal with evolving feature spaces over streams. There is very little work in this area, mainly pertaining to document streams. For example, in sensor networks, the number of sensors is variable (usually increasing) over time.

An important aspect of any learning algorithm is the hypothesis evaluation criteria. Most of evaluation methods and metrics were designed for the static case and provide a single measurement about the quality of the hypothesis. In the streaming context, we are much more interested in how the evaluation metric evolves over time. Results from the *sequential statistics* [16] may be much more appropriate. [8] propose a general framework for assessing predictive stream learning algorithms using sequential statistics. They show that the prequential error converges to an holdout estimator when computed over sliding windows or using fading factors.

3 Conclusions

The ultimate goal of Data Mining is to develop systems and algorithms with high level of autonomy. For such, Data Mining studies the automated acquisition of domain knowledge looking for the improvement of systems performance as result of experience. These systems address the problems of data processing, modeling, prediction, clustering, and control in changing and evolving environments. They self-evolve their structure and knowledge on the environment.

The challenges and research opportunities of data streaming mining are abundant. It is one of most pleasent research areas nowadays.

Acknowledgments. This work was supported by Sibila research project (NORTE-07-0124-FEDER-000059), financed by North Portugal Regional Operational Programme (ON.2 O Novo Norte), under the National Strategic Reference Framework (NSRF), through the Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT), and by European Commission through the project MAESTRA (Grant number ICT-2013-612944).

References

1. Bifet, A., Gavaldà, R.: Mining adaptively frequent closed unlabeled rooted trees in data streams. In: Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, Las Vegas, USA, pp. 34–42 (2008)
2. Bifet, A., Gavaldà, R.: Adaptive XML tree classification on evolving data streams. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part I. LNCS, vol. 5781, pp. 147–162. Springer, Heidelberg (2009)
3. Cauwenberghs, G., Poggio, T.: Incremental and decremental support vector machine learning. In: Proceedings of the Neural Information Processing Systems (2000)
4. Chen, R., Sivakumar, K., Kargupta, H.: Collective mining of Bayesian networks from heterogeneous data. Knowledge and Information Systems Journal 6(2), 164–187 (2004)
5. Gaber, M., Yu, P.S.: A framework for resource-aware knowledge discovery in data streams: a holistic approach with its application to clustering. In: ACM Symposium Applied Computing, pp. 649–656. ACM Press (2006)
6. Medhat, M., Gaber, M., Krishnaswamy, S., Zaslavsky, A.: Cost-efficient mining techniques for data streams. In: Proceedings of the Second Workshop on Australasian Information Security, pp. 109–114. Australian Computer Society, Inc. (2004)
7. Gama, J.: Knowledge Discovery from Data Streams. Data Mining and Knowledge Discovery. Chapman & Hall CRC Press, Atlanta (2010)
8. Gama, J., Sebastião, R., Rodrigues, P.P.: Issues in evaluation of stream learning algorithms. In: KDD, pp. 329–338 (2009)
9. Hulten, G., Domingos, P.: Catching up with the data: research issues in mining data streams. In: Proc. of Workshop on Research Issues in Data Mining and Knowledge Discovery, Santa Barbara, USA (2001)
10. Kargupta, H., Joshi, A., Sivakumar, K., Yesha, Y.: Data Mining: Next Generation Challenges and Future Directions. AAAI Press and MIT Press (2004)
11. Kargupta, H., Park, B.H.: Mining decision trees from data streams in a mobile environment. In: IEEE International Conference on Data Mining, pp. 281–288. IEEE Computer Society, San Jose (2001)
12. Kargupta, H., Park, B.H., Dutta, H.: Orthogonal decision trees. IEEE Transactions on Knowledge and Data Engineering 18, 1028–1042 (2006)
13. Kifer, D., Ben-David, S., Gehrke, J.: Detecting change in data streams. In: Proceedings of the International Conference on Very Large Data Bases, pp. 180–191. Morgan Kaufmann, Toronto (2004)
14. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc., San Mateo (1993)
15. Sharfman, I., Schuster, A., Keren, D.: A geometric approach to monitoring threshold functions over distributed data streams. ACM Transactions Database Systems 32(4), 301–312 (2007)
16. Wald, A.: Sequential Analysis. John Wiley and Sons, Inc. (1947)

Agreement Technologies and Multi-agent Environments*

Mirjana Ivanović

Department of Mathematics and Informatics,
Faculty of Sciences, University of Novi Sad, Serbia
mira@dmi.uns.ac.rs

Abstract. Agreements as crucial social concepts are present in all human interactions and without them there is no cooperation in social systems. As a consequence of rapid development of different disciplines, agreement and processes for reaching agreements between different kinds of agents, getting a subject of perspective research activities. Agreement Technologies refer as well to computer systems in which autonomous software agents negotiate with one another, in order to come to mutually acceptable agreements.

The goals of this paper are to present the essential issues in Agreement Technologies and highlight its influence on multi-agent environments.

1 Introduction

One of the most important social skills human beings possess is perhaps their ability to explicitly *reach agreements* with each other. A world without agreement would be incredible. Human social skills represent an intriguing challenge for researchers in artificial intelligence: can they build *computers* that are capable of exhibiting these skills? Can they develop software systems that can *reach agreements* with each other on behalf humans? These questions present deep research challenges and has led to the emergence of a new research field, *Agreement Technologies* [27]. Agreement Technologies (AT) refer to computer systems in which autonomous software agents negotiate with one another, typically on behalf of humans, in order to come to mutually acceptable agreements.

In meanwhile a lot of high-quality research activities and initiatives emerged and significant scientific results are achieved in this area. One among most important initiatives in the area of Agreement Technologies is surely realization of big COST Action IC0801 on Agreement Technologies [20].

The rest of the paper is organized as follows. In Section 2, basic concepts of Agreement Technologies are briefly presented. Section 3 brings wider view on these concepts and their role in multi-agent environments. Last section concludes the paper.

* The work is partially supported by Ministry of Education and Science of the Republic of Serbia, through project no. OI174023: "Intelligent techniques and their integration into wide-spectrum decision support".

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_2

2 Agreement Technologies

Nowadays in different working environments people are supported by specific software components - *agents* to stress their capability of representing human interests. Such systems are built, enacted, and managed away from rigid and centralized client-server architectures, towards more flexible and decentralized means of interaction. In next-generation open distributed systems interactions between *computational agents* are based on the concept of *agreement* where two key elements are needed: a normative context that defines rules; an interaction mechanism by means of which agreements are first established, and then enacted [21]. AT paradigm is characterized by: autonomy, interaction, mobility and openness, and supported by technologies: semantic alignment, negotiation, argumentation, virtual organizations, and learning.

2.1 A Computing Perspective of Agreement Technologies

Nowadays, agreement and all the processes and mechanisms involved in reaching agreements between different kinds of agents, are also a subject of intensive research.

Software agents as specific software components are able to solve complex tasks, interact in sophisticated ways, and posses higher levels of intelligence. Services, agents, peers, or nodes in distributed software systems usually imply different degrees of openness and autonomy. Interactions between them can be abstracted to the establishment of *agreements for execution*, and *execution of agreements*.

Traditional software components remain *unchanged* at execution-time. But when software systems become open, adaptive and autonomic software components need to interact with others and adjust to changes that appear in the environment. Accordingly agreements have to be changed *dynamically* at run-time. In a long term interoperation agreements can evolve by further interaction between the computational entities. So agreements could be seen as basic run-time structures that determine if a certain interaction is correct [21]. It introduces new term “interaction-awareness” where software components explicitly represent and reason about agreements and their associated processes. There are several key dimensions where new solutions for the establishment of agreements need to be developed [2]: *Semantic Technologies*, *Norms*, *Organizations*, *Argumentation* and *Negotiation*, and *Trust*.

2.2 Agreements between Software Agents

Crucial elements of open distributed systems are *software agents* characterized by: **autonomy**, **social ability**, **reactivity**, **proactiveness**. Interactions between a software agent and with its environment must be supported by a quite complex program which includes sophisticated activities: reasoning, learning, or planning. So software agents in next-generation open distributed systems must be inevitably based on agreements including a normative model and an interaction model [21].

Agreement Technologies are getting unavoidable in contemporary systems and characteristic areas of applications are *E-Commerce*, *Transportation Management* and *E-Governance*. Researchers in the area forecast that AT will play essential role in future *smart energy grids* [23].

3 Key Dimensions of Agreement Technologies

There are several key dimensions that characterize AT: Semantic Technologies, Norms, Organizations and Institutions, Argumentation and Negotiation, and Trust.

3.1 Semantics in Agreement Technologies

Over the last several years Web has got rather matured and consists of several standards endorsed by the World Wide Web consortium: XML, RDF, Ontologies and OWL, RIF, XQuery, SPARQL. In AT these standards include new elements:

1. *Policies, Norms and the Semantic Web "Trust Layer"* - Rules and constraints that model intended behaviors represent in fact policies. Necessary standards are protocols to exchange policies and also rules languages that support describing and exchanging policies (as RIF - Rule Interchange Format and XACML - eXtensible Access Control Markup Language).

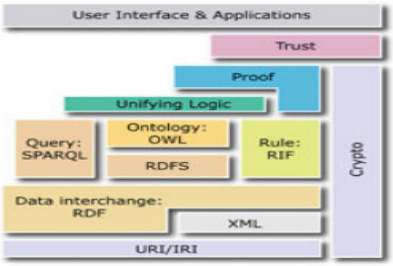

Fig. 1. Semantic web layers (2009)

Agreed policies in a community represent norms but they also can be something individual (as mail filtering policies). Formalization of (private and organizational) policies and (community) norms is important for different applications.

2. *Evolution of Norms and Organizational Changes* – Usually the evolution of norms and policies and organizational change are connected to merging and aligning existing policies and norms. Description Logics based ontology languages are not sufficient to express semantic models and policies and it is necessary to use other formalisms.

3. *Semantic Web Languages versus Norm-Based or Organization-Based Programming Languages* – "trust layer" of the Semantic Web is still in immature stage. Different protocols and languages (as P3P, XACML) are developing and it is necessary to resolve how to embed rule based and formal descriptions of and norms.

4. *Implicit Versus Explicit Norms on the Semantic Web* - Best practices and norms on the Web are not (yet) made explicit.

Logical formalisms for AT - Semantic Web standards serve for representing the knowledge of local agents, in order to achieve a goal in agreement with other agents. In distributed, open and heterogeneous systems that use AT, formalisms of Semantic Web safer of limitations. Autonomous agents define their knowledge according to their own beliefs. Semantic Web standards do not provide the means to compartment knowledge from distinct sources, so conclusions reached when using the global knowledge of disagreeing agents could be inconsistent.

Recently a number of logical formalisms in order to handle the situations appeared. They usually extend classical or Semantic Web logics [30] and the common name for them is *contextual logics* or *distributed logics* or *modular ontology languages* [13].

3.2 Norms in Agreement Technologies

Norms recently have been an issue of growing interest in agent environments and systems. They started to be important mechanisms to regulate electronic institutions and electronic commerce and also to deal with coordination and security. Study of norms, as interdisciplinary approach, includes different views and caused an innovative understanding of norms and their dynamics. Deontic logic is highly connected to norms. It is the field of logic that is concerned with obligation, permission, and related concepts. On the other hand, it is a formal system that attempts to capture the essential logical features of these concepts. Several key research questions and dilemmas connected to deontic logic are: Norm Without Truth, Reasoning About Norm Violation, Normative Conflicts, Revision of a Set of Norms, Time and Action Issues, Norm Emergence and Games, Permissive, Knowledge and Intentions.

In [5] authors proposed 'BOID' architecture that incorporates interaction between beliefs, obligations, intentions and desires in the formation of agent goals. Essential issues discussed here is that the interaction between 'internal' and 'external' motivations (deriving from norms of the agent's social context) points out several types of agents (benevolent and egocentric agent).

Constitutive Norms - In legal and social theory there are different types of norms: regulative norms describing obligations, prohibitions and permissions; constitutive norms that support 'institutional' actions - making of contracts, the issuing of fines. Constitutive norms are extremely important mechanism [4] to normative reasoning in dynamic and uncertain environments. Characteristic example is realization of agent communication in electronic contracting.

Early works of application of norms and cooperation in software systems were concentrated on simulation [3]. In meanwhile study of social phenomena had become prominent and interconnection between the social sciences and artificial intelligence born new discipline devoted to multi-agent systems. Also research in normative multi-agent systems is boosting and there is main assumption that norms are specified by the institution and all the agents in the society know about these norms ahead of

time [1]. Alternatively, researchers interested in the emergence of norms do not assume that agents know the norms in advance.

Recent works on model agents interactions based on cooperation or coordination [26] studying how norms emerge. Agents are supposed to perform few actions (e.g. cooperate and defect) and research is concentrated on studying mechanisms that facilitate small number of actions that an agent is capable of performing. An interesting approach is presented in [25] where authors propose a data-mining for the identification of norms. Quantity of domain knowledge and prior knowledge about norms an agent possesses may play significant role in norm identification.

Another limitation of current simulation-based works on norms is the lack of consideration of all three aspects of active learning on the part of an agent: learning based on doing, observing and communicating. Most studies that investigate norm emergence using simulations employing simple games have only used learning based on doing. But it is expectable that in future research authors will integrate these three types of learning in different applicable domains. Also an interesting approach recently appeared in multi-agent systems is to provide agents with the ability to identify the presence of norms through sanctions and rewards. A promising research area for the study of norms could be inclusion of humans so in different simulations agents can learn from human agents and software agents can recommend norms to humans.

3.3 Organizations and Institutions in Agreement Technologies

Open multi-agent systems and Agreement Technologies are promising technologies for organizations and institutions. Complex task or problem in organizations can be solved by appropriate declarative specifications to a number of agents, agents can work together as teams in order to solve delegated task in reaching the goals of the organization. Besides, the notion of institution has been used within the agent community to model and implement a variety of socio-technical systems. During the interaction among autonomous agents norm compliance could be ensured. Organizational perspective proposes that the joint activity inside Multi-Agent Systems regulated by a consistent body of formally specified norms, plans, mechanisms and/or structures will achieve appropriate tasks. An organizational model consists of a conceptual framework (Organization Modeling Language) in which organizational specification can be enacted on a traditional multi-agent platform or by using some organization management infrastructure (OMI) [10], [16].

Agents have to know how to access the services of the infrastructure and to make requests according to the available organizational specification. Such agents possesses

Organization Awareness skills making them able to contemplate the organization and decide whether or not to enter such a structure, to change it by setting in place a reorganization process and whether or not to comply with the different rights and duties promoted by the organization. Multi-Agent organizations exhibit basic traits that may be part of the organizational models: system structure i.e. elements that form the system and the relationships interconnecting these elements; static/kinetic perspectives: time independent/dependent description of the system.

In modern complex sociotechnical systems it is not possible to possess and keep updated all the information about the environment. Agent-oriented modeling [28] presents a holistic approach for analyzing and designing organizations consisting of humans and technical components (agents). They are active entities that can act in the environment, perceive events, and reason [28] in *sociotechnical organizations* consisting of human and software agents.

Recently several different organizational models have been developed. A lot of interesting examples of organizational model appear recently: Moise (Model of Organization for multI-agent SystEms) [15], AGR [11], TAEMS [19], ISLANDER [10], OperA [8], AGRE [11], MOISEInst [12], ODML [14], TEAM [29], AUML [22], MAS-ML [7]. For these models different modeling dimensions are presented in Table 1.

Table 1. Organization modeling dimension in some organizational models

Model	Structure	Interaction	Function	Norms	Environment	Evolution	Evaluation	Ontology
AGR	+	+	−	−	−	−	−	−
TAEMS	−	−	+	−	+	−	+	−
ISLANDER	+	+	−	+	−	−	−	+
OperA	+	+	+	+	−	−	−	+
AGRE	+	+	−	−	+	−	−	−
MOISEInst	+	−	+	+	−	+	−	−
ODML	+	−	−	−	−	−	+	−
STEAM	+	−	+	−	−	−	−	−
AUML	+	+	+	−	+	−	−	−
MAS-ML	+	+	+	+	+	−	−	−
*M*oise	+	−	+	+	−	+	−	−
VOM	+	+	+	+	+	−	−	+
Agent-oriented	+	+	+	+−	+	−	−	−
AAOL	+	+−	+−	+	+	+	+−	−

These and some additional dimensions (Organizational Environment, Organizational Evolution, Organizational Evaluation, and Organizational Ontologies) are widely present in existing organizational models.

3.4 Augmentation and Negotiation in Agreement Technologies

As other AT concepts, argumentation is also initially studied in philosophy and law. The theory of argumentation is interdisciplinary research area (include philosophy, communication studies, linguistics, psychology and artificial intelligence). In last decade argumentation has been researched extensively in computing especially for inference, decision making and decision support, dialogue, and negotiation. Generally speaking argumentation focuses on interactions where different parties plead for and against some conclusion. They are unavoidable in situations when incomplete,

possibly inconsistent information exists and for the resolution of conflicts and differences of opinion amongst different parties. Agreement also benefits from negotiation, especially when autonomous agents have conflicting interests/desires.

The nature of argumentation is predominantly modular and most formal theories of argumentation adopt that: (1) arguments are constructed in some underlying logic; (2) interactions between arguments are defined; (3) given the network of interacting arguments, the winning arguments are evaluated.

Recent work in computer science community has illustrated the potential for implementations of logical models of argumentation, and the wide range of their application in different software systems.

Furthermore any non-trivial process resulting in an agreement presupposes some kind of conflict and the need to resolve the conflict. Such conflicts may arise between different parties/agents involved in wide range of negotiating situations. In these dialogues, the reasons or arguments for offers, stated beliefs, or proposed actions can be usefully used to further the goal of the dialogue. Nowadays the key area of research is online negotiations involving automated software agents. In e-commerce systems in a handshaking protocol, a seller would simply successively make offers and have these either rejected or accepted. The exchange of arguments provides for agreements that would not be reached in simple handshaking protocols. Having it facts in mind it is clear that argumentation may be of significant value in AT.

Interesting is concept of Argument Web. The plethora of argument visualization and mapping tools [18] testifies to the enabling function of argumentation-based models for human clarification and understanding, and for promoting rational reasoning and debate. The development of such tools is a consequence of existence of pile of discussion forums on the web, and the lack of support for checking the relevance and rationality of online discussion and debate. Such tools offer possibility of reuse of *readymade* arguments authored online.

3.5 Trust and Reputation in Agreement Technologies

Computational trust and reputation mechanisms at the moment have reached certain level of maturity. Appearance of the multi-agent systems paradigm initiated an evolution in the kind of topics explored by researchers in this area. Trust and reputation models can not be treated as black boxes isolated from any other process performed by the agent. Computational trust and reputation have to be considered together with the other elements of the agents' environments.

Trust is a social construct present in everyday life. Always a person needs to interact with another person or group a certain kind of decision about trust has to be made.

As trust has vital role in society, it is interesting research areas that include apart from sociology, philosophy, economics, management, and political science also computer science community, particularly researchers from multi-agent systems [20]. Equipping intelligent agents with ability to estimate the trustworthiness of interacting partners is crucial in improving their social interactions [24]. This means that agents use *computational trust models* to assist their trust-based decisions. Trust theory

offers a diversity of notions and concepts that reveals a "degree of confusion and ambiguity that plagues current definitions of trust" [6]. This makes easier a job of computer scientists when they attempt to formalize models of computational trust in decision making processes of artificial entities. Trust could be considered twofold, first as a decision and not an act, and second as a multi-layer concept that includes disposition and decision [6]. Also it is not necessarily mutual or reciprocal. [9] introduces *situational trust* by defining trust as a measurable belief that the truster has on the competence of the trustee in behaving in a dependably way, in a given period of time, within a given context and relative to a specific task.

So to construct robust computational trust models, it is necessary to understand how trust forms and evolves. This will allow intelligent agents to promote their own trust-worthiness, and to allow them to correctly predict others' trustworthiness even in case of new partnerships.

Reputation is again a social concept as complex as trust. Interrelation between trust and reputation is rather ambiguous: reputation is an antecedent of trust, and it may or may not influence the trust; the process of reputation building is subject to specific social influences.

So it is possible to see trust and reputation as isolated constructs therefore reputation does not influence trust.

Recently in the distributed artificial intelligence several computational trust models have been proposed with intention to allow intelligent agents to make trust-based decisions. Most of them have focused on the aggregation of past evidence about the agent under evaluation in order to estimate its trustworthiness.

Although *computational reputation* is a field that has its own set of research questions different researchers have proposed models of computational trust and reputation that integrate both social concepts, assuming the perspective of reputation as an antecedent of trust [17], [24].

4 Conclusions

The paper brings some key concepts, dilemmas and aspects of usage of Agreement Technologies in open distributed environments predominantly based on multi-agent systems. These define environments that are based on norms, argumentations and trust within which agents interact. Agreement Technologies are obviously contemporary, interesting and promising research area. Its multidisciplinary and interdisciplinary character offer great future possibilities for applications in more intelligent and sophisticated artificial societies.

References

1. Aldewereld, H., Dignum, F., García-Camino, A., Noriega, P., Rodríguez-Aguilar, J.A., Sierra, C.: Operationalisation of norms for usage in electronic institutions. In: Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS), pp. 223–225. ACM, New York (2006)

2. Argente, E., Boissier, O., Carrascosa, C., Fornara, N., McBurney, P., Noriega, P., Ricci, A., Sabater-Mir, J., Schumacher, M., Tampitsikas, C., Taveter, K., Vizzari, G., Vouros, G.A.: Environment and Agreement Technologies. In: AT 2012, pp. 260–261 (2012)
3. Axelrod, R.M.: The evolution of cooperation. Basic Books, New York (1984)
4. Boella, G., van der Torre, L.: Constitutive norms in the design of normative multi-agent systems. In: Toni, F., Torroni, P. (eds.) CLIMA 2005. LNCS (LNAI), vol. 3900, pp. 303–319. Springer, Heidelberg (2006)
5. Broersen, J., Dastani, M., van der Torre, L.: Beliefs, obligations, intentions and desires as components in an agent architecture. International Journal of Intelligent Systems 20(9), 893–920 (2005)
6. Castelfranchi, C., Falcone, R.: Trust theory: A socio-cognitive and computational model. Wiley Series in Agent Technology. Wiley, Chichester (2010)
7. da Silva, V.T., Choren, R., de Lucena, C.J.P.: A UML based approach for modeling and implementing multi-agent systems. In: International Joint Conference on Proceedings of the Autonomous Agents and Multi-agent Systems, vol. 2, pp. 914–921. IEEE Computer Society, Los Alamitos (2004)
8. Dignum, V.: A model for organizational interaction: Based on agents, founded in logic. Ph.D. thesis, Universiteit Utrecht (2004)
9. Dimitrakos, T.: System models, e-risks and e-trust. In: Proceedings of the IFIP conference on towards the E-society: E-Commerce, E-Business, E-Government, I3E 2001, pp. 45–58. Kluwer, Deventer (2001)
10. Esteva, M., Rodríguez-Aguilar, J.A., Rosell, B., Arcos, J.L.: AMELI: An agent-based middleware for electronic institutions. In: Jennings, N.R., Sierra, C., Sonenberg, L., Tambe, M. (eds.) Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS), pp. 236–243. ACM, New York (2004)
11. Ferber, J., Gutknecht, O., Michel, F.: From agents to organizations: An organizational view of multi-agent systems. In: Giorgini, P., Müller, J.P., Odell, J.J. (eds.) AOSE 2003. LNCS, vol. 2935, pp. 214–230. Springer, Heidelberg (2004)
12. Gâteau, B., Boissier, O., Khadraoui, D., Dubois, E.: Moiseinst: An organizational model for specifying rights and duties of autonomous agents. In: Third European workshop on multi-agent systems (EUMAS 2005), Brussels, pp. 484–485 (2005)
13. Homola, M.: Distributed description logics revisited. In: Calvanese, D., Franconi, E., Haarslev, V., Lembo, D., Motik, B., Tessaris, S., Turhan, A.Y. (eds.) Proceedings of the 20th International Workshop on Description Logics, DL 2007, Brixen, Bressanone, Italy, June 8-10. Bolzano University Press (2007), http://ceur-ws.org/Vol-250/paper_51.pdf
14. Horling, B., Lesser, V.: A survey of multi-agent organizational paradigms. The Knowledge Engineering Review 19(4), 281–316 (2005)
15. Hübner, J.F., Sichman, J., Boissier, O.: A model for the structural, functional, and deontic specification of organizations in multi-agent systems. In: Bittencourt, G., Ramalho, G.L. (eds.) SBIA 2002. LNCS (LNAI), vol. 2507, pp. 118–128. Springer, Heidelberg (2002)
16. Hübner, J.F., Boissier, O., Kitio, R., Ricci, A.: Instrumenting multi-agent organisations with organisational artifacts and agents. Journal of Autonomous Agents and Multi-Agent Systems (2009)
17. Huynh, T.D., Jennings, N.R., Shadbolt, N.R.: An integrated trust and reputation model for open multi-agent systems. Autonomous Agents and Multi-Agent Systems 13, 119–154 (2006)

18. Kirschner, P.A., Buckingham Shum, S.J., Carr, C.S.: Visualizing argumentation:Software tools for collaborative and educational sense-making. Springer, London (2003), http://oro.open.ac.uk/12107
19. Lesser, V., Decker, K., Wagner, T., Carver, N., Garvey, A., Horling, B., Neiman, D., Podorozhny, R., NagendraPrasad, M., Raja, A., Vincent, R., Xuan, P., Zhang, X.: Evolution of the gpgp/taems domain-independent coordination framework. Autonomous Agents and Multi-Agent Systems 9(1), 87–143 (2004)
20. Ossowski, S.: Agreement Technologies. Law, Governance and Technology Series, vol. 8 (2013)
21. Ossowski, S., Sierra, C., Botti, V.: Agreement Technologies: A Computing perspective. In: Ossowski, S. (ed.). Law, Governance and Technology Series, vol. 8, pp. 3–16 (2013)
22. Van Dyke Parunak, H., Odell, J.J.: Representing social structures in UML. In: Wooldridge, M.J., Weiß, G., Ciancarini, P. (eds.) AOSE 2001. LNCS, vol. 2222, pp. 1–16. Springer, Heidelberg (2002)
23. Ramchurn, S., Vytelingum, P., Rogers, A., Jennings, N.: Putting the "Smarts" into the smart grid: A grand challenge for artificial intelligence. Communications of the ACM 55(4), 86–97 (2012)
24. Sabater-Mir, J., Paolucci, M.L.: On Representation and aggregation of social evaluations in computational trust and reputation models. International Journal of Approximate Reasoning 46(3), 458–483 (2007)
25. Savarimuthu, B.T.R., Cranefield, S., Purvis, M.A., Purvis, M.K.: Obligation norm identification in agent societies. Journal of Artificial Societies and Social Simulation 13(4) (2010), http://jasss.soc.surrey.ac.uk/13/4/3.html
26. Sen, S., Airiau, S.: Emergence of norms through social learning. In: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI), pp. 1507–1512. AAAI Press, Menlo Park (2007)
27. Shaheen Fatima, S., Wooldridge, M., Jennings, N.R.: Optimal Negotiation of Multiple Issues in Incomplete Information Settings. In: AAMAS 2004, pp. 1080-1087 (2004)
28. Sterling, L., Taveter, K.: The art of agent-oriented modeling. MIT, Cambridge (2009)
29. Tambe, M., Adibi, J., Alonaizon, Y., Erdem, A., Kaminka, G.A., Marsella, S., Muslea, I.: Building agent teams using an explicit teamwork model and learning. Artificial Intelligence 110(2), 215–239 (1999)
30. Zimmermann, A.: Integrated distributed description logics. In: Calvanese, D., Franconi, E., Haarslev, V., Lembo, D., Motik, B., Tessaris, S., Turhan, A.Y. (eds.) Proceedings of the 20th International Workshop on Description logics, DL 2007, Brixen, Bressanone, Italy, June 8-10, 2007, pp. 507–514. Bolzano University Press (2007)

A Review of Obstacles Observed while Applying Optimisation and Information Systems in Practice

Damir Kalpic

Applied Computing Department,
Faculty of Electrical Engineering and Computing,
University of Zagreb, Croatia
damir.kalpic@fer.hr

Abstract. For many years, as a natural extension to their teaching activities and scientific research, the author and his team have been offering development of non-standard and complex software applications and information systems, and information technology related consultancy. They have experienced in their home country the communist times, time of war, transitional economy and finally the status of a country in European Union. Through all these times, optimization and/or software support to decision-making have been among their preferred tasks, often actively offered to users. During these years, a significant number of cases have demonstrated obstacles that were hindering this activity. The paper attempts an analysis of these obstacles, which are rarely clearly visible and nearly never explicitly expressed. Some of them may have become obsolete, although there is evidence in revered literature that in even the most developed economies similar wrong practices still exist. Emerging of Big Data and Cloud services may remove some obstacles but it is probably not a cure for all. The reasons for obstacles principally may be located in wrong, insufficient or contrary motivation, in ignorance, and in natural resistance to any change. Motivating management practices can help in surmounting some of the obstacles.

Keywords: Information systems, Optimization, Production systems, Production costs, Linear programming, Motivation, Fractal management, Cloud computing, Big Data, Analytics.

1 Introduction

Writing software for those users who are facing practical real-life problems has been the orientation of the author's team since their employment as young electrical engineers. Although they had been placed in the Department of Applied Mathematics of the Faculty of Electrical Engineering, more than four decades ago they turned into applicative programmers, as result of strong market demand. The most appreciated challenge has always been to solve the problems beyond the direct users' demand, hoping to increase the user's satisfaction on

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_3

one hand but also to open possibilities to scientific research and publishing on the other hand. Attempts to increase the project income cannot be neglected either. We experienced a number of success stories but we have still resented the fact that win-win situations had not been granted. Multiple obstructive factors were usually present, but for the sake of simplicity and clarity, in this paper the reasons will be attributed in each case to a principal factor alone. In terms of factorial analysis, the author's estimation would be that most of the variance should be attributed to this principal factor. The cases presented in this paper derive predominantly from a limited author's personal experience, coming from a tiny, semi developed country. Much of this experience stems from communist times, a system that is most probably left behind [1]. However, the impulse for writing this paper came from the evidence that even the most developed countries may face problems with wrongly set goals. For example, in [2] a case is described where the salespersons in USA were paid according to the overall income they had achieved, resulting in damage for the company, because they maximised their income by selling a large quantity of goods below their production price. The rate of information system successful implementations, even if being solidly based on software engineering methods, is still significantly lower than in any other engineering profession. Even Ireland, for long time observed as ideal in introduction of information technologies, is no exemption [3]. Continuous attempts to improve the development and implementation methods have made the situation better. However, in [4] the authors discuss the difference between real progress and hype. If new methods are devised, but without any technological novelty supporting it, the hype cycles become more frequent, more expensive and useless. The paper attempts to highlight the reasons for difficulty or failure that are not attributable to deficient technology or methodologies but rather on factors that can be tamed if enough of good will and full engagement of the stakeholders were exercised. The lack of this good will may derive from wrong motivation, from ignorance, and resistance to change. In final effect, there could always be just a single reason, stemming from someone's ignorance. That "someone" is not necessarily the direct user. The most absurd user behaviour can surprisingly be explained with wrongly set goals or requirements and accordingly motives coming from some higher authority, be it an ignorant manager, corporate leadership, incompetent government leadership or an ill-conceived political system. The above statements shall be illustrated with examples coming mostly from the author's personal experience. The names of the concrete companies or persons shall be omitted for ethical reasons. Finally, the emerging Big Data and its implications on removing the obstacles are speculated.

2 The Obstacles on the State Level

2.1 Inflation

Inflation brings the motives for economic subjects to behave outright opposite to what is generally expected in Operational Research literature. In last years of the

former Yugoslavia, inflation was staggering and bureaucratic rules strict. Therefore, the actual monetary transfers among ordinary citizens were performed, or at least calculated, in some hard currency, traditionally in Deutsche Mark. Meanwhile, the legal regulation ordered for official bookkeeping in companies to be recorded in local inflation-affected currency, the dinars. Some examples of practices, harmful for the whole society, but reflecting also to computerisation issues follow.

Rushing of Expenditures and Delaying of Income. Computerisation of import of foreign journals as reported in [5] was the first local multiuser microcomputer-based business application, developed for the major Croatian publisher and books and journals trader. Besides the many reasonable and logical requirements, there was also one that should help the user to postpone charging its local customers in local currency for the subscription to foreign journals. At the same time, the company was rushing to pay in hard currency in advance to foreign suppliers. It is clear that such practice results in net monetary loss for the country. However, for the company, the situation was reversed. The hard currency, necessary to pay the foreign journals, would be immediately converted from dinars and recorded in dinars in bookkeeping at the moment of payment. With a year of delay, an equivalent amount in dinars of the already paid price in hard currency would be charged from the customers, but according to the actual exchange rate. With inflation in order of magnitude of 100% p.a., the numerical differences in amounts expressed in dinars, and the corresponding phony earnings were impressive.

Maximization of Waste. A well known local factory producing cutlery and stainless steel dishes had a department of some 5 professionals who designed the cutting schemes for large stainless plates in order to produce cutlery from stripes and dishes from the principal part of the plate. They ordered from the author's team software to help them in their work. A PC based program was developed, surpassing the users' expectations [6]. The proposed cutting schemes corresponded to the demand of certain products and the utilisation of the stainless steel as raw material went up to 98%. As result, the company paid for development, but they never even installed this software. The reason can be twofold. The professionals in the design department may be out of work and substituted by the software, while the official explanation to the authors was that they actually earn most due to waste. The reason was also in inflation. The stainless plates as the raw material were imported and paid in hard currency. The cost was recorded in the company's bookkeeping in dinars at the current exchange rate. Some time later, leftovers were re-exported for correspondingly lesser amounts in hard currency. However, after having converted and recorded that amount into the local inflated currency, it turned to be a greater amount in dinars than was the purchase price of the whole plate. Again, phony earnings took place.

2.2 Wrong Policies on the State Level

Socialization of Losses. To a well established cattle food factory in the former Yugoslavia the author's team offered affordable optimisation software for microcomputer, based on proprietary developments [7]. For an illustration, using the provided data, the authors demonstrated a possibility for serious savings, while even increasing the quality of the output product. The author's offer was turned down because at that time existed a compulsory solidarity agreement forcing all the enterprises within specific branch in the whole country to respect similar salary structures and amounts. Any surplus in earnings would be taxed for this solidarity fund. The addressed company had no interest to improve its business process for the sake of this solidarity. They would rather enjoy the possibility to be comfortable behaving suboptimally, knowing that in case of unexpected difficulty, they would have enough reserve to neutralise it.

2.3 Public Procurement Act

Public Procurement Act has its aim as a guardian against corruption and wasting of tax payers' money by public servants and state owned institutions. However, the author's local experience implies that it may be counterproductive in numerous cases. Only two examples from our own experience shall be presented here.

Automatic Coding of Census Data. In 1990 the author's team conducted a very successful project of automated coding of the census data [8]. It was applied in Croatia and in Bosnia and Herzegovina in the wake of breaking of Yugoslavia. The author's team developed algorithms appropriate for the idiosyncrasies of the very flective Croatian language, where words are subject to substantial changes through cases, genders and tenses. The results achieved had met the highest expectations. The job was completed after a few months. Ten years later, the census was also performed but the author's offer to significantly upgrade the program according to technological developments, was rejected in favour of a slightly cheaper off-the shelf Canadian program, which was probably excellent for English, but hardly applicable for the Croatian language. The State Statistics Institute defended their wrong decision even if the delay of results of the last census, performed in 2011, was measured in years. For them, it seems to be most important to avoid any possible legal remark concerning the procurement, regardless of the actual costs and benefits.

Computerisation of Higher Education. The authors' team developed Student administration and Subsidised students aliment software [9], but they were reluctant towards development of the legally regulated administrative part, regarding it as a task for dedicated bookkeeping-oriented software companies, rather than for a university-based team. Therefore, public competition was announced for procurement of this software. The author was member of the committee for procurement. In order to make the bidding outcome feasible, the

committee fixed the ERP platform in its requirements specification. The author agreed, even if he used to be very critical about the same platform [10]. If it were not fixed, a myriad of bidders would show up and using legal instruments, could indefinitely postpone the procurement. The result of applying this platform was mostly successful at the author's institution that served as pilot project, but it faced difficulties on the University level [11] due to lack of proper motivation and hardly hidden obstruction, deriving probably from reluctance to change. However, nobody in charge would take the risk of openly declaring this hypothesis. It turns out to be too dangerous to criticise some behaviour and offend someone, or to obtain the feared label of incorrectness. The project will be probably left to die-off *naturally*.

Computerisation of Croatian Forestry. The cooperation of the author's team with the Croatian Forestry began in 2005. Forestry is a large and profitable state-owned institution. Their information system was developed through years by forestry engineers who evolved into programmers. The setback of their solution was fragmentation and obsolete platform. The advantage was in good understanding and support for processes in forestry. The forestry professionals, familiar to some extent both with forestry and computing, had already visited some countries with supposed model forestry, like Austria and Finland but they found out that their ways differ significantly from the situation in Croatia and their software could not serve the local purpose. General purpose ERP systems were offered as well, principally SAP, but they could not provide the essential professional functionality as required by the Forestry. The author took the view that the knowledge of the local developers and their deep understanding of forestry processes should be used as advantage [10]. The proposed idea was to update the computing knowledge of the Forestry's programmers, to teach them how to design databases and how to use the development environment based on C# and .NET, produced by the author's team. After the strategy had been accepted in the Forestry, the first problem arose when subsequent steps had to be performed. New bidding was necessary to obey the Public Procurement Act. There was no serious competition expected, who would be able, and what is even more important, who would be willing to take over such a demanding, risky and non-standard, unique, and non-repeatable task. The tender had to be formulated in the way that nobody could be accused that it was modelled in favour of the author's team. The author's team was selected as the most favourable, but the competition, even if far from being able to offer anything worth mentioning, managed to annihilate the bidding and postpone the project for more than half a year. At the next bidding, they did not even participate, as they understood that real work is required, rather than some formal design or recycling of standard texts teeming with references to the most modern methodologies. Neither punishment nor black list for such behaviour is applied. Similar situations were repeating but with different protagonists, on the edge of blackmailing. Some irrelevant company wished to be engaged as subcontractor; otherwise they threatened to abolish the bidding. In this way, serious delays were happening,

but in 2010 the education of the Forestry staff for the author's team development platform was completed. However, the final bidding for the completion of the new version of their information system, that would proceed under our couching and with using our development tool for which the Forestry staff had been educated, finished with the selection of another company who had offered a significantly lower price, obviously without being aware of the complexity of the work ahead. The author's unofficial but quite probable information is that the project was abandoned. It would not come as surprise to us to be called to resume the project after few years, while a decade would be lost due to inadequate legislation.

3 The Obstacles on the Corporate Level

3.1 Wrong Corporate Policy between Departments

Oil Transportation. In late 80-ties of the past century an important local oil supplier ordered software for minimisation of transportation costs. A database and linear programming software were incorporated into the application made on purpose to meet the users' demands. Transportation plans were made each month. The users were steadily demonstrating their dissatisfaction. They kept complaining that the plans were unrealistic and that they have to edit them in order to make them feasible. The author tried desperately to find the reasons for their dissatisfaction but it was mostly futile. Any further improvement in the optimisation was encountered by grim faces. The author believed to have discovered the reason for dissatisfaction after a user's statement that their drivers have got the habit to go one route while the program sends them every time elsewhere. A multi-criteria feature was added to the program to allow compromise between the minimum cost and respecting of habits. The only useful result of this effort was an internationally refereed publication [12]. As the users understood that the author's efforts to please them shall not be abandoned, they finally admitted that the salaries of their department had been negotiated as a percentage of all the costs of the fuel, as delivered to the customer, including the transportation costs. Minimisation of transportation costs lowered their salaries. They ordered the program against their own interest, upon insisting of the refineries. As outcome, they wished that everything stayed as it was before, but that the rest of the company could believe that the solutions were optimal.

Oil Procurement. Using the oil purchases data in a major refinery, the conditions on the spot oil market were analysed. It was not attempted to forecast future selling prices, as it was known to be a hard problem. Instead, the policy of replenishment of oil reserved was targeted [13]. Historical data were partly used to train a neural network and the other part was used to test it. Testing has shown that significant savings could be achieved if oil replenishments had proceeded according to neural network's results. However, there was no motivation or interest in the company to try using this decision making software tool. The decision makers' interests were probably somewhere else.

Insurance Company. As reported in [14], the author's team performed a quick analysis of reasons for failure of a serious computerisation project in the major Croatian insurance company. The project was stalled by the company top management after considerable amount of time and money had been spent. The amount remained secret to the author. The author's team knew the stakeholders and harboured some prejudice about what might have been the cause of failure. The project was performed by outsourcers. The project leader was known for his fascination with formal methods and tools. The implementing company was also known for absolute preference of CASE tools even if at that time they were hardly meeting the expectations. At the first glance, these two reasons could explain the failure. However, the author had to admit to himself that, nevertheless, the project leader was an experienced person and the software development company was among the best in its branch. The initial project design, containing meta-data model, indicated a high quality approach. Therefore, the author's relatively weak prejudice could not explain the outright failure of the project. Some investigation was indispensable. It turned out that the company management engaged outsourcers, even if a strong in-house computing team was available, without having asked for their opinion. This move had caused passivity among the in-house IT professionals. On the other hand, the company had its branches and offices all over the country. Local branch directors enjoyed great level of independence and were mostly out of control. They were engaging local computing companies at their choice and were creating business reports that could hardly be checked. Without accusing anyone, and without even attempting to gather any accusing evidence, it seemed obvious that non-existence of a high quality up to date and accurate information system was not among their priorities. So nearly nobody inside the company did care about the new information system development. After having completed a good and very advanced design, the designers were engaged in another project. Inexperienced programmers were sent-in instead, to complete the job relying heavily on CASE tools. They were not allowed to depart from the original project even if the prospective users were suggesting them to. One of the most important applications, collecting the routine insurance payments from customers, instead of being simple, it was designed to cover all the business activities of the company, as it would be appropriate for some isolated offices. It was completely inappropriate for the majority of administrative force. It prolonged the duration of the most common routine processes for an order of magnitude. That was the final trigger for the top management to stop the project. The author's team tried to learn something from this experience and applied it when, few years later, a new strategy for computerisation of the same company was ordered. In this new strategy, which is according to available information, accepted in practice and in active use, the role and motivation of the insiders was in no way neglected. It was even stressed that the chief information officer (CIO) should take part in the highest board of directors in the company, as was confirmed in [15]. Beforehand, CIO was mostly regarded as a person to provide technical assistance

to the business and not like nowadays, as a major stakeholder in improving the business performance.

The author's team experience regarding how to manage themselves was also taken into account [16]. At the time when the author's team was still close to the "magic" number of 7 commensurable members, fractal management was possible [17]. Flat organisation, mutual evaluation and personal income determination by secret mutual balloting were for some time highly motivating. It was all dismissed when the author's team grew over 20 members, including also young and inexperienced ones, so that a split into multiple groups and regression to more classical management methods had to be done.

3.2 Wrongly Set Goals

Minimization of Costs. Minimisation of costs on the level of states is a disputed goal in today's global economy. Many authorities insist that it can only ruin the economy. Application in an ordinary enterprise can turn senseless. A brief notice in the daily press few years ago reported about a prize-winning Cuban factory for achieving the maximum savings in comparison to the previous year. The prize had to be returned after the authorities found out that the factory achieved that goal by reducing any activity to the level zero.

Maximization of Income. Maximization of income is a very dangerous goal. Among the first professional experiences the author had [18], was the production planning in metal industry using linear programming. For the sake of the model formulation, from the raw data about manufacturing, the contribution was calculated for every product and for each variant of its production. It was the first time for that factory to notice that the most favoured product, a very expensive wire, was incurring direct manufacturing costs significantly exceeding its high and seemingly attractive selling price. Therefore, the efforts in the past to produce as much as possible of that product, wrongly favoured by the sales department, were leading directly to economic damage.

Focusing on Irrelevant Goals. The automobile parts factory that engaged the author's team in difficult time of war in Croatia [19], initially had the goal of properly distributing the fixed costs among the final products. The author had already experience with such a task. In [20] the production mix was achieved using linear programming for maximisation of contribution, defined as the difference between selling price and direct manufacturing costs, for each version of every article. The linear programming model did not contain any fixed costs. However, the middle management wanted to present to their superiors how successful they were. Their envisaged argument would be that all the products were profitable. For that purpose, they wanted the fixed costs distributed to the quantities of final articles in the optimum production mix. The distribution was attempted according to different criteria: machines engagement, material consumption, direct costs, selling prices and finally contribution. The at that time

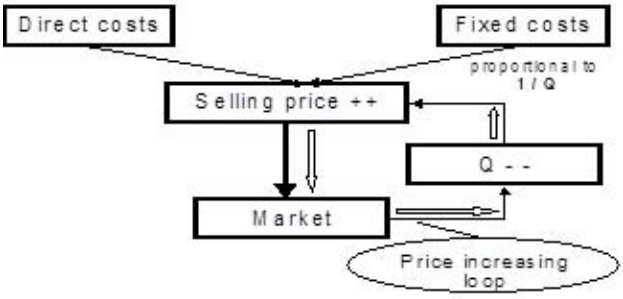

Fig. 1. Harmful consequence of distribution of fixed costs to the quantity Q of the final article

young and inexperienced author did not expect much interest for presentation of his paper, so he tried (unsuccessfully!) to explain to the audience that he just wanted to participate at a conference in a nice place and that the whole topic was senseless. As the total contribution was surpassing the total fixed costs, it was obvious that the production as a whole was profitable. It was also obvious that if the alleged profitability of every final article should be proofed, the fixed costs had to be distributed proportionally to the articles' contribution. Except for bluffing, there is no use of such fixed costs distribution.

Distribution of fixed costs to the final products can make sense in investment programs, but in operational planning, it may be even harmful, as shown in Figure 1 taken from [21] and described further on.

Instead that the sales department finds out what quantities and at what price could be sold, the selling price is calculated allegedly exactly. Let Q be the planned quantity of some final article to be produced and sold in the next planning period. If the fixed costs that someone arbitrarily adjoins to this final article are F, than each piece of it accrues the fixed cost F/Q. Let us add to it the cost d, required for manufacturing of a piece of that final article. The consideration usually proceeds that the cost of a piece of this article is $d + F/Q$. In order to gain certain percentage p, the selling price is calculated as $s = (d + F/Q) * (1 + p)$. This selling price would assure the covering of fixed and direct costs, and it would bring profit proportional to p. Such reasoning would be appropriate in an autarchic monopolist market, but some people have not yet noticed that such conditions were nonexistent today. What happens instead, is that competition offers a substitute article at lower price and the quantity sold of the considered final product drops. Keeping the above-described algorithm alive, the factory divides the same fixed costs to a smaller quantity, what increases the selling price, the sales drop further, until the loop ends with manufacturing extinguished. Had they been lowering the selling price as long as it is greater than the direct manufacturing cost, the article would increase the overall contribution. As long as this contribution could cover the fixed costs, the manufacturing would have survived.

The author was lucky that the mentioned automobile parts industry already had their own experience of extinguishing a contribution-bringing manufacturing, due to the thinking as illustrated in Figure 1. They remembered well how wrong that decision was and that it had preserved only the burden of covering the fixed costs. Thanks partly to that, and to an extremely difficult situation in the time of war, the factory top management accepted the idea of production planning with linear programming, based on direct costing, with maximizing of the contribution, as the difference between income and direct cost. The factory is nowadays well and alive, present with shares on the Croatian Stock exchange, while many industries have perished in the meantime. Among merits for this information systems and operational research success, the motivation should not be forgotten. The times were extremely hard and no other exit was in sight, but to succeed.

4 Obstacles to Proper Use of Data

In early days of computerisation and especially in less developed countries, the complaint mostly heard was that there were no available reliable data. At the time when data were entered from paper documents, a posteriori, and served practically only for delayed bookkeeping, the argument was mostly true. However, in this way were dismissed some optimisation attempts in production planning. It served sometimes as an alibi for not doing something against partial interests, as the already illustrated reluctance to bring optimal decisions. When data started to be entered as part of the work process and not for nearly archive purposes, the argument about lack of reliable data has waned. New obstacles emerged. Data privacy is in author's opinion one of the mostly abused arguments, hiding the proper motives, as follows:

Fear of Finding the Truth. Two major software solutions aimed at the higher education were developed on the author's department. One was the Student administration system [9] where all the relevant events in a student's curriculum are recorded; from the student's achievement in the secondary school, success on the faculty entrance examination, what is recently substituted with equally valid results of the state matura (leaving certificate from the secondary school), enrolment of courses to recorded success in all the examinations. Under the pretext of data privacy, all these data are hidden and accessible only to very few persons bearing high functions in higher education, like the dean and vice-dean for education. They are usually highly engaged, and can hardly find time to analyse these data. In this way, even the mentors are not aware of the qualities of their students, not to tell about the other population. Most of the students study on taxpayers money or are heavily subsidised, even if they are paying something. Their achievements or sometimes "achievements" are secret not only to a general taxpayer but also to their parents who most often sustain them [22].

In parallel to Student administration software, a credit card system for subsidised aliment of students was introduced [23]. Immediately it had shown interesting evidence of cheating practices. Some students would only buy packed

resalable products like pudding or yoghurt and would earn on subsidised prises. Instead of using the system to eradicate fraudulent practices, data privacy was introduced so that the abusers were protected and financed further on by the taxpayers.

Few years ago both systems demonstrated the power of data secrecy when some students of a large faculty initiated a student strike demanding no participation in payment even for repeating students and asking for protracted subsidised aliment. This strike caused a serious concern among some higher state officials who took the students' demands very seriously and did not dare to oppose them directly. However, in author's opinion, it would have been enough to present the study achievements of the strikers publicly on Internet. It would most probably defuse their protest while becoming obvious why they insisted to enjoy protracted subventions from the taxpayers.

Attempt to Avoid Legal Order. High concern about the data privacy can well serve to protect the villains. In 2002 in Croatia the personal identification number was practically banned, already for decades well established and in heavy everyday use. It was done under the pretext that that number was revealing all the personal secrets. The author was engaged in a government committee to solve the problems that would emerge without this identification number [24]. As the only immediate remedy, he suggested to keep the existing ID number or, if that were inacceptable, because that number revealed the date of birth and gender of the person, to introduce immediately a non-revealing number on the new identification cards. This new non-revealing number was introduced, not on the new identification cards, but seven years later after much damage and costs and after having granted a significant grace period to the abundant tax evaders.

Attempt to Do Nothing. High concern for the data privacy can well serve to hide the intent to do nothing. State administration in Croatia still requires from citizens a lot of visits to offices and transferring of papers, instead of transferring data in electronic form. The situation is improving, but without the alleged fear for data privacy, it could have been faster and more efficient. The author witnessed an interesting discussion on a recent conference; one participant expressed his concerns that in computerisation of a hospital, data privacy might be endangered. Another participant stated that those projects, where at the very beginning too much concern is expressed regarding data privacy, are usually doomed to fail.

Attempt to Avoid Responsibility. For the author it is difficult to understand data security restrictions imposed in some medical applications. For example, there was a requirement to restrict for doctors data access to digital patient records, according to their specialisation. The author strongly opposed such approach constructing an example where an ophthalmologist would try in wane to cure the vision of his or her patient, not being aware of the patient's diabetes.

Instead of aiming towards a more holistic approach, which is probably lacking in the Western medicine, data privacy is used to foster further fragmentation and possibly reduce the doctors' responsibility?

5 The Emerging Role of Clouds, Big Data and Analytics

Nowadays, a long time cherished excuse, while obstructing computerisation, was that there were no data available. This had been recognised by some professionals already decades ago and they would counter with stating that there was no motivation rather than no data. The author agrees with the latter ones and is happy to see incredible amounts of data emerge. The introduction of clouds in computing resembles to establishment of utilities like electrical power, gas, water etc. Telecommunications might be the closest example. If major accidents are avoided, computing in cloud will free the people of fears regarding the loss or abuse of their data if they are stored remotely, as it is the case with their money. It is well known that credibility of banks, if once tarnished, requires much effort and time to be reinstated. Supposedly, the situation with clouds could be similar.

Big data and analytics might require a revision, or to put it better, restoration of approach to profession and education. There were times when general education and education accordingly, were required and highly appreciated. Rather recently, it was proclaimed that a Renaissance person were an obsolete concept, not needed anymore. The tendency was towards highly specialised experts who were described as "knowing everything about nothing". General picture is again in demand. Big data enable statistical analyses, predictions and forecasts even if the exact mathematical model or causalities and reasons to explain certain observations are not known [25]. Emergence of enormous amount of data proves that quantitative change leads to qualitative change, something that already the old philosopher Karl Marx had noticed. The role of sophisticated statistical methods can increase but also change focus. Instead of making conclusions based on samples, all the data shall be examined and some formerly unnoticed details discovered [25]. The quantity of data allows for less accuracy and less insisting on correctness of each stored fact. It becomes less important to cleanse the data, what for long time has been a never ending task. Relational databases with fixed schema allegedly give way to noSQL. Data may be changed constantly and arriving from many sources. Data consistency in time cannot be assured anymore.

In addition to number crunching, analytics applied to Big Data yields the hints how to create new values. A professional in this field should be able to transfer convincingly the information to a heterogeneous set of stakeholders. Therefore, communication skills are also required. Additional education in that aspect may also improve the communication within our societies, where difference in opinion is often regarded as open hostility.

Natural language processing and computer translation among many languages was substantially improved due to enormous amount of entered texts, although many of these texts are far from being perfectly correct. Still sometimes the

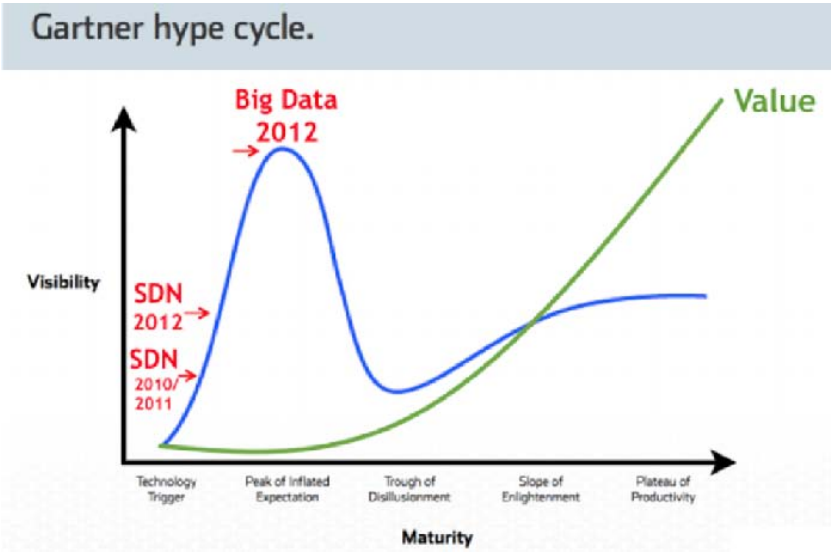

Fig. 2. Position of Big Data in the Gartner hype cycle [27]

translations are outright funny and can be well illustrated to people who are familiar with both involved languages.

Data privacy which has been sometimes misused in order to undermine certain computerisation projects, changes its aspects as ever more people voluntarily submit their data to social networks and other data collecting systems like loyalty cards. Now an individual can be endangered due to recognition of his or her behaviour patterns leading to unfavourable predictions, like being a too risky person to grant a loan. Due to Big Data, it has become of less importance to understand the causal relationships. The question What replaces the traditional Why. In selling products, it can be found that some very different articles go together, even if there is no hint why is it so. One can predict future behaviours from pure statistic models. New processing technologies like Hadoop have become attractive. Instead of classical data warehousing built with ETL procedures, huge amounts of data are processed where they are. The accuracy suffers but where this is not critical, the benefits of speed prevail. Surely, the hype cycle [4] is also present and illustrated in Figure 2. In [26] the criticism regarding Big Data is present. Cases are presented where not understanding Why turned out to be detrimental. While the values of shares and investment funds may stochastically vary in time, handling someones bank account as a statistical entity depending on Big Data would be hardly satisfactory. Accurate, well structured and correct data are still needed.

The current situation regarding Big Data is well illustrated in the sentence: "Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it" [28].

We can expect in the near future some disappointment over the inflated expectations of Big Data and then after few years Big Data would reach its stability level and coexist simultaneously with the accurate, well structured SQL operated data. Each of the data aspects will have its reasons, advantages and setbacks. Probably, there will always be some individuals wishing to obstruct anything the others attempt, be it by criminal acts, hacking, forgery or phony concerns for the human well being. Therefore, when something goes wrong in a computerisation project, do not forget to ask for true motivation.

6 Conclusions

All these case studies and examples are gathered here in order to document the authors' belief that hectic progress in computing is partly artificial while new tools, new methods, new programming languages and new interfaces are constantly devised. It would be absurd to negate the progress, but technical improvements have limited scope and require time to be properly absorbed in society. Some are new ideas and concepts and do positively affect our lives. Clouds, Big Data and Analytics might be among them. Nevertheless, there is surely high motivation for creating hype among the main stakeholders in the IT industry, as they gain profits in sales of new releases. The author supposes to be not the only one who posed the question why a certain technological change was necessary, while some well-known deficiency remains untouched? The unnecessary complexity of numbering the headings in recent versions of MS Word could be a good example. Another obvious problem has faced anyone who tried to translate a PowerPoint presentation to some other language. It is hardly conceivable that one cannot simply change the language for the whole file. Instead, every text box must be separately addressed. It is difficult to understand the mind of the one who decided to determine the language automatically from the keyboard setting. Obviously, it must be someone who cannot imagine that a single person can be able to write in two different languages. At the same time innumerable unnecessary improvements have been performed.

A part of the success-increasing potential lies however not only in technology nor in methodology, but in proper motivation, willingness, and capability to establish positively-acting organisation. The computing professional's task is to understand and if necessary modify it, and try to meet the users' requirements. Computing professionals should have relationships with their users, as do the good doctors to their patients. Empathy and ethical behaviour are essential, nearly as much as the technical competence. Improvement of the latter aspect brings direct profit, while the rewards for improving of the first two should pay-off in a longer run.

References

1. Kalpic, D., Boyd, E.: The politics of irm: lessons from communism. In: IRMA Conference. pp. 72–73 (2000)
2. Buytendijk, F.: Performance leadership: The next practices to motivate your people, align stakeholders, and lead your industry. McGraw Hill Professional (2008)
3. (2008), http://www.qas.ie/company/data-quality-news/irish_it_projects_have_just_10_success_rate_903.htm
4. Fenn, J., Raskino, M.: Mastering the hype cycle. Harvard Business, Cambridge (2008)
5. Kalpić, D., Popovi, K.: Microcomputer based software for import of journals. In: 8th International Conference Computer at the University Proceeding. SRCE, Cavtat (1986) (in croatian)
6. Mornar, V., Kalpić, D.: Two stage two dimensional cutting stock problem. In: 9th International Conference Computer at the University Proceeding. SRCE, Cavtat (1989)
7. Kalpić, D., Mornar, V.: Interactive multicriterial linear programming system. In: EURO VIII Abstracts, Lisbon, Portugal, pp. 123–124 (1986)
8. Kalpic, D.: Automated coding of census data. Journal of Official Statistics 10, 449–463 (1994)
9. Kalpić, D., Baranović, M., Mornar, V., Krajcar, S.: Development of an integral university management system. In: International Conference on System Engineering, Communications and Information Technologies, ICSECIT (2001)
10. Kalpić, D., Fertalj, K.: Development of a new information system for croatian forestry. In: ISOne World 2007-Engaging Academia and Enterprise Agendas (2007)
11. Mornar, V., Fertalj, K., Kalpić, D.: Introduction of sap erp system into a heterogeneous academic community. In: Power Control and Optimization: Proceedings of the 3rd Global Conference on Power Control and Optimization, vol. 1239, pp. 388–395. AIP Publishing (2010)
12. Kalpić, D., Kenda, S., Mornar, V.: Microcomputer multicriteria transportation system. In: EURO IX, TIMS XXVIII, Joint International Conference, Abstracts, Paris, France, pp. 218–219 (1988)
13. Domijan, P., Kalpić, D., Petrović, I.: Speculative trading strategy of buying crude oil for refinery on spot market. Forecasting Financial Markets (June 2004)
14. Kalpić, D., Fertalj, K., Mornar, V.: Analysis of reasons for failure of a major information system project. In: Kamel, S. (ed.) BITWorld 2001 Conference Proceedings CD-ROM, pp. 1–8. The American University in Cairo, Cairo (2001)
15. Kitzis, E., Broadbent, M.: The new cio leader: setting the agenda and delivering results. Harvard Business School Publishing, Massachusetts (2005)

16. Kalpić, D., Baranović, M., Fertalj, K.: How to organise a university based r&d and teaching group in computing? a case study. In: Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics, vol. 2, pp. 174–181 (1997)
17. Shin, W.: Paradigm shift in the corporation: The fractal factory. In: Proceedings of the Sixth International Conference on Manufacturing Engineering: Manufacturing; a Global Perspective, p. 275. Institution of Engineers, Australia (1995)
18. urek, M., Kalpić, D., Karin, R., Maaovi, P., Epi, A.: Optimum production mix on computer. In: Proceedings of Informatica 1974, Bled, Slovenia (1974) (in croatian)
19. Kalpić, D., Baranovic, M., Mornar, V.: Case study based on a multi-period multi-criteria production planning model. European Journal of Operational Research 87, 658–669 (1995)
20. Jankovi, K., Epi, A., Urek, M., Kalpić, D.: Production model and distribution of fixed costs among the final production mix. In: Proceedings of Informatica 1976, Bled, Slovenia (1976) (in croatian)
21. Kalpić, D.: Reliance on own expertise vs. imported solutions-case study of a small medium developed country. In: The 3rd Conference on Documentation & Electronic Archiving "Knowledge Investment & Management for Decision Support" (2005)
22. Kalpic, D.: Computerisation, data privacy and scientific excellence; where are we going? In: 30th International Conference on Information Technology Interfaces, ITI 2008, pp. 89–96. IEEE (2008)
23. Mornar, V., Fertalj, K., Kalpic, D., Krajcar, S.: Credit card system for subsidized nourishment of university students. Annals of Cases on Information Technology 4(2002), 468–486 (2002)
24. Kalpic, D.: I vladina grupa za matini broj (a government team for the personal identification number) (2003)
25. Mayer-Schönberger, V., Cukier, K.: Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt (2013)
26. Harford, T.: Big data: are we making a big mistake. Financial Times Magazine (2014)
27. (2014), http://ipcarrier.blogspot.com/2013/03/demystifying-software-defined-network.html
28. Big data is like teenage sex: 4 ways to unravel the mystery (2014), http://www.linkedin.com/today/post/article/20140326154716-2967511--big-data-is-like-teenage-sex-4-ways-to-unravel-the- mystery

Magnetic Response Properties of Aqueous Aluminum(III) Ion: A Hybrid Statistical Physics Quantum Mechanical Approach Implementing the Map-Reduce Computational Technique

Bojana Koteska[1], Anastas Mishev[1], and Ljupco Pejov[2]

[1] Faculty of Computer Science and Engineering,
Rugjer Boskovikj 16, 1000 Skopje, Macedonia
{bojana.koteska,anastas.mishev}@finki.ukim.mk
[2] Faculty of Natural Sciences & Mathematics,
Institute of Chemistry, 1000 Skopje, Macedonia
ljupcop@iunona.pmf.ukim.edu.mk

Abstract. In the present study, we make a computation of magnetic properties of aqueous Al^{3+} ion. To account for the fluctuating character of the condensed-phase environment, we first carry out a statistical physics Monte Carlo simulation of Al^{3+} aqueous solutions, followed by subsequent quantum mechanical computations of magnetic response properties of charge-embedded clusters with varying size and complexity. In particular, we address in details the issue of proper representation of the bulk solvent long-range electrostatic influence on these properties, by the averaged solvent electrostatic configuration (ASEC) approach, which is much simpler and computationally less demanding than the more widely used averaged solvent electrostatic potential (ASEP) methodology. In our particular case, we implement the ASEC computational method using the map-reduce technique, which appears to be extraordinarily suitable for the computations in question. We consider both the fundamental aspects concerning the development of computational method and the computational aspects related to the efficiency of the computational process. In particular, we address the application of the map-reduce computational technique to the particular phases of computation within the developed methodology.

Keywords: Map-Reduce computational technique, aqueous Al^{3+} ion, hybrid statistical physics, Monte Carlo simulation, quantum mechanical approach.

1 Introduction

Thorough understanding of the properties of ionic/molecular species in condensed phases is crucial for better understanding and realistic modeling of processes taking place in biochemical, geochemical and other environments. However, development of reliable and robust theoretical models that explicitly include the influence

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_4

of condensed phase environment, in particular its dynamical characteristics, appears to be far from a trivial task. One side of the problem is to find a fundamentally correct way to include the condensed-phase environment in the model, while the other side is a purely computational aspect of the problem. Related to the later problem is the need for large computational resources, development of new efficient algorithm for data analysis as well as the effective use of computational resources (which is again, a programming-related problem). In the present study, we address a particularly relevant problem computation of magnetic properties of aqueous Al^{3+} ion. We consider both the fundamental aspects concerning the development of computational method and the computational aspects related to the efficiency of the computational process. In particular, we address the application of the map-reduce computation technique to the particular phases of computation within the developed methodology.

Aqueous Al^{3+} is a particularly relevant system to many scientific areas. These include geochemistry, biochemical sciences as well as environmental chemistry. Numerous aspects related to hydration of Al^{3+} in dilute and concentrated water solutions have been studied. However, most of the methods that aim to explicitly include the condensed-phase influence on the ion's properties were based either on cluster or cluster + polarizable continuum approaches. In the first (cluster) approach, actually only the nearest-neighbor in-liquid environment is explicitly included in the model. In other words, potential energy hypersurfaces of the free clusters including the Al^{3+} ion plus nearest-neighbor water molecules are investigated and the properties of such clusters are computed, aiming to reproduce the in-liquid behaviour of the system in question. Of course, accounting solely for only those solvent molecules that reside in the nearest neighborhood of the Al^{3+} ion can hardly lead to satisfactory results if one is interested of the complete solvent influence on these properties. It is well known that a strongly dipolar liquid (such as water) exerts a substantial at least electrostatic influence on highly charged ion such as Al^{3+}, manifested both through ion-multipole (dipole, quadrupole etc.) and ion-induced multipole interactions. Placing the ion plus the nearest neighborhood environment within the rest of the solvent treated as a polarizable continuum is of course a much more realistic model variant. However, even with this approach, much of the bulk solvent's fundamental characteristics are left out from the model. For example, treating the bulk solvent as a simple polarizable continuum disregards the specific noncovalent intermolecular interactions between water molecules residing within the first hydration shell (i.e. the actual nearest neighbors to the ion) and those residing within the second shell. Such specific interactions, which are of hydrogen bonding type in this particular case, could lead to rather significant alterations in the electronic density of the second-shell water molecules, much larger than those predicted by the simpler cluster + PCM approach.

In this paper, we aim to propose a model which could enable a much more realistic description of this aqueous ionic system, on the basis of a hybrid statistical physics quantum mechanical approach. The main idea of the approach is to account explicitly for the dynamical character of the in-liquid environment, i.e.

not to disregard the thermal motion of the molecules within the liquid phase. After the statistical physics model of the liquid is generated, we then implement more rigorous quantum mechanical methodology for computation of Al^{3+} ion magnetic properties accounting for the liquid environment at different levels of sophistication and approximation. As mentioned before, we also address significant computation-related issues, which are crucial for efficient usage of the high-performance computing architectures for these purposes. We actually focus on the development of a method for computation of in-liquid magnetic properties of aqueous Al^{3+}, or, more precisely, of the $Al(H_2O)_6{}^{3+}$ species in a liquid environment - $Al(H_2O)_6{}^{3+}(aq)$. These hydrated species have served as a standard for comparison of magnetic properties of other possible aqueous Al(III) species, i.e. as a sort of an internal standard with respect to which e.g. the average ^{27}Al isotropic shielding constants are referred to, i.e. with respect to which the shifts are computed. It is therefore of certain both fundamental as well as practical importance to develop a robust and reliable method for computation of $Al(H_2O)_6{}^{3+}(aq)$ magnetic properties, with proper inclusion of the in-liquid aqueous environment. Of course, one possibility is the finite-cluster approach, where finite clusters of increasing size would be treated either in a static or dynamic manner, as explained in more details before. On the other hand, a much more plausible approach would be to choose a single sort of averaged configuration generated from a series of statistical physics simulations, for which a rather high-level computation could be performed.

2 Related Work

Map-Reduce model and Hadoop have been used for solving different scientific problems such as data analysis, simulations, data mining, sorting, etc. The Map-Reduce implementation in Google inc. is presented in [6]. The authors confirm that the implementation is highly scalable because it supports the execution of upwards of one thousand Map-Reduce jobs and it processes terabytes of data on thousands of machines. They also use Map-Reduce model for data mining, machine learning, data generation, etc. The scalability and speedup that can be achieved with the use of the Map-Reduce model are discussed in [9]. The authors perform High Energy Physics data analysis and Kmeans clustering using the Map-Reduce model. Also they developed a streaming-based Map-Reduce implementation and they compared its performance with Hadoop. In their paper [22], the authors propose an improved MK-means algorithm for large-scale meteorological data based on Map-Reduce model. In [12], the author proposed a new solution for molecular dynamics simulation based on the Map-Reduce technique and Hadoop. The purpose of this simulation is to predict the execution time of a given molecular dynamics simulation system. In [25], the authors describe a Hadoop based cloud scientific computing application that processes sequences of microscope images of lilivg cells. A structured programming framework (Genome Analysis Toolkit (GATK)) designed to ease the development of efficient and robust analysis tools for next-generation DNA sequencers using the Map-Reduce

model is presented in [18]. A Hadoop plugin for execution of the logical queries over array-based data models is given in [3]. The main goal is to reduce the data transfers and unnecessary reads.

3 Implementation of the ASEC Computational Method Using the Map-Reduce Technique

3.1 Map-Reduce Model and Hadoop

MapReduce is a programming model for processing large data sets in parallel [17]. The map reduce programs consists of finite sequence of rounds, each containing three phases [23]: **Map phase** - maps each single key-value pair to the machines in the run-time system as a new multiset of key-value pairs where each value is a substring of the original value; **Shuffle and Sort phase** - sorts and transfers the map output to the reducers; **Reduce phase** - computers some function on the data on each machine (merges all the intermediate values associated with the same key).

Map-Reduce programs implement the Mapper and Reducer interfaces to provide the map and reduce functions as specified below. Thereby, values with the same key are reduced together [16].
method **Map**(key k, value v) $\rightarrow$ EMIT(key k', value v')
method **Reduce**(key k, value v) $\rightarrow$ EMIT (key k', value $[v', v_2, v_3...]$

Formally, each round of a Map-Reduce program is a finite sequence of 2-tuples (M_i, R_i). Each tuple is consisted of map and reduce functions and it can be written as $((M_1, R_1), (M_2, R_2), ..., (M_n, R_n))$ where M_i is a mapper, R_i is a reducer, $1 \leq i \leq n$ and n is an integer number. Let the Map-Reduce program input which is a multiset of (key;value) pairs be denoted by U_0 and the output which is a multiset of (key;value) pairs of the i-th round be denoted by U_i. The Map-Reduce program executes for $r = 1, ...n$. The **Map**, **Shuffle and Sort** and **Reduce** phases are performed in each iteration (for every r). The **Map** phase feeds each (key;value) pair $(k; v)$ in U_{r-1} to the mapper M_r and runs it. The output of the mapper M_r is a sequence of (key;value) pairs $(k_1; v_1)$, $(k_2; v_2)$,... and it is defined as: $U'_r = \cup_{(k;v)\in U_{r-1}} M_r((k; v))$. The **Shuffle and Sort** phase constructs $V_{k,r}$ (values such that $(k; v_i) \in U'_r$) from U'_r for each k. The **Reduce** phase feeds the k and some arbitrary permutation of $V_{k;r}$ to the separate instance of the reducer R_r and runs it for each k. The output of the reducer is a sequence of 2-tuples $(k; v'_1)$, $(k; v'_2)$,... and $U_r = \cup_k R_r((k; V_{k,r}))$ which is a multiset of (key;value) pairs produced by the reducer R_r [19] [13].

The Map-Reduce model is suitable when the computational problem can be split into smaller independent computations and the intermediate results should be merged later in order to get the final result. The model automatically supports parallel programming and the programmer is only responsible for writing the map and reduce functions. The Map-Reduce model is also suitable for processing scientific data volumes and clustering algorithms used in chemistry, biology, physics which are computing intensive operations and the use of parallelization techniques is key for achieving efficient data analyzes. This model

provides robustness, simplicity and has less synchronization constraints which supersede the additional overheads. The disadvantage of the model is that the programmer cannot affect the efficiency of the parallelism [13] [9].

Apache Hadoop is an open source software data-processing library used for distributed and parallel processing of large data sets. It is used by many companies including Facebook, Microsoft, Cloudera, Amazon, Yahoo, etc. This project includes four different modules: Hadoop Common (utilities that support the other Hadoop modules), Hadoop Distributed File System (distributed file system that provides high-throughput access to data), Hadoop YARN (framework for job scheduling and cluster resource management) and Hadoop MapReduce (A YARN-based system for parallel processing of large data sets) [10]. The processing of the data in Hadoop can be done in two ways: by using the Map-Reduce directly or by using high-level languages and translating into Map-reduce jobs later [21].

3.2 The Algorithm

The purpose of our algorithm is to implement the averaged solvent electrostatic configuration (ASEC) computational method using the map-reduce technique which is much simpler and computationally less demanding than the more widely used averaged solvent electrostatic potential (ASEP) methodology. There are total 3000 configurations given in the input file. Each configuration is defined by 1 atom of Al (described by x, y and z coordinates) and 3000 molecules of H_2O (each H atom and O atom described by x, y and z coordinates). The goal is to find the averaged configuration consisting of 1 atom of Al and 3000 molecules of H_2O. In order to combine the appropriate atoms (for example first oxygen atom from first configuration with first oxygen atom from the all other configurations), we enumerate all atoms, thereby adding indexes to them.

We define two additional classes: **Composite_key** and **Atom**. The class **Composite_key** implements the interface **WritableComparable** and overrides the three methods: **readFields(DataInput in)**, **write(DataOutput out)** and **compareTo(compositekey o)** [11]. There are two instance variables defined in the **Composite_key** class: index and atom_name. Each key of the $< key, value >$ tuple is an object of the class **Composite_key** and each value is an object of the class **Atom**. The class **Atom** implements the interface **Writable** and it has three instance variables of the type double: x, y and z. Any key or value type in the Hadoop Map-Reduce framework implements the interface **Writable** (or the interface **WritableComparable**). The pseudo code of the map and reduce methods used in our algorithm is given below.

```
Method Map(LongWritable key, Text value):
 // key: input key  value: input_value
 Composite_key k;
 Atom v;
 for each line in value:
 {
```

```
    String [] array=line.split(" ");
    k=new Composite_key(array[1], array[2]);
    v=new Atom (array[3], array[4], array[5]);// x, y and z
    EmitIntermidiate(k,v)
 };

Method Reduce(Composite_key k, Iterator<Atom> interm_vals):
 // k: key  interm_vals: intermediate values -
 // list of all values(Atoms) grouped by k
 double sumx=0.0, sumy=0.0, sumz=0.0;
 Atom result;
 for each v in interm_vals:
 {
    sumx += v.x;
    sumy += v.y;
    sumz += v.z;
 }
 result.x=sumx/length(interm_vals);
 result.y=sumy/length(interm_vals);
 result.z=sumz/length(interm_vals);
 Emit(k, result);
```

4 Statistical Physics Simulations

To generate the structure of the Al(III) ion aqueous solution, statistical physics simulations were carried out by the Monte Carlo (MC) method, applying the Metropolis algorithm. For this purpose, the statistical mechanics code DICE was used [4]. All MC simulations were performed in the isothermal-isobaric (NPT) ensemble, at T = 298 K, P = 1 atm, using the experimental density of liquid water of 0.9966 g cm^{-3} at these conditions. In each MC simulation, a single Al^{3+} ion was surrounded by 3000 water molecules in a cubic box with side length of approximately 45 $\mathring{A}$, imposing periodic boundary conditions. Long-range corrections (LRC) to the interaction energy were calculated for interacting atomic pairs between which the distance is larger than the cutoff radius defined as half of the unit cell length. The Lennard-Jones contribution to the interaction energy beyond this distance was estimated assuming uniform density distribution in the liquid (i.e. g(r)$\approx$ 1), while the electrostatic contribution was estimated by the reaction field method involving the dipolar interactions. In all MC simulations carried out in the present study, intermolecular interactions were described by a sum of Lennard-Jones 12-6 site-site interaction energies plus Coulomb terms:

$$U_{ab} = \sum_{i}^{a} \sum_{j}^{b} 4\varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \quad (1)$$

where i and j are sites in interacting molecular systems a and b and r_{ij} is the interatomic distance between sites i and j. The following combination rules were

used to generate two-site Lennard-Jones parameters ε_{ij} and σ_{ij} from the single-site ones:

$$\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j} \tag{2}$$

$$\sigma_{ij} = \sqrt{\sigma_i \sigma_j} \tag{3}$$

For water, we have used the SPC model potential parameters [2].

5 Quantum Mechanical Computations of Magnetic Properties of the Aqueous Al(III) Species

Nuclear magnetic shielding tensors are defined as mixed second derivative of the energy (E) with respect to the magnetic moment of the X-th nucleus ($\overrightarrow{m}_x$) and the external magnetic field ($\overrightarrow{B}$) [8]:

$$\sigma_x^{\alpha\beta} = \frac{\partial^2 E}{\partial B^\alpha \partial m_x^\beta}$$

where the Greek superscripts denote the corresponding vector or tensor components. We have computed the ^{27}Al isotropic shielding values as:

$$\sigma_{\text{iso}} = \frac{1}{3}(\sigma_{11} + \sigma_{22} + \sigma_{33})$$

To achieve gauge invariance, two approaches were used: the GIAO (gauge independent atomic orbital) and the CSGT (continuous set of gauge transformations) method. GIAO methodology is based on using explicitly field-dependent wavefunctions [7,24,20]. The CSGT method, on the other hand, is based on the expression for the shielding tensor components for the X-th nucleus in terms of the induced first-order electronic current density; accurate calculations of the last quantity are achieved by performing a gauge transformation for each point in space [14,15,1]. All quantum mechanical calculations were carried out with the B3-LYP combination of exchange and correlation functional, as well as the second order Moller-Plesset perturbation theory (MP2), using the rather large 6-311++G(3*df*,3*pd*) basis set for orbital expansion.

6 Influence of the Aqueous Environment

The simplest purely electrostatic model of the solvent influence on the Al^{3+} ion magnetic properties consists of treating all of the solvent molecules from the in-liquid environment as being built up by point charges. There are several possible ways to generate such solvent representation. One alternative is to compute the Al^{3+} ion magnetic properties in various "momentary" (or through-phase-space) in-liquid environments, taken as snapshots from the MC simulations, in which all of the water molecules residing within a sphere with a radius e.g. equal to the half of the unit-cell side length (or even larger, accounting for the periodic boundary

conditions) are represented by point charges placed at the hydrogen and oxygen atomic positions as generated by the statistical physics simulation. Another alternative, which we actually implement and use in the present study, is the averaged solvent electrostatic configuration (ASEC) developed by Coutinho and collaborators [5]. It consists of superimposing the solvent atomic charges taken from M statistically uncorrelated MC-generated configurations, each scaled by 1/M. The single configuration that is used thus actually consists of a single Al^{3+} ion, surrounded by 100·x water molecules, where x is the number of water molecules included in the configuration, chosen such that the O_{water}...Al distance is smaller than a threshold value of r_{tresh}.. The maximum value of r_{tresh}. is $a/2$, where a is the MC cell side length. The actual charges of oxygen and hydrogen atoms within the solvent water molecules were equal to $q_O/100$ and $q_H/100$, where q_O and q_H are the corresponding SPC model charges which have been actually used throughout the MC simulations. In this context, we test the convergence of the computed ^{27}Al isotropic shielding constant with the threshold distance value and the computational methods used for achievement of gauge invariance. Table 1 summarizes the obtained results. As can be seen, with such large basis set as used in our computations, both GIAO and CSGT isotropic shielding values may be safely considered as being converged. As for the convergence with the r_{tresh} value, it can be seen that the results are already converged upon inclusion of all water molecules the oxygen atoms of which reside within a sphere of a radius equal to 5 $\mathring{A}$ around the central Al(III) ion. Since all of the discussed values are excellently converged, the dispersion of the distribution of the corresponding values obtained from a series of computations for 100 different environments of the Al^{3+} cation are rather small, the average values being in excellent agreement with those computed with the ASEC approach. In the currently studied case, thus, the ASEC approach is an excellent approximation to obtain the average isotropic shielding values of the aqueous Al(III) ion, relying on a quantum mechanical calculation of only a single in-liquid configuration (though such configuration is surely unphysical one).

Table 1. ^{37}Al isotropic shielding values (expressed in ppm) computed at B3LYP/6-311++G(3*df*, 3*pd*) level of theory, with GIAO and CSGT approaches to achieve gauge invariance, at various r_{tresh} values (expressed in $\mathring{A}$, see text for details)

	Free ion	5 $\mathring{A}$	10 $\mathring{A}$	15 $\mathring{A}$	20 $\mathring{A}$
GIAO	766.2468	766.2421	766.2421	766.2421	766.2421
CSGT	766.2477	766.2430	766.2430	766.2430	766.2430

However, considering all of the solvent water molecules as point charges in the context of computation of Al(III) magnetic properties in the liquid medium is a rather crude approximation, especially if one wants to obtain quantitative agreement with the experimental data for certain other Al-containing complex aqueous species. For such species, often the isotropic shielding values are expressed as shifts with respect to the aqueous Al^{3+} species, the later one thus

serving as a sort of "internal standard". This statement especially holds in the case of those water molecules residing in the closest vicinity of the Al^{3+} ion, i.e. within the first hydration shell. Interaction of these particular solvent molecules with the central ion is much different and much more complex than the pure Coulombic interaction assumed in the simple ASEC approach. While ASEC approach is surely good enough to approximate the residual influence of the remaining bulk solvent molecules on the overall isotropic shielding (and other magnetic properties) of the central Al^{3+} ion, at least the first-shell waters need to be treated in a different way, i.e. describing them by a full quantum mechanical wavefunction (instead of point charges). To demonstrate the validity of the previous statement, we have carried out test calculations at B3LYP/6-311++G(3*df*,3*pd*) CSGT and GIAO levels of theory for the $Al(H_2O)_6{}^{3+}$ species embedded in "bulk solvent" molecules treated as point charges, taken from several snapshots from the equilibrated MC run. One typical configuration of this type is shown in Fig. 1.

Fig. 1. A typical configuration taken from an equilibriated MC run, showing a particular arrangement of the first-shell water molecules around the central Al^{3+} ion

7 Conclusion and Future Work

Such test calculations have shown that the computed isotropic shielding values for the central Al(III) species differ significantly from those obtained by the simpler ASEC approach. Calculations for several thousands of such configurations are currently in progress, in order to obtain the exact distribution of isotropic shielding values. At the same time, a development of the procedure for averaging the configuration of the first-shell waters from several thousands of statistically independent configurations from MC run is in progress as well. While the configuration of the remaining part of the solvent molecules, representing the "bulk" solvent may be safely treated by the elaborated ASEC approach, the choice of average "first salvation shell" is a rather specific and non-unique computational task. One approach is to simply average the Cartesian coordinates of each of the first-shell water molecules for a series of MC-generated configurations (after appropriate coordinate system transformation if necessary). Such approach seems to be justified as the solvent molecules residing in the first hydration shell around the Al^{3+} ion are rather tightly bound, i.e. they practically do not interchange with the second-shell waters throughout the simulation. This can be clearly seen

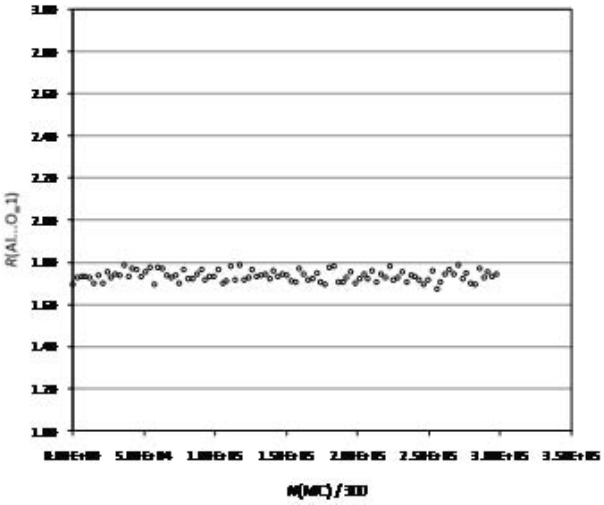

Fig. 2. The Al...O_w distance variation for one of the first-shell waters around Al^{3+} ion as a function of the MC step

in Fig. 2, where the Al...O_w distance variation for one of the first-shell waters around Al^{3+} ion is shown as function of the MC step. However, such approach has certain ambiguities, that need to be solved in an unique way. A useful alternative would be to average the orientational parameters of each of the first shell waters around Al(III), upon suitable coordinate transformation for each MC-generated configuration. Development of such approach is in progress.

References

1. Bader, R.F.W., Keith, T.A.: Properties of atoms in molecules: Magnetic susceptibilities. The Journal of Chemical Physics 99(5) (1993)
2. Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., Hermans, J.: Intermolecular forces. In: Pullman, B. (ed.). Reidel, Dordrecht (1981)
3. Buck, J.B., Watkins, N., LeFevre, J., Ioannidou, K., Maltzahn, C., Polyzotis, N., Brandt, S.: Scihadoop: Array-based query processing in hadoop. In: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2011, pp. 66:1–66:11. ACM, New York (2011)
4. Coutinho, K., Canuto, S.: Dice: A monte carlo program for molecular liquid simulation. University of São Paulo, Brazil (1997)
5. Coutinho, K., Georg, H., Fonseca, T., Ludwig, V., Canuto, S.: An efficient statistically converged average configuration for solvent effects. Chemical Physics Letters 437(13), 148–152 (2007)
6. Dean, J., Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. Commun. ACM 51(1), 107–113 (2008)
7. Ditchfield, R.: Self-consistent perturbation theory of diamagnetism. Molecular Physics 27(4), 789–807 (1974)

8. Dykstra, C.: Quantum chemistry and molecular spectroscopy. Prentice Hall PTR (1992)
9. Ekanayake, J., Pallickara, S., Fox, G.: Mapreduce for data intensive scientific analyses. In: Proceedings of the 2008 Fourth IEEE International Conference on eScience, ESCIENCE 2008, pp. 277–284. IEEE Computer Society, Washington, DC (2008)
10. Foundation, T.A.S.: Apache hadoop, `http://hadoop.apache.org/`
11. Foundation, T.A.S.: Apache hadoop, `http://hadoop.apache.org/docs/r2.3.0/api/org/apache/hadoop/io/WritableComparable.html`
12. He, C.: Molecular Dynamics Simulation Based on Hadoop Mapreduce. Ph.D. thesis, Computer Science and Engineering, Department of University of Nebraska-Lincoln, Lincoln, Nebraska (May 2011)
13. Karloff, H., Suri, S., Vassilvitskii, S.: A model of computation for mapreduce. In: Proceedings of the Twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, pp. 938–948. Society for Industrial and Applied Mathematics, Philadelphia (2010)
14. Keith, T., Bader, R.: Calculation of magnetic response properties using atoms in molecules. Chemical Physics Letters 194(12), 1–8 (1992)
15. Keith, T.A., Bader, R.F.: Calculation of magnetic response properties using a continuous set of gauge transformations. Chemical Physics Letters 210(13), 223–231 (1993)
16. Licari, D.: Mapreduce (November 2010)
17. Lmmel, R.: Google's mapreduce programming model revisited. Science of Computer Programming 70(1), 1–30 (2008)
18. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A.: The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research (9), 1297–1303 (2010)
19. Spangler, T.: Algorithms for Grid Graphs in the MapReduce Model. Master's thesis, University of Nebraska-Lincoln (2013)
20. Tossell, J.: Nuclear magnetic shieldings and molecular structure. NATO ASI series: Mathematical and Physical Sciences. Kluwer Academic Publishers (1993)
21. Wang, G.: Evaluating MapReduce System Performance: A Simulation Approach. Ph.D. thesis, Faculty of the Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA (August 2012)
22. Wei Fang, V.S., Sheng, X.W., Pan, W.: Meteorological data analysis using mapreduce. The Scientific World Journal 2014, 10 (February 2014)
23. White, T.: Hadoop: The Definitive Guide, 1st edn. O'Reilly Media, Inc. (2009)
24. Wolinski, K., Hinton, J.F., Pulay, P.: Efficient implementation of the gauge-independent atomic orbital method for nmr chemical shift calculations. Journal of the American Chemical Society 112(23), 8251–8260 (1990)
25. Zhang, C., De Sterck, H., Aboulnaga, A., Djambazian, H., Sladek, R.: Case study of scientific data processing on a cloud using hadoop. In: Mewhort, D.J.K., Cann, N.M., Slater, G.W., Naughton, T.J. (eds.) HPCS 2009. LNCS, vol. 5976, pp. 400–415. Springer, Heidelberg (2010)

Opportunities and Challenges for Green HPC

Sonja Filiposka[1,2], Anastas Mishev[1], and Carlos Juiz[2]

[1] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, R. Macedonia
{sonja.filiposka,anastas.mishev}@finki.ukim.mk
[2] Architecture and Performance of Computer and Communication Systems Department, University of the Balearic Islands, Palma de Mallorca, Spain
cjuiz@uib.es

Abstract. Recognizing the unsustainability in designing power hungry HPC systems, in the recent years an effort towards an energy and performance efficient HPC design is on the rise. Based on the available data on top green HPC systems, in this paper we analyze the main components of the HPC system with the attempt to provide an insight into the green efficiency in line with the technology development. Using global scale measurements, our goal is to discover the most promising designs and pinpoint on the existing main challenges that should direct future research efforts.

Keywords: energy efficiency, green computing, HPC, power consumption, performances.

1 Introduction

Today's top high-performance computing (HPC), i.e. supercomputers, are being designed with the aim to achieve the highest possible performance in terms of the number of 64-bit floating-point operations per second (flops). Their architecture has evolved from early custom design systems to the current typical clusters made out of multisocket, multicore systems. In terms of performance these systems are ranked twice a year and the results are published as the Top-500[1] list using the High Performance Linpack benchmark (HPL)[2] as the referent benchmark for the ranking.

A historical analysis of these published data can put light to some interesting developments in the HPC architecture as well as shed light on the possibilities of the future trends and main research challenges. Thus, one noticeable example was easily highlighted in the 2009 list: the advent of HPC clusters based on accelerator cores (co-processors) as the dominant petascale architecture. In order to achieve more operations per second current architectures seem to be moving away from the traditional clusters of homogenous nodes to clusters of heterogeneous nodes. This trend that

[1] http://top500.org, current list: November 2013.
[2] http://www.netlib.org/benchmark/hpl

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_5

started a few years back has led to the last ranking set with the first two most powerful supercomputers to be a heterogeneous cluster with Xeon Phi co-processors and a MPP with NVidia GPUs. However, on 4 out of the top 10 on the list, the homogenous PowerPC technology still reins with its BlueGenes. Adding a GPU to a conventional HPC cluster node can quadruple its peak performance, or even increase it by an order of magnitude when using 32-bit arithmetic [1]. But the increased peak performance doesn't necessarily vouch for sustained application performance.

While in the past the only metric that was valued were flops, over the years the HPC community has acknowledged that this kind of design leads to supercomputers that consume a lot of power. Thus, the currently top power hungry supercomputer consumes over 19 MW. Hence, a shift in the goals and design has been made towards the so-called green HPC systems whose goal is high performance with small power consumption. This initiative has been supported with the rising popularity of a green ranking metric: the performance versus power ratio, published as the Green-500 list [2]. The benchmark used here is also Linpack, where performance and power consumption are measured are expressed as flops per watt. The data published in this list allows for forecasting the future trends towards HPC architectures that are power consumption friendly while still aiming for the exascale computing power.

Detailed inspection and cross comparison of these systems can unearth a lot information about the power efficient architectures that should be utilized and developed in the future in order to further improve the green HPC initiative. The main goal of this paper is to infer the future of efficiency and power consumption of HPC systems by analyzing the technologies developed today showcased by the leadership-class computer systems. The objective is to provide an insight of the performance versus power trends for the current best architectures and to determine the direction in terms of processors, interconnections, system family and alike that shows a steady state improvement and paves the way for the future power aware efficient HPC systems.

2 Green Potential Analysis of Current Top HPC Systems

Since increased power consumption of the HPC system inevitable leads to high energy spent on cooling equipment as well as increased construction costs and problems with the system reliability and availability, the green HPC initiative is a major concern for researchers and vendors. Thus, today, power management and power effective architecture has become essential for HPC systems. The overall energy efficiency of supercomputers has improved rapidly in the first years of the green HPC initiative. Improvements are observed over the full range of machines, and are due to the "plucking" of low hanging fruit in energy efficiency, e.g., using existing low-power microprocessors [3]. After the initial burst stage, further improvements would require novel energy-centric architectural designs, and hence, would take longer to achieve.

One of the even more pronounced problems with power consumption in HPC systems today is the inefficient use of their energy consumption due to the reduced performance when compared to their theoretical peak. The measured performance

obtained with HPL (Rmax) is drastically different across systems (ranging from 28% to 81% of the projected theoretical peak), while the real performances of the scientific applications are only 10% of Rmax [4]. As it can be seen in Fig. 1 the gap between the max and peak performance is getting more pronounced in the last years while the number of flops is exponentially rising, especially when accelerators are introduced into the system. The timeline analysis given in Fig. 2 clearly shows that although efforts are being made in making the HPC systems more power efficient, this is actually not a case since the power consumption in average and maximum values is increasing even more steeply in the last years compared to the overall green performance.

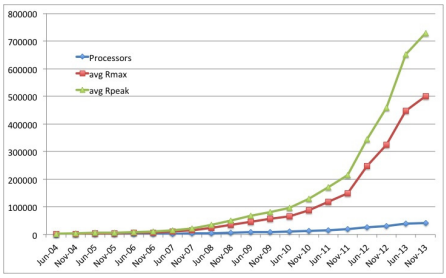

Fig. 1. Timeline of the average performances of top supercomputers

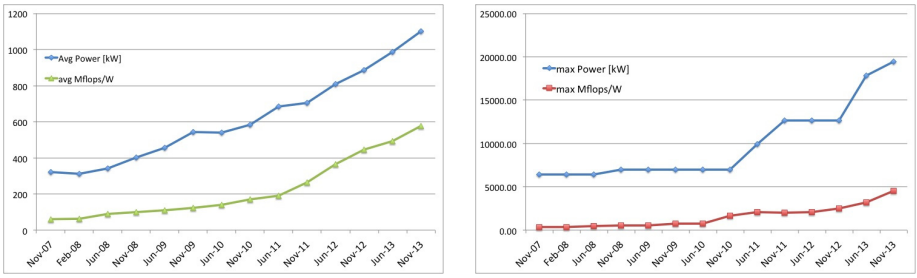

Fig. 2. Timeline of (a) average and (b) maximum power consumption and performances of top supercomputers

However, the analyzed timelines also suggest the existence of notable trends in the HPC evolving architecture. This global performance versus power consumption overview can be used as a starting point for a thorough analysis of the influence of each part of the HPC architecture over the system green performances, which in the end can serve as a guideline for deciding on the future HPC architectures as well as a topic for further improvement in the research field. Thus, in the following subsections we investigate the influence of each significant part of the HPC system design options from the performance and power consumption point of view based on the available detailed descriptions of the top supercomputers.

2.1 Processor Family

One of the first decisions made when designing an HPC system is the processor family and generation. Today's top HPC systems are mostly based on the Intel SandyBridge (61%) with the Intel Xeon 5 processor generation, followed by the older Intel Nehalem (13%) and the newest Intel IvyBridge (7%), which is a minor improvement: almost all of the SandyBridge implementations come with 8 cores per processor socket, while IvyBridge increases to 10 and 12. The AMD processor family, on the other hand, is based on Opterons that with 12 or 16 cores per socket, and the PowerPC family with PowerBQC processor utilizes exclusively 16 cores per socket.

When compared on the basis of performance/power, as given in Fig. 3, it can be seen that the PowerPC processor family offers the best green performances and is outperformed only by Intel SandyBridge and IvyBridge in the cases when a huge number of co-processor cores are used. These results confirm that the design of IBM's PowerPC is made with two goals in mind: low energy consumption and high performance. It can also be easily seen that the leap from the previous Power to the new PowerPC family is tremendous. One of the major changes towards this goal is the reduced processor frequency from 3.8 to 1.6 GHz. These simple, power-efficient processors originally developed for embedded systems are the basis of the top performing IBM BlueGene/Q. They also include several task specific acceleration engines.

Although the PowerPC is the most powerful and green friendly processor architecture so far, its major drawback is the custom interconnection and design that creates a software and hardware "isolated" effect many are trying to get away from. Thus, the most general choice is the Intel based alternative with its less than half typical performances. With a close inspection of the Intel SandyBridge, one can perceive four different levels of performances. While the last performance peak is due to the introduction of co-processors as was already mentioned, the first three different levels are mainly due to the different interconnection types used (namely Gigabit Ethernet, 10G Ethernet and InfiniBand), whose influence is discussed later on.

An important observation are the performances of one Xeon processor generation with a different number of cores per socket, or different frequency. A comparative analysis has shown that as the number of cores per socket rises the performance/power ratio for the system is expected to rise as well, while as the frequency rises the performance/power ratio falls. These two trends will pave the way for future processor generations that are expected to decrease their frequency and increase the number of cores per socket in order to achieve better performances for lower energy consumption. This is already somewhat the case with the Intel IvyBridge family with an increased number of cores per socket and lowered frequency. Thus, almost half of the representatives from this family have a performance/power ratio above 1000.

As for the HPCs based on the AMD processor family, they are usually designed using the MPP architecture, unlike the typical Intel Cluster. Increasing the number of cores per socket increases their performance/power ratio as well. In order to reach a higher performance/power ratio (above 1300) they need to be combined with co-processors. They are also usually seen in combination with a Cray interconnect which provides better performances than other types of interconnect for this family.

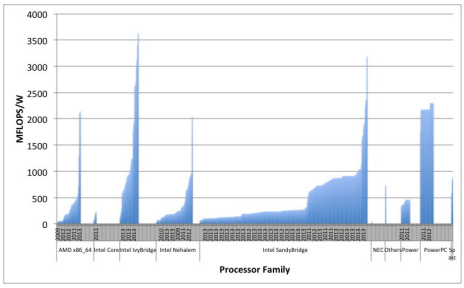

Fig. 3. Performance/power categorization of different HPC processor families

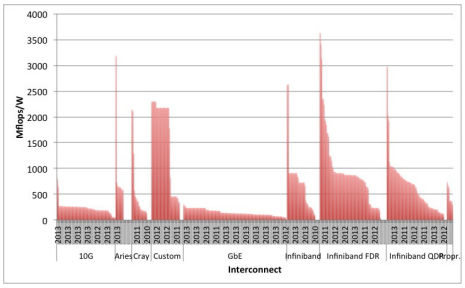

Fig. 4. Interconnect type influence on the supercomputer performances

2.2 Interconnect Type

As already noted, another important characteristic of a HPC system is the type of interconnection used. Today's top performing systems are mainly based on the InfiniBand technology (41%), followed by Gigabit Ethernet (27%) and 10G Ethernet (16%). The rest of the interconnect types are mainly custom based for the specific systems like Cray interconnect, or the custom interconnect in the IBM BlueGenes.

From the green perspective, in Fig. 4 the performance is presented for different interconnect type families. The results show that IBM's Custom interconnect is the most performance/power efficient solution. However, this type of interconnect is used only in combination with a Power-family processor, which is its major compatibility drawback. Also, due to its torus-like topology, this interconnect requires more programming effort to be utilized to its maximum.

The second best performing interconnect type is the InfiniBand with its different implementations. InfiniBand has gained wide acceptance in HPC mainly due to its high bandwidth and in particular due to its low latency and high flexibility. The InfiniBand technology is seen as the successor of the common Gigabit and 10G Ethernet [5] and, as it can be seen in the figure, highly outperforms its predecessors in terms of green performance. However, it is always found in combination with the Intel Xeon processor family, while some of the Ethernet based solutions are designed using AMD Opteron. InfiniBand seems to be significantly better than 10G or Gigabit Ethernet from the energy efficiency perspective. For instance, in the case of systems based on the Xeon E5-2680 8C 2.700GHz processor, InfiniBand gives in average 3.5 times better performance/power ratio compared to 10G, and is 2.7 times better compared to Gigabit Ethernet. However, this is not always the case. Among the top HPC systems there are a number of examples that show that InfiniBand does not always work well with NVidia co-processors. For an example, when comparing two systems that differ only in that the first one is designed with and the second without co-processors, it turns out that the performance/power ratio drops rapidly (around 3.5 times) when co-processors are introduced mainly due to the lower performances of the systems that drop significantly below the peak. This example must raise a flag of

careful inspection of the system since, despite expectations, adding co-processors into the system will not always boost the performances and green behavior. We must note however, that there is an example (namely, the CSIRO GPU Cluster), which is a successful example of mixing InfiniBand and NVidia co-processors.

Thus, in the given figure, the best and worst performing InfiniBand based examples are the ones with co-processors. The performance interconnection between InfiniBand and GPUs is just becoming a hot topic in the research community [6]. It is important to note that the Ethernet based interconnections never exhibited this problem when co-processors are introduced into the system. On the contrary, examples show that systems based on 10G Ethernet interconnection perform as well as InfiniBand based solutions in the cases when the 10G based system is built using a great number of cores. This is another important remark regarding InfiniBand, namely the InfiniBand based systems are usually built using a smaller number of cores with rare examples of systems with a great number of cores, which is mainly due to the complexities of its flat fabric. Also, with InfiniBand Remote DMA the cores are free from overseeing the network data read/write, which boosts the system performances without the need to add more cores.

2.3 Accelerators / Co-processors

Adding accelerators, or co-processors, is the current trend for achieving green HPC performances that started in November 2009 when a heterogeneous HPC system has appeared for the first time in the top500. In the next year the trend has gotten momentum and a fast rise of this architecture has been foreseen [7], as presented in Fig. 5. Thus, it is somewhat unexpected to see a small decrease in the number of supercomputers based on this architecture in the last year. The architectures and programming models of co-processors may differ from CPUs and vary among different co-processor types. This heterogeneity leads to challenging problems in implementing and porting of application operations and obtaining best performance. Currently the heterogeneous HPC systems constitute 10% of the top 500, the worst one of which has maximum performance that is only 28% of its theoretical peak.

The current average efficiency in terms of sustained versus peak performance of the top supercomputers that are constructed using NVIDA GPUs is around 0.57, which is still a lot behind the average of 0.7 of the ones without any accelerators. However, the top supercomputers that utilize Xeon Phi accelerators seem to be catching up with an average efficiency of 0.69, although they are a lot less popular in the community. One must always bear in mind that these differences are a lot more pronounced when considering real workloads. Still, we must not forget that out of the total of 52 such systems, 11 are in the current highest top according to the green500 list followed by the IBM's BlueGene/Q architecture as their opposite side in possible future direction.

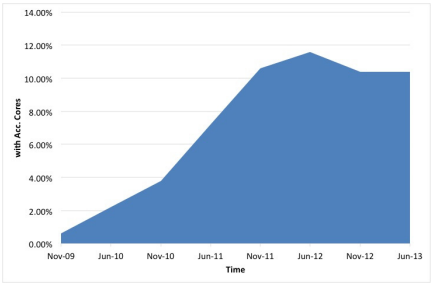

Fig. 5. Timeline of the amount of accelerator cores used for boosting performances

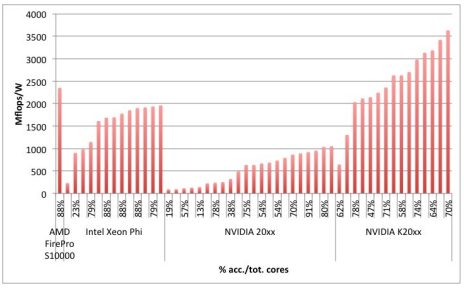

Fig. 6. Green performances of different heterogeneous HPC systems

It is of practical interest to analyze how the amount of share of co-processors in the number of total cores impacts the performance/power ratio of the system, as well as how are the main competitive accelerators performing. This analysis is presented in Fig. 6. It is clear that the accelerator share in almost all of the systems is well above 50%, with a typical 70% for NVidia and 88% for Xeon Phi. As it is presented, with careful systems design the expected achieved green performance can be above 1000 Mflops/W, with potential to reach staggering 3500 Mflops/W and more.

There seem to be two actual choices for a co-processor in the today's systems: the Intel many integrated core (MIC) Xeon Phi and NVidia's most popular K20xx options. It is evident that NVidia's new Kepler architecture improves the GPU's performance significantly, mainly due to the new streaming multiprocessor SMX [8]. However, experimental cross comparison [8] shows that different types of co-processors are more appropriate for specific data access patterns and types of parallelism. The MIC's performance compares well with that of the GPU when regular operations and computation patterns are used. The GPU is more efficient for those operations that perform irregular data access and heavily use atomic operations. The programming tools and languages employed for code development for a MIC are the same as those used for CPUs. This is a significant advantage as compared to GPUs. For the MIC, auto vectorization is performed by the compiler, which however needs additional guidance when complex pointer manipulations are used [9].

Since the expectation is for the GPUs to carry out a substantial portion of the calculations, host memory, PCIe bus, and network interconnect performance characteristics need to be matched with the GPU performance in order to maintain a well-balanced system [10]. InfiniBand QDR interconnect is highly desirable to match the GPU-to-host bandwidth. Host memory also needs to at least match the amount of memory on the GPUs in order to enable their full utilization, and simplification of the development of MPI-based applications is necessary. However, many challenges remain open in order to make the accelerated HPC systems truly energy efficient with practical performances a lot closer to their theoretical peak compared to today. Accelerator technologies are difficult to program and unsuitable for some workloads [11]. More important, however, is the inability of many applications to efficiently map to accelerator architectures. At the high end, suitability for a wide range of applications is a must. These two issues call into question accelerator viability in the largest exascale machines.

2.4 System Family Impact

In Fig. 7 the power consumption related to the number of total cores for the most prominent system families of today's HPC supercomputers is presented. The results show that there are three current trends that depict the green status and scaling of the different system families. The most prominent example are the HP Cluster Platform system families that are all consistently following a linear increase in the total power with the rising number of total cores. This example is also the least performing one since the toll of more power needed for increasing the number of cores (and thus performances) is the highest of all compared. However, there are two members of this group that show low power consumption in combination with a large number of cores (circles on the figure). This "out of normal" behavior is due to the fact that these systems are supported by a great number of co-processors, which on the other hand require a lot less power per core compared to a "pure" core in the system.

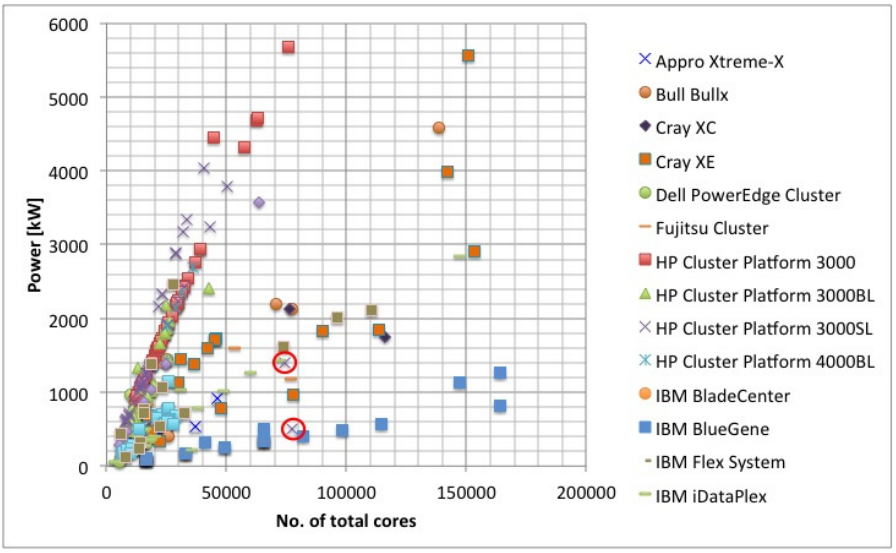

Fig. 7. Power consumption related to number of total cores for different system families

Another obvious trend that strongly relates to the system family are the IBM BlueGenes. The figure clearly shows that IBM BlueGenes scale extremely well with only slight increase of power demand for a great increase of total number of cores, which in further accentuates the excellent properties of this system family since it never relies on increasing its performances by adding accelerators or co-processors. Furthermore, the Mflops/W ratio for these system family is consistently rising over the years with around 370 for the systems using PowerPC 4C processors in 2008, 450 when using Power7 8C in 2011, to a staggering 2200 when using the PowerBQC in 2012-13. It is also of great importance that these systems are scaling with the same Mflops/W ratio when keeping all of the parameters the same and simply increasing only the number of total cores. All of these characteristics make them the

most effective and consistent green HPC system design, with the BlueGenes being on 6 out of the first 10 positions in the green500, keeping high positions in the top500 list as well.

The rest of the system families seem to fall somewhere between the worst and best extremes. Here we find the rest of the supercomputers manufactured by IBM, as well as Cray supercomputers and the SGI ICEs.

To establish the level of impact the difference in system family has over the rest of the system parameters (like processor, interconnection, co-processors) we made a comparison of three different supercomputers that differ in the system family only. This effectively means that the design difference of these systems is in the enclosure, which defines the physical placement of the cores, as well as fans and cooling among other parameters. Our analysis shows that direct liquid-cooling system [12] of the electronic components more than doubles the Mflops/W compared to the other similar configuration. The method of implementation of the internal air-cooling system also strongly influences the efficiency. Systems with shared chassis fans show less efficiency than ones with tightly coupled fans. Thus, the enclosure type has great impact on the overall system performances and has to be chosen very carefully in order to minimize the power consumption while providing maximum system performances. The results also show that thermal aware schedulers are very important for achieving the green goal. Thus major future efforts should be focused on this challenge.

3 Conclusion

Ignoring power consumption as a design constraint results in an HPC system with high operational costs and diminished reliability, which often translates into lost productivity in the long run. Thus, power consumption has become an increasingly important issue in HPC and has focused efforts on design of green systems.

The main focus, however, has been on improving the energy efficiency of computation. A major event in this field has been the introduction of the accelerators in the heterogeneous systems. Beyond being cost-effective, HPC accelerators also have the potential to significantly reduce space, power, and cooling demands. On the other hand they present a number of new challenges in terms of the application development process, job scheduling and resource management, and security.

The analysis of green efficiency of HPC systems presented in this paper has pointed to the conclusion that in order to achieve the best performances for the minimum invested power consumption, the overall designed system must be well balanced. From this point of view, we attempted to point out the connections between the major system components and their influence on the overall green performances of the system. Our results have uncovered the main challenges in future HPC design: low power high performing processors preferably based on embedded systems, interconnect with high bandwidth and low latency that will work well with accelerator cores, increased usability of the raw computational power of the accelerator cores by improving the programmability, and system family that employs liquid based cooling of the individual elements together with thermal balanced schedulers and data access types.

Acknowledgements. This work has been partially supported by the EUROWEB project funded by the Erasmus Mundus Action II programme of the EC, and by the Spanish Ministry of Economy and Competitiveness under grant TIN2011-23889.

References

1. Kindratenko, V., Trancoso, P.: Trends in High-Performance Computing. In: Novel Architectures, Computing in Science and Engineering. IEEE (2011)
2. Feng, W., Cameron, K.W.: The Green500 List: Encouraging Sustainable Supercomputing. IEEE Computer, 50–55 (2007)
3. Feng, W., Lin, H.: The Green500 List: Year Two (2009), `http://www.green500.org`
4. Liu, Y., Zhu, H.: A survey of the research on power management techniques for high-performance systems. Soft. Pract. Exper. 40, 943–964 (2010)
5. Bortolotti, D., Carbone, A., Galli, D., Lax, I., Marconi, U., Peco, G., Perazzini, S., Vagnoni, V.M., Zangoli, M.: Comparison of UDP Transmission Performance Between IP-Over-InfiniBand and 10-Gigabit Ethernet. IEEE Tran. on Nuc. Sci. 58(4), 1606–1612 (2011)
6. Reano, C., Mayo, R., Quintana-Orti, E.S., Silla, F., Duato, J., Pena, A.J.: Influence of InfiniBand FDR on the performance of remote GPU virtualization. In: IEEE International Conference on Cluster Computing (CLUSTER), pp. 1–8 (2013)
7. Krieder, S.J., Raicu, I.: An Overview of Current and Future Computing Accelerator Architectures. In: 1st Greater Chicago Area System Research Workshop Poster Session (2012)
8. Jeong, H., et al.: Performance of Kepler GTX Titan GPUs and Xeon Phi System. Journal of Computational Physics, PoS (LATTICE 2013), 423 (2013)
9. Teodoro, G., Kurc, T., Kong, J., Cooper, L., Saltz, J.: Comparative Performance Analysis of Intel Xeon Phi, GPU, and CPU. In: Distributed, Parallel and Cluster Computing. Cornell University Library (2013)
10. Kindratenko, V.V., Enos, J.J., Shi, G., Showerman, M.T., Arnold, G.W., Stone, J.E., Phillips, J.C., Hwu, W.: GPU Clusters for High-Performance Computing. In: IEEE International Conference on Cluster Computing and Workshops (2009)
11. Hermmert, S.: Green HPC From Nice to Necessity. IEEE Comp. in Sci. and Eng. (2010)
12. Loken, C., et al.: SciNet: Lessons Learned from Building a Power-efficient Top-20 System and Data Centre. J. Phys. Conf. Ser. 256, 012026 (2010)

Exploratory Analysis of Communities in Co-authorship Networks: A Case Study

Miloš Savić[1], Mirjana Ivanović[1], Miloš Radovanović[1], Zoran Ognjanović[2], Aleksandar Pejović[2], and Tatjana Jakšić Krüger[2]

[1] Department of Mathematics and Informatics, Faculty of Sciences
University of Novi Sad, Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia
{svc,mira,radacha}@dmi.uns.ac.rs
[2] Mathematical Institute of the Serbian Academy of Sciences and Arts
Kneza Mihaila 36, 11001 Beograd, Serbia
{zorano,pejovica,tatjana}@mi.sanu.ac.rs

Abstract. Digital libraries enable worldwide access to scientific results, but also provide a valuable source of information that can be used to investigate patterns and trends in scientific collaboration. The Electronic Library of the Mathematical Institute of the Serbian Academy of Sciences and Arts (eLib) digitizes the most prominent mathematical journals printed in Serbia. Using eLib bibliographical records we constructed a co-authorship network representing collaborations between authors who published their papers in eLib journals in the period from 1932 to 2011. In this paper we apply community detection techniques in order to examine the structure of the eLib co-authorship network. Such study reveals characteristic patterns of scientific collaboration in Serbian mathematical journals, and helps us to understand the (self-)organization of the eLib community of authors.

Keywords: digital library, co-authorship network, Serbian mathematical journals, scientific collaboration, social network analysis, community detection.

1 Introduction

It has long been realized that the analysis of co-authorship graphs can help us to understand the structure and evolution of corresponding academic societies. Those networks can also be used to develop models for ranking most influential authors in a database [8], to automatically determine the most appropriate reviewers for a manuscript [21], or even to predict future research collaborations [12]. Nodes in a co-authorship network represent researchers – people who published at least one research paper. Two researchers are connected by an undirected link if they authored at least one paper together, with or without other coauthors. Additionally, link weights can be introduced in order to express the strength of collaboration: two researchers are connected by a link of weight w if they co-authored exactly w different research papers.

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_6

Community structure is a typical feature of social networks [16,4]. A community (cluster or module) is a part of a network (group of nodes) where internal connections are denser than external ones. Uncovering communities helps us to understand the structure of the network, to identify cohesive subgroups, and to draw a readable map of the network.

This study explores structural properties of the co-authorship network that is formed from bibliographical records contained in the Electronic Library of the Mathematical Institute of the Serbian Academy of Sciences and Arts – eLib [13]. ELib started as a response to the increasing requirement for easier access to old issues of the journal *Publications de l'Institut Mathématique*. Currently, eLib digitizes 12 mathematical journals printed in Serbia. Therefore, the nature of the bibliographic data enables us to investigate the structure of scientific collaboration characteristic to authors who publish their results in Serbian mathematical journals.

The rest of the paper is structured as the folows. Related work is presented in Section 2. Section 3 describes the methodology that is used to examine structural properties of the network and identify cohesive subroups of co-authors. The obtained results are presented and discussed in Section 4. Finally, the last section concludes the paper.

2 Related Work

A more recent resurgence of interest in networks of scientists and scientific papers was sparked by the observation of power-law degree distributions in various types of real-world networks [1] including networks of scientific collaboration [14,15]. It is also observed that the largest connected component in collaboration networks tends to take up the majority of the network [14]. Collaboration networks also exhibit expected short paths between arbitrary researchers [15], i.e. they tend to be "small worlds."

The body of work most relevant to our study involves collaboration networks in the field of mathematics. Studies of collaboration networks focused around Paul Erdős include [9] and [2]. More general analysis of mathematics collaboration networks is performed by Grossmann [10,11] who examined statistical properties of the network derived from Mathematical Reviews (MR). Brunson et al. [5] studied the evolution of the MR network, identifying two points of drastic reorganization of the network, as well as increased collaboration between mathematics researchers in more recent times.

Communities in co-authorship graphs may indicate groups of people with common research interest. For example, Girvan and Newman [7] used community detection techniques to identify groups corresponding to different research divisions at the Santa Fe institute. In our previous work [23] we studied statistical properties and evolution of the eLib co-authorship graph. The same article presented the methodology that is used to extract the network. This paper continues the work presented in [23]. Namely, in this paper we investigate the structure of the network using community detection methods.

3 Exploratory Analysis

The analysis of structure of scientific collaboration in eLib journals is based on standard methods and metrics used in analysis of social networks. Firstly, we performed connected component analysis in order to isolate disjoint components of the network and to determine whether the network contains so called *giant connected component*. A connected component of an undirected network is a set of mutually reachable nodes, i.e. there is a path connecting each two nodes in the component. Giant connected component is a component that encompasses the vast majority of nodes. Secondly, we distinguish between two types of components in a co-authorship network: non-trivial and trivial components. A component of a co-authorship network is considered trivial if it is a complete sub-graph of the network and the weight of each link is equal to one. In other words, trivial components represent research collaborations that have not evolved in the examined time period.

We use different metrics to quantify nodes (authors) in the eLib co-authorship network. *Degree centrality* (DC) of author A is the number of links incident to A, i.e. the number of other authors with whom A collaborated. *Betweenness centrality* (BC) of A is the number of shortest paths between any pairs of nodes that pass through A. Unlike DC which is a local centrality measure, BC quantifies the centrality of a node considering the whole network. Nodes with high BC tend to be the most important actors in the network since they connect different groups of nodes and may control the flow of information in the network. To measure author productivity we use the normal counting method, i.e. the productivity of A is equal to the number of publications A (co-)authored. *Timespan* of author A is the number of years that passed from the publication of A's first article to the publication of A's last article in eLib journals.

An important advance in community detection was made by Girvan and Newman [17] who introduced a measure called *modularity* to estimate the quality of a partition of a network into communities. For weighted networks modularity Q is defined as

$$Q = \sum_{c=1}^{n_c} \left[\frac{W_c}{W} - \left(\frac{S_c}{2W} \right)^2 \right],$$

where n_c is the number of communities in the partition, W_c is the sum of weights of intra-community links of community c, S_c is the total weight of links incident to nodes in c, and W is the total weight of links in the network. In other words, modularity accumulates the difference between the total weight of links within a cluster and the expected total weight in an equivalent network with links placed at random. In this paper we use the Louvain method for community detection [3] to identify cohesive subgroups in the eLib co-authorship graph. Initially, we investigated the performance of five different community detection techniques on the largest connected component and showed that the Louvain method is the most suitable for our case study. The method uses a greedy multi-resolution approach to maximize Q starting from the partition where all nodes are put in different communities. When Q is optimized locally the algorithm builds the

coarse-grained description of the network (network of communities), and then repeats the same procedure until a maximum of modularity is attained. Although widely used, the modularity measure has a weakness known as the resolution limit problem – community detection techniques based on modularity maximization may fail to to identify modules smaller than a scale which depends on the total size of the network. Therefore, the application of modularity maximization methods requires investigation of the quality of obtained community partitions. In order to assess the reliability of the community detection method we use the definition of community proposed by Radicchi et al. [19] adopted for weighted networks. Namely, a community is called *Radicchi strong* if for each node in the community the sum of weights of links within the community (strength of intra-community links) is higher than the sum of weights of links connecting the node with the rest of graph (strength of inter-community links).

4 Results and Discussion

In total 6480 research papers were published in eLib journals from 1932 to 2011. The majority of articles are single-authored papers: 4836 papers (74.63% of the total number of papers) are written by exactly one author. This situation is not surprising for mathematical journals, since researchers in mathematics and humanities usually engage in solitary work, while laboratory scientists tend to write articles with many co-authors.

The total number of authors that published papers in eLib journals during the examined period is 3597. Therefore, the co-authorship network formed from eLib bibliographic records contains 3597 nodes (authors). Those authors are connected by a significantly smaller number of links (2766) which means that there is a large number of authors (33% of the total number of authors) who have not collaborated with other eLib authors by publishing articles in eLib journals.

4.1 Connected Components

Connected component analysis revealed that the eLib co-authorship network is extremely fragmented: it contains 625 connected components (excluding isolated nodes) neither of which is a giant connected component. Additionally, the network contains nearly the same number of trivial and non-trivial components: 319 components are trivial (51.04%), while 306 of them are non-trivial. The average size of non-trivial components is 6.42, while the standard deviation is 17.83. This means that the eLib co-authorship graph contains components whose size is drastically larger than the average. In total, 19 components have size that is greater or equal to ten, while six of them have size greater than 20 authors. The largest connected component encompasses 249 authors, which is 6% of the total number of authors. The number of papers published by authors from the largest component is 997, which is 15.38% of the total number of papers, and the maximal number of papers per component.

4.2 Community Structure of Largest Connected Components

In order to select the best community detection method for our case study we initially investigated performance of five different community detection methods on the largest connected component. Results are presented in Table 1. It can be observed that the Louvain method shows the best performance for our network: this method reveals a community partition having the highest modularity and the largest percentage of Radicchi strong communities.

Table 1. Comparative analysis of performance of different community detection methods applied to the largest connected component: C – the number of detected communities, Q – modularity score, Strong – the percentage of Radicchi strong communities

Method	C	Q	Strong [%]	Reference
Girvan-Newman edge betweenness	11	0.813	72.7	[7]
Walktrap	23	0.824	82.6	[18]
Infomap	30	0.802	66.7	[22]
Label propagation	29	0.803	79.3	[20]
Louvain	16	0.834	93.7	[3]

Since the Louvain method shows the best performance on the largest connected component we selected this method to investigate the community structure of ten largest connected components in the network. Results are summarized in Table 2. It can be observed that for each component the value of the modularity measure Q is higher than 0.3. Usually a value of Q larger than 0.3 is considered as a clear indication that the network possesses community organization according to the modularity based definition of community [6]. Moreover, the modularity score of the five largest eLib components is even higher than 0.5, and the largest component has the largest value of modularity.

Table 2. Results of community detection for ten largest connected components in the eLib co-authorship graph: N – the number of nodes in the component, Q – modularity score, C – the number of detected communities

N	Q	C	N	Q	C
249	0.834	16	21	0.503	4
74	0.716	8	19	0.486	3
37	0.507	4	19	0.500	4
27	0.531	5	18	0.435	5
25	0.583	4	17	0.334	3

To investigate the quality of obtained community partitions we examine in detail the communities detected in the three largest connected components. Figure 1 shows the visualization of the largest connected component after community detection, while Table 3 provides a description of the obtained communities.

The largest cohesive subgroup is organized around Ivan Gutman who is the best connected eLib author and the most productive author. The central figure in the second largest community is Žarko Mijajlović who is the most central author according to the betweenness centrality metric. The third largest community which is organized around Jovan Karamata (1902–1967) encompasses the oldest generation of authors present in eLib journals, also including Paul Erdős. From this community the whole component started to emerge: the first collaboration among eLib authors is the collaboration between Jovan Karamata and Hermann Wendelin which was established in 1934. It can be observed that for each detected community the number of intra-community links (denoted by "IntraL" in Table 3) is significantly higher than the number of inter-community links (denoted by "InterL"). The same holds also for the sum of weights of intra-community ("IntraW") and inter-community ("InterW") links which means that the overall strength of collaboration among members of each community is higher than the strength of collaboration among authors belonging to different communities. Moreover, each of the detected communities, except community C6, is Radicchi strong which means that each author from a community collaborates more often with authors from his/her community than with authors from other communities. In case of community C6 there are only two authors who are not Radicchi strong: (1) Slobodan Simić has 9 joint publications with members of his community and 10 joint publications with members of communities C1 and C5, and (2) Vlajko Kocić has 1 joint publication with Slobodan Simić and 3 joint publication with Jovan Kečkić who belongs to community C5. For the majority of detected communities (all of them except for C3, C5 and C6) the author having the highest degree centrality in the community (shown in Table 3) is at the same time the author who is most central according to the betweenness centrality metric.

Figure 2 shows the structure of the second largest connected component after community detection. The characteristics of the partition are given in Table 4. It can be observed that for each detected community the number of intra-community links is significantly higher than the number of inter-community links. The same also holds for the sum of weights of this two types of links. Moreover, each detected community is Radicchi strong which clearly suggests that the applied community detection technique produced a good partition into communities. The authors having the highest degree centrality in communities denoted by C1, C4, C5, C6 and C8 are Serbian mathematicians affiliated with the University of Novi Sad. Community C5 is organized around Bogoljub Stanković, a Serbian Academician from Novi Sad, who is the author with the maximal value of timespan for the whole network in the examined time period: the first paper of Bogoljub Stanković published in eLib journals is from 1953, while the last one is from 2011. For 6 out of 8 communities (all except C2 and C7) the author having the highest degree in the component is also the author with the highest betweenness centrality. The authors having the maximal betweenness centrality in C2 and C7 are Miroslava Petrović-Torgašev and Ratko Tošić, respectively.

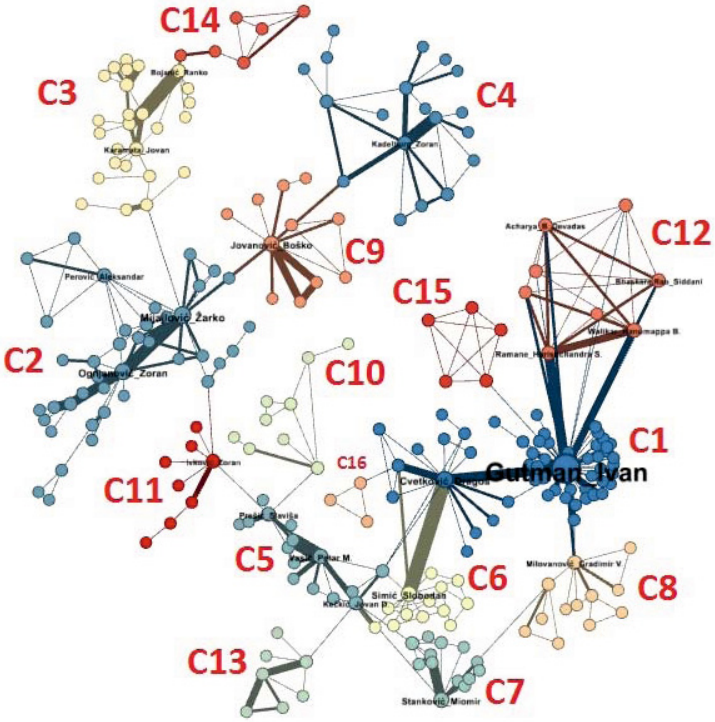

Fig. 1. Visualization of the largest connected component in the eLib co-authorship graph. Nodes from the same community are in the same color. Additionally, each community is marked with an appropriate identifier (C1, C2, etc.) used in Table 3.

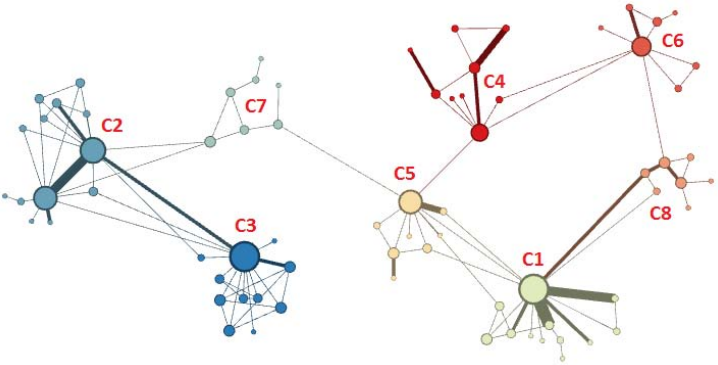

Fig. 2. Visualization of the second largest connected component in the eLib co-authorship graph after community detection

Table 3. Description of detected communities for the largest connected eLib component

Community	Size	Max. degree author	IntraL	InterL	IntraW	InterW	Strong
C1	54	Ivan Gutman (50)	82	15	108	33	yes
C2	40	Žarko Mijajlović (16)	66	4	106	6	yes
C3	26	Jovan Karamata (8)	35	2	64	3	yes
C4	19	Zoran Kadelburg (7)	25	1	42	2	yes
C5	15	Petar M. Vasić (10)	23	8	42	10	yes
C6	13	Slobodan Simić (12)	20	4	20	13	no
C7	13	Miomir Stanković (11)	23	3	34	3	yes
C8	12	Gradimir Milovanović (8)	15	3	18	4	yes
C9	11	Boško Jovanović (12)	14	3	26	6	yes
C10	9	Jovan Petrić (5)	10	1	11	1	yes
C11	8	Zoran Ivković (8)	7	2	9	2	yes
C12	8	Ramane Harishchandra (8)	21	8	31	18	yes
C13	7	Svetozar Milić (5)	8	1	16	1	yes
C14	6	Snežana Pejović (4)	8	1	10	2	yes
C15	5	Song Zhang (5)	10	1	10	1	yes
C16	3	Bolian Liu (3)	3	1	3	1	yes

Table 4. Description of detected communities for the second largest connected eLib component

Community	Size	Max. degree author	IntraL	InterL	IntraW	InterW	Strong
C1	14	Stevan Pilipović (13)	18	5	28	6	yes
C2	13	Leopold Verstraelen (11)	19	6	25	7	yes
C3	11	Ryszard Deszcz (13)	19	4	20	5	yes
C4	9	Dragoslav Herceg (7)	10	3	14	3	yes
C5	8	Bogoljub Stanković (10)	9	6	12	6	yes
C6	7	Djurdjica Takači (8)	8	3	9	3	yes
C7	7	Mirjana Djorić (4)	7	3	7	3	yes
C8	5	Arpad Takači (5)	5	2	6	3	yes

The third largest connected component in the eLib co-authorship network encompasses eLib authors who published their papers in two eLib journals: "Computer Science and Information Systems" and "Review of the National Center for Digization". The scope of mentioned journals is not purely mathematical, but oriented to applications of mathematics and computer science, where the number of authors per paper is generally higher compared to pure mathematical research. Consequently, this component is denser than the previously two described connected components. The details of obtained communities for the third largest component are provided in Table 5. It can be observed that all detected communities are Radicchi strong. Additionally, for each component the author having the highest degree centrality has the highest betweenness centrality.

Table 5. Description of detected communities for the third largest connected eLib component

Community	Size	Max. degree author	IntraL	InterL	IntraW	InterW	Strong
C1	12	Pedro Henriques (13)	25	16	49	19	yes
C2	11	Ivan Luković (10)	18	4	21	4	yes
C3	9	Marjan Mernik (17)	23	17	33	20	yes
C4	5	Bryant R. Barrett (5)	10	5	10	5	yes

5 Concluding Remarks

The project of the electronic library of the Mathematical Institute of the Serbian Academy of Sciences and Arts (eLib) was founded in order to provide online presence and long-term preservation of mathematical journals printed in Serbia. In this study we used eLib bibliographical records to construct the co-authorship network of eLib authors and to identify cohesive subgroups in the network.

Analysis of connected components of the network revealed that the network contains a large number of components. The majority of them are isolated authors or small trivial components, but there is also a small number of relatively large, non-trivial components of connected authors. The main contribution of this article is that we showed that the largest connected components of the eLib co-authorship graph possess clear community structure. This means that authors belonging to the largest components are organized into non-overlapping cohesive subgroups. Additionally, we showed that the majority of identified groups tend to be strong in the sense that each author from a group collaborates more often with authors from his/her group than with authors from other groups.

Acknowledgments. Miloš Savić, Mirjana Ivanović and Miloš Radovanović gratefully acknowledge the support of this work by the Serbian Ministry of Education, Science and Technological Development through project no. OI174023. Zoran Ognjanović, Aleksandar Pejović and Tatjana Jakšić Kruger gratefully acknowledge the support of this work by the Serbian Ministry of Education, Science and Technological Development through project no. III44006.

References

1. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (1999)
2. Batagelj, V., Mrvar, A.: Some analyses of Erdős collaboration graph. Social Networks 22(2), 173–186 (2000)
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (2008)
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.: Complex networks: Structure and dynamics. Physics Reports 424, 175–308 (2006)

5. Brunson, J.C., Fassino, S., McInnes, A., Narayan, M., Richardson, B., Franck, C., Ion, P., Laubenbacher, R.: Evolutionary events in a mathematical sciences research collaboration network. ArXiv e-prints (2012)
6. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. Proceedings of the National Academy of Sciences 104(1), 36–41 (2007)
7. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99(12), 7821–7826 (2002)
8. Gollapalli, S.D., Mitra, P., Giles, C.L.: Ranking authors in digital libraries. In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, JCDL 2011, pp. 251–254. ACM, New York (2011)
9. Grossman, J.W., Ion, P.D.F.: On a portion of the well known collaboration graph. Congressus Numerantium 108, 129–131 (1995)
10. Grossman, J.: The evolution of the mathematical research collaboration graph. Congressus Numerantium 158, 201–212 (2002)
11. Grossman, J.: Patterns of collaboration in mathematical research. SIAM News 35(9), 8–9 (2002)
12. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM 2003, pp. 556–559. ACM, New York (2003)
13. Mijajlović, Z., Ognjanović, Z., Pejović, A.: Digitization of mathematical editions in Serbia. Mathematics in Computer Science 3(3), 251–263 (2010)
14. Newman, M.E.J.: Scientific collaboration networks I: Network construction and fundamental results. Physical Review E 64, 016131 (2001)
15. Newman, M.E.J.: Scientific collaboration networks II: Shortest paths, weighted networks, and centrality. Physical Review E 64, 016132 (2001)
16. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45, 167–256 (2003)
17. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69(2), 026113 (2004)
18. Pons, P., Latapy, M.: Computing communities in large networks using random walks. Journal of Graph Algorithms and Applications 10(2), 191–218 (2006)
19. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proceedings of the National Academy of Sciences of the United States of America 101, 2658–2663 (2004)
20. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E 76, 036106 (2007)
21. Rodriguez, M.A., Bollen, J.: An algorithm to determine peer-reviewers. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, pp. 319–328. ACM, New York (2008)
22. Rosvall, M., Bergstrom, C.T.: Maps of information flow reveal community structure in complex networks. Proceedings of the National Academy of Sciences of the United States of America 105, 1118–1123 (2007)
23. Savić, M., Ivanović, M., Radovanović, M., Ognjanović, Z., Pejović, A., Jakšić Krüger, T.: The structure and evolution of scientific collaboration in Serbian mathematical journals. Scientometrics (to appear, 2014)

Employing Personal Health Records for Population Health Management

Ana Kostadinovska[1], Gert-Jan de Vries[2], Gijs Geleijnse[2], and Katerina Zdravkova[1]

[1] Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University, Skopje, Macedonia
kostadinovska.ana@gmail.com, katerina.zdravkova@finki.ukim.mk
[2] Philips Research – Healthcare, Eindhoven, The Netherlands
{gj.de.vries,gijs.geleijnse}@philips.com

Abstract. Linking various sources of medical data provides a wealth of data to researchers. Trends in society, however, have raised privacy concerns, leading to an increasing awareness of the value of data and data ownership. Personal Health Records address this concern by explicitly giving ownership of data to the patient and enabling the patient to choose whom to provide access to their data. We explored whether this paradigm still allows for population health management, including data analysis of large samples of patients, and built a working prototype to demonstrate this functionality. The creation and application of a readmission risk model for cardiac patients was used as carrier application to illustrate the functionality of our prototype platform.

Keywords: Personal Health Records, Population Health Management.

1 Introduction

Modern technology more and more enables gathering, storage and coupling of various datasets. Telecom providers store usage and location of mobile phones that we carry all day, internet companies store and analyze patterns of web usage, banks are using spending patterns for targeted advertisements, and there are many more examples. By coupling such databases, even more rich information can be obtained, which can be used to our advantage, however, privacy concerns are becoming more and more apparent [1,2]. While some applications can be rather harmless, concerns are more serious when health related data are involved. Coupling various sources of healthcare data, such as hospital information systems, healthcare insurance data, home monitoring devices, general practitioner databases, etc. may enable more precise and personalized care, at lower cost. Unsurprisingly, concerns about ownership and privacy of health data do exist [3]. In most current healthcare information systems, the gatherer of the data, e.g., a hospital, insurance company or GP is considered the owner of the data. More and more people become aware of the value of their data and would like to have additional control of the access to their personal data. Personal Health Records (PHR) meet this need and aim at collecting healthcare data from these various

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_7

sources, whilst empowering the patient as the owner of the data to decide who to give an authorization to have access to his/her data [4].

In the PHR model, the patient is the only stakeholder with access to the full and holistic overview of the data. This allows for richer data analysis than in current health care data models. To that end, it is important to be able to analyze data form multiple patients. The decentralized ownership in PHRs, however, makes it more difficult to collect such a dataset. Currently, PHRs do not provide a solution to this problem.

We studied options for using PHR data for such research purposes, whilst maintaining the PHR philosophy of empowering the patient. We created a working prototype framework, which we termed an intelligent PHR, which runs on top of Microsoft HealthVault [5] and can apply implemented services on the available data. Examples of such services are statistical analysis methods to perform descriptive or predictive analysis, or the application of developed predictive models. We developed predictive risk models using a dataset of cardiac patients. These risk models were implemented in the intelligent PHR to demonstrate its functionality and can be used for both personal and population level risk prediction using PHR data.

In the remainder of this paper, we will first describe the state of art with respect to PHRs and reveal how the care for cardiac patients can benefit from PHRs after which the methods used to develop the system on top of an existing PHR are presented, followed by its architecture. The paper concludes with a brief discussion and conclusion.

2 Personal Health Records

A PHR is a system of health-related information of a patient, which is managed, shared and controlled by the patient (rather than individual care providers). It contains data from various sources: e.g., clinical data measured by a health care organization, but also home monitoring data, measured by patients themselves. It is a form of an EHR (Electronic Health Record), but, in contrast to traditional EHRs, PHRs are not hosted and managed by a health care organization, but managed by patients. That is, a PHR is accessible online by the patients and by anyone they specifically gave consent to access their information. Therefore, it has the potential to collect a richer dataset by enabling the collection, monitoring and organization of health data on a daily basis, and sharing and querying health and personal information [6]. The information collected in a PHR might include: personal information of the patient, lab results, symptoms, vitals, exercise and dietary habits, health goals (such as to stop smoking) and data from devices (such as electronic weight scales).

Another important difference is that PHRs are aimed not only for patients in a clinical context, as EHRs are typically focused on, but also for (former) patients in other contexts as well as healthy individuals. Hence, PHRs also allow individuals to manage their health and wellbeing by monitoring appropriate vital signs. A particular group of interest is chronic patients, who after an acute phase during which they

receive intensive medical care, enter a period of chronic care including self-care which involves close self-monitoring of their condition. To provide pro-active longitudinal care, predictive models that assess future care needs may be of use. In the following we will elaborate why PHRs are particularly interesting for chronic patients, and in particular for cardiac patients.

There are several PHR systems available, including My HealtheVet, MyChart, My Health Manager, Microsoft HealthVault, Health Space, Dossia, Tolven. Out of these, we selected Microsoft HealthVault [5]. Microsoft launched HealthVault, as an interconnected PHR system, in October 2007 in the United States and nowadays is available also in United Kingdom, Canada and Germany. It is defined as a "Cloud-based platform designed to put people in control of their health data" and enables its users to manage their own PHR and was designed to put the users in full control of their health data. Patient level services can be implemented in HealthVault, but the platform currently does not support population level applications and analyses.

2.1 Datasets for Cardiac Patients

Chronic diseases become increasingly prevalent in Western populations; illustrated by the fact that for example 49% of the US population in 2005 had at least one chronic condition [7]. Cardiac conditions form one of the most prevalent chronic diseases and are characterized by high mortality and readmission rates. In 2009, 30-day readmission rates in the US were 17.1% after a heart attack, with average costs of re-hospitalization of $13,200 [8].

Care for these patients involves a plurality of aspects, including medical interventions, medication, daily monitoring of vitals, regular follow-up checks, but also lifestyle and dietary changes. For this reason, there are many stakeholders involved and lots of different places where data is gathered. One central place where data is contained could really benefit the care for these cardiac patients. In addition, especially when lifestyle and dietary changes are required, patient engagement is key to success. By giving patients a central role in their health data management, PHRs have the ability to further motivate patients to engage in their health management.

Although the quality of care for patients with cardiac conditions has made enormous progress over the past decades, cardiac patients are still often admitted to the hospital [9, 10], with even higher rates for heart attack patients [11], which triggered research of predictive risk models [12, 13, 14]. With such predictive risk models, it is possible to predict adverse events in an early stage and thereby enable early intervention before a costly adverse event happens. It is believed that many readmissions can be prevented by better (planned) care as well as an earlier detection of the onset of worsening symptoms [9, 10, 15]. The research on such models is still in an exploratory phase, and will therefore benefit from the collection of as much data as possible through PHRs. In the framework that we propose on top of PHRs, the development of new risk models and the application of existing risk models can be seen as examples of services that require input data from at least one patient.

3 Methods

The design process we followed for the development of a system on top of a PHR to enable population based management within the PHR paradigm consists of the following steps. First we defined the stakeholders involved in using the system. Second, we created use cases, and third, we designed the architecture of the system. In order to present its functionality, we also created and applied risk models to our cardiac dataset. As part of the initiation of the design process, we sketched the context in which the intelligent PHR would be implemented. Furthermore, we needed to understand usage of the system.

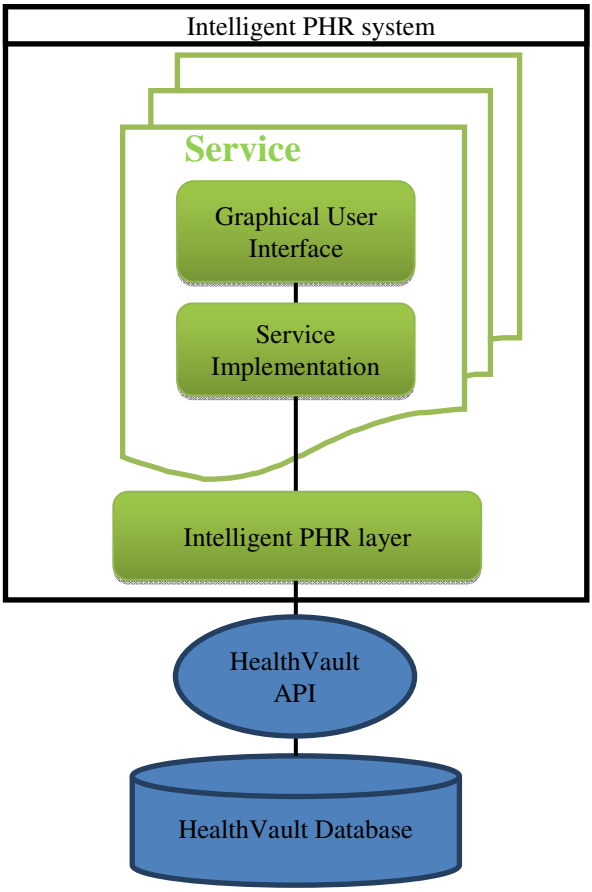

Fig. 1. High level architecture of the intelligent PHR system

3.1 Implementation Context of the Intelligent PHR System

The aim of the intelligent PHR system is to enable services to make use of Microsoft HealthVault on a population level. Therefore, the intention of the system is to extend the functionality of Microsoft HealthVault while maintaining the philosophy of PHR that patients are in charge of their data. These services will make use of data obtained through the HealthVault API. Microsoft HealthVault allows the retrieval of information at real time, such that there is no need for a local storage in the intelligent PHR. This high level architecture is depicted in Figure 1.

3.2 Usage of the Intelligent PHR System

With the intention of finding a solution for the problem we stated, we should mainly focus on the needs and the unmet demands of the stakeholders. The stakeholders that have the greatest effect as well as the biggest benefit from the improvement and adoption of the PHR systems are patients, health workers and researchers.

Patients are motivated to use these systems mainly because they are in personal control of their health. Using a PHR system, they can manage their lifelong health information and their chronic diseases together with their care givers, and also their health can be easily monitored by their family. The need of continuous communication with their care givers, not only in the hospital but also at home, has an essential role in the prevention of the readmissions and worsening of their health conditions.

Health workers are focused on providing the best care to patients while minimizing costs. Using the available applications in the PHR system, they can support their patients' care by monitoring clinical and laboratory data in their PHR record. Online consultations, scheduling and medication refill are benefits that can lead to better health condition of the patients, reduced readmission rates and thereby reduced healthcare costs.

Researchers are interested in analyzing population level data and the development of predictive models which can be applied by patients and health workers. By predicting adverse events using risk models, early intervention can be done to reduce the impact of adverse events or ultimately perhaps prevent them.

In order to design an intelligent PHR system we created use cases based on the needs of these three stakeholders for using services during or after the hospitalization of the patient. These use cases describe the usage of the services in the intelligent PHR system. The difference in the usage depends on the actors in the use cases and where they can use the services. These services may range from generic data inference services to specific risk models. As an example service in our system, we focused on risk model services. We also took into account that patients can be in different care locations (e.g., in the hospital or at home) while using the system, posing different requirements to the system.

First, a cardiologist, during hospitalization of a patient, wants to be able to evaluate the outcome of a single or compare multiple risk models using the PHR of the patient. These risk models can be of great help for the cardiology department to stratify patients, since many undesirable events can be prevented by delivering additional care and support to those at high risk for an early adverse event.

Second, the risk models can not only be used in hospital, but also during care at home. After the patient is dismissed from the hospital, the same functionalities at home are available to health workers involved, in order to prevent adverse events that can occur to the patient. Furthermore, after dismissing the patients from hospital, the patients can take better care of themselves by evaluating the results of the risk models that are calculated by the health worker. For example, the awareness of being at high risk for a hospitalization may help to adhere to lifestyle changes or support therapy adherence. In Section 4 we will elaborate the most important use cases of applying a service, such as the application of a risk model, to a set of patients.

3.3 Development of an Example Service

As an example service, we will create and apply risk models for the prediction of readmission within one year from hospitalization for ACS patients. For that purpose, we used a dataset that contained a variety of features including demographics, medical history, medication usage, vitals, and lab values. We performed feature selection using Paired t-tests [16] to identify which features distinguish readmission from no-readmission to enough extent. We applied a liberal threshold to the significance level (p-value < 0.4) found in the test to include all the features with a lower p-value than the threshold as input in our risk model. The features were normalized using z-score normalization before applying a machine learning techniques to develop a classifier. To that end, we trained models using two different types of Learning Vector Quantization (LVQ) algorithms, namely Generalized LVQ (GLVQ) and Robust Soft LVQ (RSLVQ) [17]. This type of classifier uses prototypes that are defined in the original data space to represent the classes, which allows inspection and interpretation of the knowledge gained by the classifier in terms of the original data space. We used one prototype per class, and used 10-fold cross validation to estimate generalization performance, measured by accuracy.

To the best of our knowledge, no readmission risk models have been developed for ACS patients, however there are a limited number of mortality risk models. Analogue to the approach by Auble et al., who benchmarked against heart failure risk models that were designed for readmission to predict mortality instead [18], we used the Thrombolysis in Myocardial Infarction (TIMI) STEMI model [19] as a reference.

4 Results

We created an intelligent PHR system that allows patients and care givers to make use of PHRs in Microsoft HealthVault, by applying smart services to the data. The architecture provides the bridge between the PHRs and any service that uses data from the PHRs. The architecture enables the patients to manage their health information, and allows selected care givers to access this information and communicate with the patient. It uses proven means to access selected information in a secure and privacy preserving way. Having built this architecture, the usage of intelligent algorithms can be explored, to provide meaningful decision support for clinician and patient.

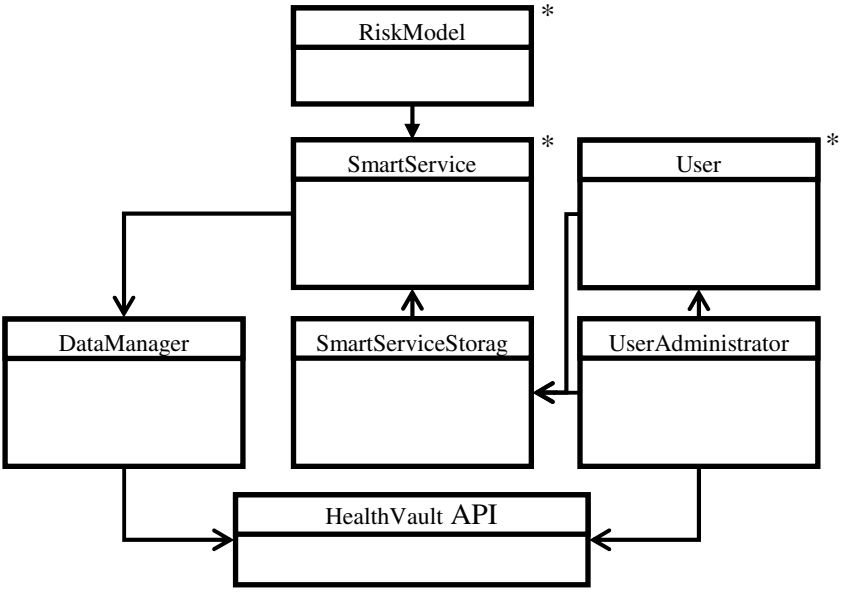

Fig. 2. Architecture of the intelligent PHR system (stars indicate multiply instantiated classes)

The architecture of the intelligent PHR is presented in Figure 2. In order to safeguard the PHR principle of patients being in charge of their own data, we implemented some user management that is required to ensure that only selected (by the patient) users can access a patient's data. The process that includes getting permission from the patient for accessing the necessary data from a user (e.g., care giver) consists of several steps, which are outlined further in the paper using the use case of applying a population level service to data of a set of patients. Given that a healthcare professional has selected a service that he wants to apply to a selected set of patients:

- For each of the patients in the selected set: Get an authorization code from Microsoft HealthVault, by creating a connect request that is based on the patient's ID, friendly name and secret question and answer. Data access is requested for the combination of the particular healthcare professional and selected service.
- Send an email to each selected patient, containing the identity code, a link to Microsoft HealthVault[1] and an information letter on the purpose of the data usage. Via a separate medium (e.g., by phone, traditional mail or a by email to a secondary email address) the secret question and answer are also provided to the patient.
- Through the opt-in mechanism of Microsoft HealthVault, the patients can now provide authorization through the following steps:
  - Go to the provided link and enter the identity code provided by the application.
  - Enter the secret answer to the required secret question.
  - Select the HealthVault record to be used by the application and authorizes it.

[1] `https://account.healthvault-ppe.com/PatientWelcome.aspx`

- Periodically check whether the patients completed the authorization. This periodical check is performed, until the patients give consent or until the request expires.
- After the authorization is completed, the intelligent PHR can pass the data in the patient's Microsoft HealthVault record to the service.
- When at least one patient has provided consent, the healthcare professional can apply service to the data of the patients who provided consent. The intelligent PHR sends data requests to each patient, collects the data and applies the service. The result is passed to the healthcare professional using the GUI of the service.

4.1 Example Service: Readmission-Risk Model

After applying the feature selection, the following set of features was included in the model:

- Albumin
- Alkaline Phosphatase
- Calcium
- Cholesterol
- Globulin
- Mean Cell Haemoglobin
- Mean Cell Volume
- Red Blood Cell Count
- Troponin I Ultra
- Non-smoking history
- Systolic Blood Pressure
- Diastolic Blood Pressure
- Heart rate
- Grip strength left hand

Based upon these features, several classifiers were trained. Table 1 shows the percentage of correctly classified readmissions in one year using the GLVQ and RSLVQ algorithm in 10-fold cross validation. The performances were better than the reference algorithm. We implemented the predictive models as smart services in the intelligent PHR, which allows the application to individual patients, but also to a set of patients, e.g., to validate the model on another patient sample.

Table 1. The percentage of correctly classified readmissions within one year for ACS patients

	Accuracy
Reference (TIMI)	65.1%
GLVQ	72.9%
RSLVQ	73.5%

5 Conclusion and Outlook

In this paper we have outlined how PHRs can be beneficial in the care for chronically ill patients, in particular cardiac patients. We identified and implemented a means to allow researchers to use PHRs to perform population level analyses whilst maintaining the PHR philosophy of empowering the patient as owner of his healthcare data deciding who gets access. By doing so, we built upon and maintained the privacy measures taken by PHR providers. We have implemented a working prototype and

used data from the cardiac domain to demonstrate its functionality. The developed risk model for readmission of AMI patients was successfully implemented and enables the calculation of patient level risks on a population of patients whose data resides in a PHR. Although we focused on chronic cardiac patients, there the intelligent PHR framework can in principle be used in the care for any other type of patient; however, we foresee most added value for patients with chronic diseases.

In future use of the proposed architecture on top of PHRs we foresee that researchers can provide search criteria along with a consent form to the PHR management system to screen for patients given certain in-/exclusion criteria. The PHR management system can then forward a request for participation with the consent form attached to eligible patients. Then, an opt-in mechanism, as introduced in this paper, can be used to digitally enroll patients in the study. Other topics that require further attention include integration into other PHR systems, preferably using a unified data model such as Resource Description Framework (RDF) [20]. Given that PHR data can come from any source, it would be good to have a label attached to data samples that indicates a confidence level of correctness. Settings this will not be trivial though.

Acknowledgements. The authors would like to thank the University of Sheffield for providing access to the AMI patient dataset. This work was supported by the European Commission in the context of the VPH-Share project (FP7-ICT-269978) and the authors wish to thank the whole consortium.

References

1. Flemmix, P.: Panopticon (2013), http://panopticondefilm.nl/ (last accessed: June 01, 2014)
2. MarketingCharts: Privacy A Growing Concern For Almost 2 in 3 Internet Users (2013), http://www.marketingcharts.com/wp/online/privacy-a-growing-concern-for-almost-2-in-3-internet-users-36781/ (last accessed: June 01, 2014)
3. Meingast, M., Roosta, T., Sastry, S.: Security and Privacy Issues with Health Care Information Technology. In: 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2006, pp. 5453–5458 (2006)
4. Tang, P.C., Ash, J.S., Bates, D.W., Overhage, J.M., Sands, D.Z.: Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption. J. Am. Med. Inform. Assoc. 13(2), 121–126 (2006)
5. Microsoft HealthVault, https://www.healthvault.com (last accessed: June 1, 2014)
6. Lee, M., Delaney, C., Moorhead, S.: Building a personal health record from a nursing perspective. International Journal of Medical Informatics 76 (2007)
7. Schneider, K.M., O'Donnell, B.E., Dean, D.: Prevalence of multiple chronic conditions in the United States' Medicare population. Health Qual. Life Outcomes 7, 82 (2009)
8. Stranges, E., Barrett, M., Wier, L.M., Andrews, R.M.: Readmissions for Heart Attack. Healthcare Cost Utilization Project, Statistical Brief, 140 (2009)

9. Jennings, D.L., Petricca, J.C., Yageman, L.A., O'Dell, K., Kalus, J.S.: Predictors of RehospitalizationAfter Acute Coronary Syndromes. American Journal of Health System Pharmacy 63(4) (2006)
10. Kociol, R.D., Lopes, R.D., Clare, R., Thomas, L., Mehta, R.H., Kaul, P., Pieper, K.S., Hochman, J.S., Weaver, W.D., Armstrong, P.W., Granger, C.B., Patel, M.R.: International Variation in and Factors Associated With Hospital Readmission After Myocardial Infarction. American Medical Association 307(1) (2012)
11. Andres, E., Cordero, A., Magan, P., Alegria, E., Leon, M., Luenqo, E., Botaya, R.M., Garcia Ortiz, L., Casasnovas, J.A.: Long-Term Mortality and Hospital Readmission after Acute Myocardial Infarction: an eight-year follow up study. Revista Espanola de Cardiologia 65(5) (2012)
12. Ross, J.S., Mulvey, G.K., Stauffer, B., Patlolla, V., Bernheim, S.M., Keenan, P.S., Krumholz, H.M.: Statistical Models and Patient Predictors of Readmission for Heart Failure: A Systematic Review. Arch. Intern. Med. 168(13), 1371–1386 (2008)
13. Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., Kripalani, S.: Risk Prediction Models for Hospital Readmission A Systematic Review. JAMA: The Journal of the American Medical Association 306(15), 1688–1698 (2010)
14. De Vries, J.J.G., Geleijnse, G., Tesanovic, A., Van de Ven, A.R.T.: Heart Failure Risk Models and Their Readiness for Clinical Practice. In: 2013 IEEE International Conference on Healthcare Informatics (ICHI), pp. 239–47 (2013)
15. Desai, M.M., Stauffer, B.D., Feringa, H.H., Schreiner, G.C.: Statistical Models and Patient Predictors of Readmission for Acute Myocardial Infarction: a systematic review. Circulation. Cardiovascular Quality and Outcomes 2(5) (2009)
16. McDonald, J.: Student's t-test, Handbook of Biological Statistics, pp. 118–122. Sparky House Publishing, Baltimore (2009)
17. Witoelar, A.W., Ghosh, A., de Vries, J.J.G., Hammer, B., Biehl, M.: Window-Based Example Selection in Learning Vector Quantization. Neural Computation 22(11), 2924–2961 (2011)
18. Auble, T.E., Hsieh, M., McCausland, J.B., Yealy, D.M.: Comparison of Four Clinical Prediction Rules for Estimating Risk in Heart Failure. Annals of Emergency Medicine 50(2), 127–135 (2007)
19. Morrow, D.A., Antman, E.M., Charlesworth, A., Carins, R., Murphy, S.A., de Lemons, J.A., Guigliano, R.P., McCabe, C.H., Braunwald, E.: TIMI risk score for ST-elevation myocardial infarction: A convenient, bedside, clinical score for risk assessment at presentation: An intravenous nPA for treatment of infracting myocardium early II trial substudy. Circulation 102 (2000)
20. Resource Description Framework (RDF), `http://www.w3.org/RDF` (last accessed: June 1, 2014)

A Comparison and Integration of Ontologies Suitable for Interoperability Extension of SCOR Model

Srdja Bjeladinović and Zoran Marjanović

Faculty of Organizational Sciences, University of Belgrade, Belgrade, Serbia
{srdja.bjeladinovic,zoran.marjanovic}@fon.bg.ac.rs

Abstract. Systems with numerous stakeholders are very common in contemporary business. They can grow pretty fast in all aspects, including number of different semantic domain they use, and the way they interpret them. Supply Chain Management systems (SCM) are not an exception, at all. In order to achieve unique interpretation of information between stakeholders, reference models are made. Particular popular one for SCM is SCOR (Supply Chain Operations Reference) model. However, lack of semantic richness and precision has led to the need to expand the SCOR model. In the literature and practice often used as a solution for this problem is applying semantically rich ontologies in order to contribute a precise definition, interpretation and expansion of the domains within the SCOR model. Use of ontologies raises the question of possibility of their mutual integration and interoperability, which in this paper is processed as a comparative analysis of their concepts, similarities and differences (for three commonly used ontologies: IDEON, TOVE and Enterprise). We developed a model that represents the basis for further integration of these three ontologies and for further interoperability expansion of SCOR model, therefore the SCM. This model is described at the end of this paper.

Keywords: System interoperability, SCOR model, ontologies.

1 Introduction

Number of participants inside business systems is constantly increasing. Participants use different technologies, different standards in different aspects and different data models. Consequently, two main questions emerge as crucial ones. The first one is how participants in any type of communication understand each other, while the second one is how interoperability issue between them should be resolved. The aim of this paper is to provide answer to those questions. As a case study, we will use the example of Supply Chain Management (SCM), which is no exception to any other system with mentioned issues.

SCM is a dynamic environment that consists of organizations, people, technology, activities, information and resources [1]. Typical SCM consists of many stakeholders and many participants. In order to ensure smooth cooperation among all elements within the supply chain and to retain high level of system integrity, reference models

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_8

are developed. One of them which stands out is Supply Chain Operations Reference (SCOR) model, developed by the Supply Chain Council [2]. Reference models provide needed framework for specifying unique interpretation of semantic on certain domain. Main disadvantages are fact that many different reference models can exist and that different models can have different concepts and different ways of interpretation of knowledge and information. Even if one reference model is chosen as defacto standard, still main purpose of reference model is not to be semantically rich, but to provide understandable and uniform interpretation of information. Because of that reason, even a commonly used reference model, like SCOR for SCM, can have narrow semantic scope. According to some authors [3,4], semantic enrichment of model can be achieved by using ontologies, which enable easier way for defining perceptions, formal meaning of information and also to overcome the syntax and semantic gap. Throughout extension of SCOR model with ontologies, positive influence on SCM is made. Each of the ontology's has certain peculiarities. The issues raised are the extent to which these ontologies are similar, is there a way to integrate various ontologies and how to achieve the desired interoperability when different business entities use different ontologies to expand SCOR model.

The paper is organized as follows: Section 2 describes main concepts and characteristics of SCOR model. In Section 3 of this paper a comparative analysis of commonly used ontologies for SCOR model was conducted. Ontologies concepts are also described in the Section 3, as well as relationship between them. Section 4 describes model that we have been developing and which represents the basis for further integration of this three ontologies. This model also presents base for further interoperability increase of SCOR model. Concluding remarks are provided in the Section 5, as well as direction of further research.

2 Reference Model and Their Characteristics

SCOR model (Supply Chain Operations Reference model) is reference model for SCM [15,16,17,18,19,20], developed by the Supply Chain Council. Certain authors emphasize its importance not only because it is Supply Chain Council standard, but also because of its contribution to the strategic management within the supply chain [2] [5,6]. SCOR model provides a unique framework that links business processes, metrics, best practices and technology into a single structure and also supports process of communication among the stakeholders in the supply chain in order to improve efficiency [5].

The main advantage of SCOR model is reflected in standardized set of concepts that are used by all participants in the supply chain. Every single concept, simple or complex, is specified with the set of basic building blocks. Each block has standardized and uniformed meaning, which resulting that all participants in the supply chain interpret all the information in a unique way. SCOR model has attributes. Some of these are reliability and ability to perform as expected, the response time, speed of performing tasks, ability to respond to external influences and the ability to change, the cost of the process and resource management. However,

reference models (including the SCOR model) are not developed to be semantically rich nor precise, but to provide an understandable knowledge for a particular domain. Result of applying ontology on SCOR model is more precise definition of syntax (through the introduction of formalisms), expanded number of concepts and enriched semantics.

3 Ontologies, Their Concepts and Comparison

Reference model defines a way of interpreting the information, but model is not sufficient if it is necessary to expand concepts in domain. The use of reference model for a certain domain has already been discussed in the previous section, and for the purpose of extended its concept, ontologies may be used [21].

For each ontology exists certain set of specific concepts, but also a set of concepts which are more or less common. However, the usage of different ontologies for the same reference model may jeopardize the interpretation of each model. In a situation where more participants within SCM use different ontologies to expand SCOR model, questions that should be asked are how to harmonize the concepts and meanings and how to achieve interoperability. In order to display the possibility of connecting several different ontologies through a common model, three the most widely used ontology for the expansion of the SCOR model (TOVE, IDEON and Enterprise) were compared. The following describes the three ontologies suitable for reducing interoperability problems.

3.1 Toronto Virtual Enterprise (TOVE) Ontology

TOVE ontology project was initiated in the laboratory of "Enterprise Integration", University of Toronto [8]. Purpose was to develop environment for the integration of business systems. TOVE aim is to construct a data model that is sufficiently expressive to represent all aspects of business knowledge, both at a general level and at the application level [10]. Linking structure and behaviour of the organization is the focus of this ontology [8]. The project was based on the Knowledge Interchange Format (KIF), i.e. computer language for the exchange of knowledge between the various computer programs [11], which enables the automatic deduction (extract facts from the ontology presented KIF included). One of the biggest advantages of this ontology is a formalized approach to modelling, while one of the biggest drawbacks is ambitiously defined scope, which leaves certain undefined sub-domains.

Basic concepts in TOVE ontologies are represented as entities, resources, relationship and time. Type of resource can be physical, human and information. Resources can be composed of multiple components, and may also enter into the composition of other resources (typically the products). In addition to qualitative state, also quantitative resource state can be changed. History of changes for every resource should be tracked. Time is an important concept in the category of general concepts. Another group of concepts represent actions, activities, tasks, and their condition and status. In TOVE ontology, activities are represented by a combination

of actions. In TOVE ontology organization is viewed as a set of rights and restrictions on the activities carried out by organizational agents. Organizational agent, in the narrow sense, is an individual who is a member of the organization. In a broader sense, agents can represent machines or software. Agent is a member of one or more organizational division and has one or more roles in the organization. Also, an agent can perform activities and interact with other agents, using communication links, all in order to achieve organizational goals. Rights to change the status of a particular agent are realised through concept of empowerment [8].

3.2 Enterprise Ontology

Enterprise ontology was developed within the Enterprise Project, at the Institute AIAI, University of Edinburg, in order to provide methods and computer tools for modelling business systems [9]. Development environment for Enterprise Ontology was created in order to model business systems and to integrate methods and tools. The motivation for the development of this ontology is similar to one for TOVE [10][14]. Because primary goal of this project was facilitating communication between people, less effort has been made to precise formalization of the concepts.

The basic building blocks of Enterprise Ontology are entity, relationship, state and roles. These terms are defined within group of basic concepts, so called meta-ontology. In addition to these units, Enterprise Ontology defines groups of planning and organization [10]. Participant is a term that corresponds to the agent in the TOVE ontology. Only certain entities can realize roles and perform certain actions. Such entities are people, organizational units and in some cases machines. State is determined by one or more entities. The concept of time is defined and used in the same context as in TOVE ontology. Executor of activities should be elected among the potential participants (people, machines or organizational units). Decomposition of activity is supported. Resources that are used in activities are defined in same way as in TOVE ontology. Specification of activities with a defined purpose is called a plan. Enterprise ontology introduces the concept of legal entities and organizational units. The difference between these two concepts is in fact that the legal entities have rights and responsibilities in the business world, while organizational units have only the rights and responsibilities within the organization. Legal entities include individuals and corporations. Owner of rights and responsibilities, from legal point of view, should be a legal entity.

3.3 IDEON Ontology

IDEON is united business ontology that provides the basis for designing, re-engineering, management and control of collaborative, distributed enterprises [12].

Concepts of IDEON ontology are organized into four sections: general concepts, concepts of organization, process concepts, concepts of resources and products. Novelty compared to the previous two ontologies is that each entity type sets the sensors to observe environment. Collecting of information via sensors allow an assessment of the situation, which provides organizations with ability to perform

certain processes or operations in order to achive good effects to environment. Organizational concepts of this ontology is used in order to define structure of organization. There are several types of connections between organizations. Organization can consists a number of smaller organizations. Multiple level parent-child hierarchy is provided, which leads to the branching hierarchy of organization. Two organizations may be descendants of the same organization, and in that situation, organizations cooperate in order to achieve objectives. Cooperation is modelled as a connection within the process.

The concept of resources in IDEON ontology is specialized in human and material resources, where under the human resources are considered persons with the appropriate roles. The concept of roles contains specific information for a particular position, such as required qualifications, set of responsibilities and rights. Cardinality of the relationship between people and the role is "many-many". Further, material resources can be specialized in information resources or natural materials. The basis for this specialization is the fact that information can be directly controlled by the process control system, while physical materials cannot. Product concept represents a physical product that can be sold to the customer, document, service, process or new executive information system. Resource object is an object in the possession of the organization or an object created in some of its processes.

4 Results and Findings

Interoperability of information systems depends on the quality and mutual consistency of appropriate ontologies. Differences in applied ontologies concepts can lead to semantic disagreement, which negatively affects interoperability. The negative effects of inconsistencies in the conceptualization can be reduced by additional mapping, transformation or merge of appropriate ontologies into single model [3].

Described ontologies have certain degree of similarity in concepts definition, as well as the corresponding difference. Besides the obvious differences in the way that concepts are grouped into categories and besides the existence of certain specific concepts at the level of ontology (example: the concept of strengthening the role of TOVE), common set of concepts can still be extracted. Essence of further analysis will be a comparison of ontologies based on similarities and differences of concepts which represent building blocks of each ontology. The emphasis will be on the analysis of the concepts of each ontology, finding common concepts of these ontologies, their characteristics and their relationships with other concepts. Certain concepts in appropriate ontologies can be found under different namesand in different categories. Thus the concept of resource in IDEON ontology can be found in category of products and resources in Enterprise Ontology and in category of meta-concepts in TOVE ontology. In order to neutralize the differences over the naming, all analyzed concepts are divided into three categories, where each of concept from represented ontology is include by similarity. These categories are: organizational concepts, concepts of resources and concepts of activities. For a graphical representation UML class diagram was used.

4.1 Resource Concept Comparison and Integration

The main concept of this sub-model is resource. In all three ontology concept of resources is defined in almost the same way: resource can be material, human or information.

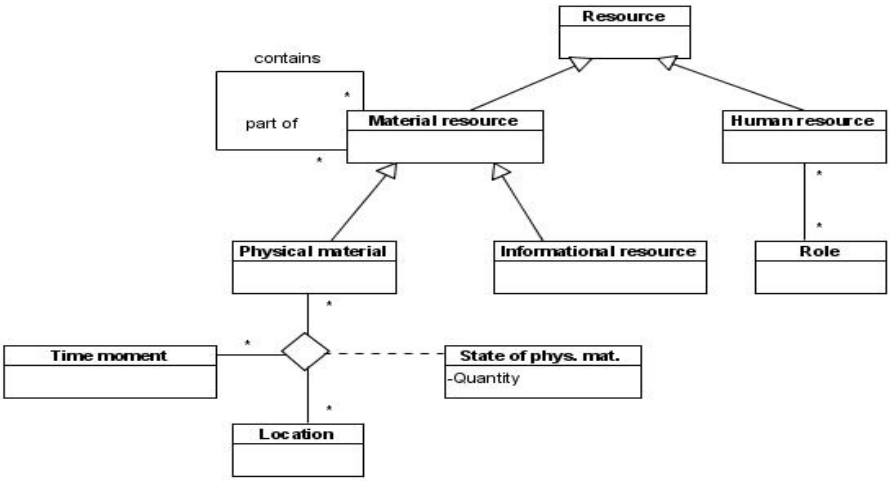

Fig. 1. Resource concept sub-model

Considering that IDEON ontology explicitly stated that material resource can be specialized to either physical or information and considering that it is not inconsistent with neither of two other ontologies specialization is defined. Specialization is done in the following way: resource can be physical or human (human records are kept on assigned roles), while material can further be specialized in physical or informational. For physical resources, beside quality characteristic also quantitative characteristic are recorded, i.e. a certain amount of physical resources at a certain location in a certain point of time. Through the concept of time it is achieved connection with the other two sub-models. Class of material resources has a connection with itself, representing components: each component can be a part of other. Only in TOVE ontology product is classified as a separate resource, and because product can be treated as a organizational resource, in this sub-model product is represented as one of many appearances of material resources. Figure1 shows explained sub-model with resource concepts.

4.2 Activity Concept Comparison and Integration

Main concepts in this category are the activities and actions. In IDEON ontology process appears as concept, which includes a number of activities. Due to the lower level of abstraction, in this model process is not shown as a concept, but as combination of its components: activities. Activities are grouped actions, whose execution gives the desired result. Action is the basic unit of work and as such, exists

within a particular activity. Activities are associated with agents (machines, people etc). Each activity is carried out in exactly one organizational unit. For connected activities each of them is executed in exactly one organizational unit, not necessarily the same. Therefore, the activity is associated with exactly one organizational unit, and with zero or more resources. Since the resources can be defined as machines (material resources) or people (human resources) and since the other resources can be used by one activity (present in all three ontologies, and particularly described in Enterprise Ontology), link between resources and activities was established in this sub-model. It indicates potential resources for the activity.

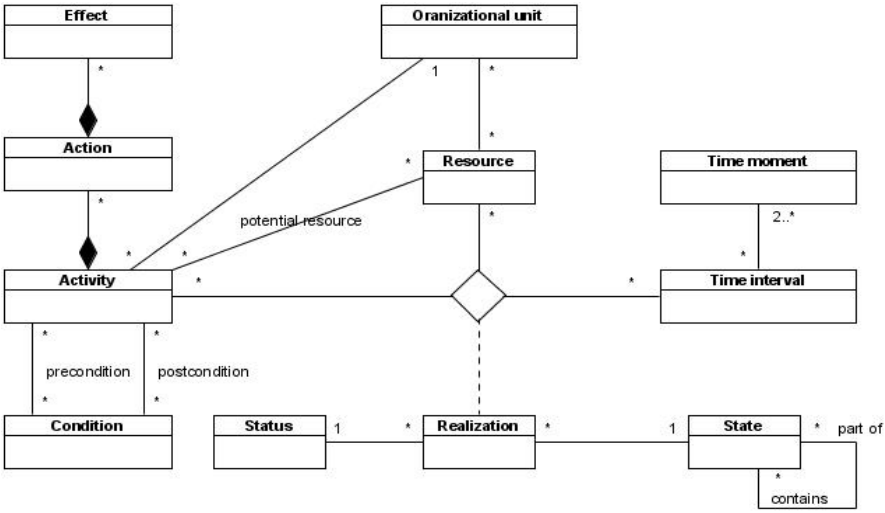

Fig. 2. Activity concept sub-model

Activity is carried out in a specific time interval, by a particular agent. Same agent may perform same action several times, but at different time intervals. Intervals are made up of several time moments, at least two different time moments (initial and final). Implementation of activities has a certain status and a certain state, which is presented with related links. Considering that TOVE introduces the possibility of complex conditions (a combination of several different), connection between state and itself is also represented. Every action, regardless of the time of its implementation, its status or its state is correlated with certain conditions (preconditions), which must be fulfilled. Conditions that should be fulfilled after the execution of the activities are presented as post conditions. Figure 2 shows described sub-model with concepts of resources.

4.3 Organizational Concept Comparison and Integration

Basic concept of this category is an agent, defined within TOVE ontology. Agent is interrelated to appropriate resources (human or material) and belongs to at least one

organizational unit. Agent can have multiple roles, but must have at least one role. The same role can be assigned to a larger number of agents. Another concept represents communication links, which are based on a specific protocol and used for communication between agents. Responsibilities and obligations are specified for each role. The model allows the modelling of hierarchical structure of agents, which specifies whether the agent has a superior agent and whether the agent has subordinate agents. This is represented by relationship which agent has to itself. Enterprise ontology introduces concepts of business entity, organization and organizational unit. Common for business entities and organizational units is that both have defined objectives which seek to achieve, using certain organizational rules. The difference is explained in Enterprise Ontology: business entities have rights and responsibilities to other stakeholders inside business, while organizational units have rights and responsibilities only to other organizational units. Figure 3 shows described sub-model with concepts of organization.

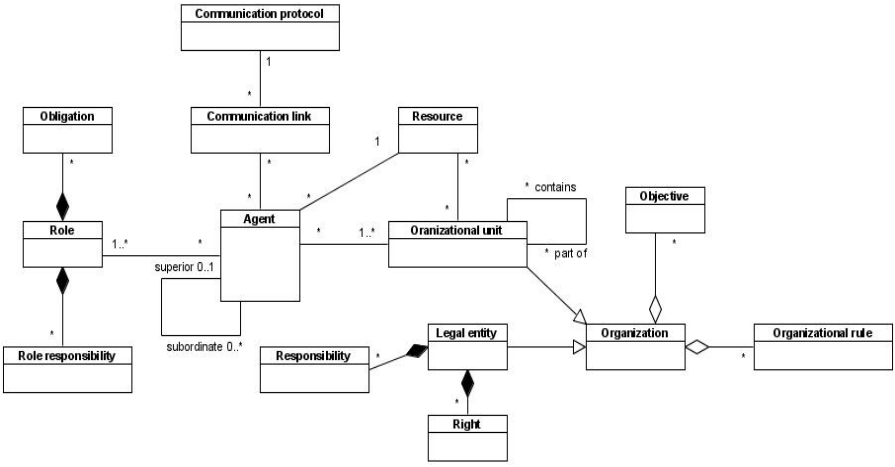

Fig. 3. Organizational concept sub-model

5 Conclusion

Ongoing development of supply chains, which has been conducted for decades, has contributed today's environment, in which partners collaboration within the supply chain can be characterized as a "network of enterprises". With the increasing number of participants, the amount of information that users share also increases, as well as number of different syntax they use and number of different ways they can interpret same semantic.

Over time need for knowledge management was perceived, and necessity for standardizing and presenting knowledge to users in a clear and intuitive way. That was one of additional factors for speeding up development of reference models for appropriate domains. Number of authors [2,3,4] [7] [13] has opinion that SCOR model was imposed as one of the leading models in the area of SCM. That model

offered to participants within the supply chain easily accessible knowledge they were requiring. However, this has opened the problem of interpretation of that knowledge. This type of reference models, including SCOR, does not have a precise way to define the semantics, because their main purpose is to make understanding of certain domain easy. Ontologies, as a way of organizing and managing knowledge, imposes as a logical choice, according to certain authors [2,3,4] [7] [13]. Growth in the number of participants in the collaboration and rapid technological development leading to the formation of a large number of ontologies, which stakeholders use to communicate. Differences in the definition of ontology concepts, their connections and increasing semantic gap between users seriously jeopardize system interoperability. Negative effects of inconsistencies in the conceptualization can be reduced by additional mapping, transformation or integration of appropriate ontologies.

In this paper we analyzed three commonly used ontologies, and it was found that despite some degree of difference, most of the concepts, along with certain adjustments and modifications can be use as a base for integration. Ontologies are very complex, and beside concepts it is also important to use descriptive logic in order to completely specify their utilizations. Integrated model with concepts of all tree commonly used ontologies was developed. In purpose of visual clarity, this model was decomposed in three sub-models (resources, activities and organizations). This model (and its sub-models) should reduce the gap between the three analyzed ontologies suitable for use in the SCOR model and also shows how much similarity exists among concepts of these tree ontologies.

Future work will consist of further integration of these ontologies and development of descriptive logic for them. Descriptive logic provides well-defined semantics and structured reasoning approach and because of that imposes as an intuitive solution for ontology languages. Implementation of descriptive logic could make benefits in further ontologies semantic standardization (defining unique way for interpreting data) and integration (making narrow differences between ontologies concepts). However, that should be only the first step in the future work. After that, work on tool which would be based on presented model and which would support descriptive logic is planned. That kind of tool could make ontologies integration more useable in practice and also could provide base for further ontologies evolution. Mentioned approach and future work should provide higher level of interoperability and further improvements for SCOR model and SCM.

References

1. Fawcett, S., Magnan, G., McCarter, M.: The Effect of People on the Supply Chain World: Some Overlooked Issues. Human Systems Management 4(24), 197–208 (2005)
2. Huan, S.H., Sheoran, S.K., Wang, G.: A review and analysis of supply chain operations reference (SCOR) model. Supply Chain Management: An International Journal 9(1), 23–29 (2004)
3. Zdravkovic, M., Panetto, H., Trajanovic, M., Aubry, A.: An Approach for Formalizing the Supply Chain Operations. Enterprise Information System 5(4), 401–421 (2011)

4. Zdravkovic, M., Trajanovic, M.: Integrated product ontologies for inter-organizational networks. Computer Science and Information Systems 6(2), 29–46 (2009)
5. Ren, C.R., Dong, J., Ding, H.W., Wang, W.: A SCOR-based framework for supply chain performance management. In: 2006 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI 2006), pp. 1130–1135. IEEE Press, New York (2006)
6. Zhou, H.G., Benton, W.C., Schilling, D.A., Milligan, G.W.: Supply Chain Integration and the SCOR Model. Journal of Business Logistics 32(4), 332–344 (2011)
7. Guarino, N., Giaretta, P.: Ontologies and knowledge bases – towards a terminological clarification. In: Towards Very Large Knowledge Bases, pp. 25–32. IOS Press, Amsterdam (1995)
8. Fox, M., Barbuceanu, M., Gruninger, M., Lin, J.: An Organization Ontology for Enterprise Modelling. Computers in Industry 29(1), 123–134 (1996)
9. Uschold, M., King, M., Moralee, S., Zorgios, Y.: The Enterprise Ontology. The Knowledge Engineering Review 13(1), 31–89 (1998)
10. Madni, A., Lin, W., Madni, C.: IDEONTM: An Extensible Ontology for Designing, Intergrating and Managing Collaborative Distributed Enterprises. Systems Engineering 4(1), 35–48 (2001)
11. Knowledge Interchange Format (January 2014), `http://logic.stanford.edu/kif/dpans.html`
12. Madni, A., Lin, W.: ISTI distributed enterprise ontology (IDEONTM): An overview. ISTI White Paper ISTI-WP-5/97-1, Santa Monica (1998)
13. Lu, Y., Panetto, H., Ni, Y., Gu, X.: Ontology Alignement for Networked Enterprises Information Systems Interoperability in Supply Chain Management. International Journal of Computer Integrated Manufacturing 26(1-2), 140–151 (2013)
14. Rajabi, Z., Minaei, B., Seyyedi, M.A.: Enterprise Architecture Development Based on Enterprise Ontology. Journal of Theoretical and Applied Electronic Commerce Research 8(2), 85–95 (2013)
15. Huang, S.H., Sheoran, S.K., Keskar, H.: Computer-assisted supply chain configuration based on supply chain operations reference (SCOR) model. Computers & Industrial Engineering 48(2), 377–394 (2005)
16. Millet, P.A., Schmitt, P., Botta-Genoulaz, V.: The SCOR model for the alignment of business processes and information systems. Enterprise Information Systems 3(4), 393–407 (2009)
17. Wang, W.Y.C., Chan, H.K., Pauleen, D.J.: Aligning business process reengineering in implementing global supply chain systems by the SCOR model. International Journal of Production Research 48(19), 5647–5669 (2010)
18. Alvarado, K., Rabelo, L.: A roadmap for the Supply Chain Operations Reference Model (SCOR). In: 25th National Conference of the American-Society-for-Engineering-Management, pp. 433–442. ASEM Press, Alexandria (2004)
19. Irfan, D., Xu, X.F., Chun, D.S.: A SCOR reference model of the supply chain management system inan enterprise. International Arab Journal of Information Technology 5(3), 288–295 (2008)
20. Medini, K., Bourey, J.P.: SCOR-based enterprise architecture methodology. International Journal of Computer Integrated Manufacturing 25(7), 594–607 (2012)
21. Orgun, B., Dras, M., Nayak, A., James, G.: Approaches for semantic interoperability between domain ontologies. Expert Systems 25(3), 179–196 (2008)

A Tracking System for the Recognition of Long Term Events in Surveillance Videos

İlkay Ulusoy[1,*], Yousef Rezaeitabar[2], and Nihan Çiçekli[3]

[1] Department of Electrical and Electronics Eng., METU, Ankara, Turkey
ilkay@metu.edu.tr
[2] Biomedical Engineering Department, METU, Ankara, Turkey
[3] Computer Engineering Department, METU, Ankara, Turkey

Abstract. Event recognition requires long term tracking of objects in the surveillance videos. However, longterm tracking typically suffers from a lack of robustness in most realistic scenarios, due to illumination changes, cluttered background, occlusions, appearance changes, etc. Therefore, most of the event recognition methods omit long-term tracking procedure, so that they can describe and recognize only short term events such as walking, running, sitting, falling, kicking, etc. To circumvent this drawback, a system is proposed in this paper, which fuses the information acquired from the foreground mask and pixel color of the frames whenever needed to handle occlusion and to achieve long term object detection, tracking and labeling. By this system, the event recognizer becomes able to discriminate long lasting events such as purse snatching, fighting, meeting, unwanted person around a car, etc. Many videos of various events and scenarios are investigated based on the spatio-temporal organization of the objects along the time and generic solutions, which are applicable for most of the problematic cases in all types of the videos and scenarios, are proposed. Finally, results are presented for well-known data sets and our data set, all of which include long term events. We observed that the performance of long term event recognition is improved with the proposed system.

Keywords: Surveillance, long term event, tracking.

1 Introduction

Surveillance cameras play an important role in public security. Therefore, there has been a considerable interest on the development of automatic video surveillance applications, which detect suspicious events online and alert immediately. Event detection techniques require two main levels of processing. In the first level, objects (people, car, package, etc.) are detected and their spatio-temporal relations are tracked using video processing techniques. In the second level, using the output obtained from the first level, possible events are searched by using event models, for which the rules have been defined or learned a priori.

* Corresponding author.

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_9

The first level processing of surveillance videos consists of many different components, each of which requires complex operations. These components are background modeling and foreground detection, motion detection, object detection, object recognition and tracking.

Event recognition has been studied widely in the literature [1,2]. In most of them, only short term events (or activities) such as walking, running, bending, falling, kicking [3-5], unattended luggage and fighting [6] are studied. Most of them present only one event type and donot consider complex scenes with many moving objects and activities over a long period of time. The performance of the long term event recognizers highly depends on the performance of the object tracking and labeling [7]. Long term tracking typically suffers from a lack of robustness in most realistic scenarios, due to illumination changes, cluttered background, occlusions, appearance changes, etc. There are some solutions for these problems in the recent literature [8,9]. However, all these solutions require multiple cameras or detailed information about the camera parameters and the scene.

In this paper, we propose a generic long term tracking and labeling framework which could be applied for long term event recognition in surveillance videos of various contents. We examine three data sets which include various long term events: PETS2006 [10], CANTATA [11] and our own dataset. Our dataset includes surveillance videos containing different events such as meeting of two or more people, fighting, purse snatching, left object, unwanted people around a car. The data sets contain both indoor and outdoor videos.

2 The Proposed Framework

First of all, object detection is applied on the input video without any preprocessing. Background modeling techniques, Mixture of Gaussian (MOG) [12] and Codebook [13], are applied and comapred. The MOG is found to be more successful and used in this study. Then, for human recognition in the foregound regions, the Deformable Parts Model [14] is used since it has been accepted as having the highest performance by the recent literature [15]. The other objects such as bag and car, are recognized by using the size information of their bounding boxes. Finally, Kalman filter method is used for labeling and tracking the detected objects.

However, if these methods are applied one after the other directly, which is called as the standard approach in this paper, the long term and robust tracking cannot be achieved. The problems occur with the standard approach and the proposed counter solutions which are generic and could be applied to all videos in this study are explained below.

2.1 Merging and Splitting of the Connected Components (Objects) in the Foreground Mask

Each connected component in the foreground mask is assumed to be a single object. The location and size of each object are computed from the bounding boxes of the foreground regions. However, when the objects come very close or they occlude each

other, their foreground masks merge. In order to continue labeling of all objects inside the mask, the merging should be detected before it happens. A similar problem happens when a connected component splits, i.e., objects move away from each other. Sample cases of merging and splitting are shown in Fig 1. Therefore, merging and splitting conditions should be identified. In this study, if a connected component in the current frame's foreground overlaps with a portion of more than one connected components in the previous frame's foreground, this is labeled as merging. To make things easier, instead of the foreground masks, the bounding boxes of the masks can be compared also. The portion is a parameter of the proposed framework and decided heuristically but kept fixed for all videos. A similar procedure is used for split detection also.

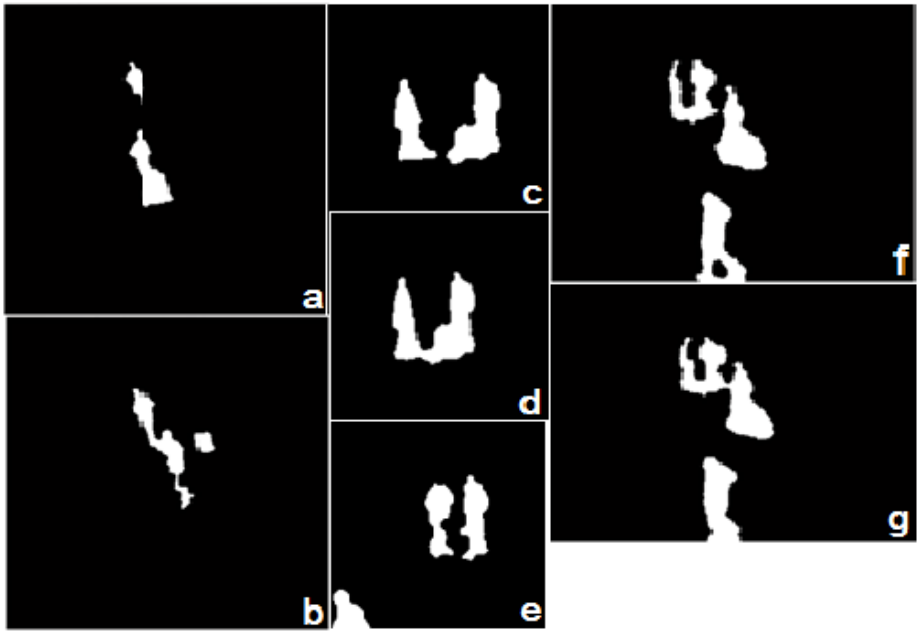

Fig. 1. (a) Two separate masks (one belongs to a person, the other one belongs to a person with a package). (b) Two separate masks (one belongs to two people, the other one belongs to a package). (c) Two separate masks (one belongs to a person, the other one belongs to a person with a package). (d) One mask (two persons and a package). (e) Three masks (each belong to a person). (f) Three masks (one belongs to two persons, the other one belongs to a person and a package, the third one belongs to a person). (g) Two masks (one belongs to three people and the other one belongs to a person).

The detection of merging and splitting can be more complicated when the merging objects are different from the splitting objects or a foreground mask, which has not been merged earlier, could split. Some examples of these problematic cases are shown in Fig 1.a and b, where two masks merge (one person mask and one person with a package mask) and then one new mask splits (one mask with two people and a mask with package only). Also in Fig 2.f and g, an already merged mask (mask with two

people) merges with another one (mask of a person). In these cases, the number of merged masks is equal to the number of split masks, but the contents of the merged and split masks are different. Thus, merging and splitting detection is applied to each frame and each foreground mask. If merging occurs, the merged objects are tagged and tracked together. If splitting occurs after merging, since objects are tracked during merged frames, they are continued to be tracked separately after splitting. If splitting occurs for a connected component, which was not detected as merged previously, split regions are handled separately for object recognition and tracking.

In order to find the actual coordinates of the objects inside a merged region, similarity checking is applied to the objects and the occlusion mask. The image of every single object and the color histogram of this image are saved and updated in each frame. Once occlusion happens, the stored information just before merging is used to find the coordinates of the objects inside the merged region. An example is shown in Fig 2.a,b,c. In the case of full occlusion, the locations of these objects are estimated by the Kalman tracking.

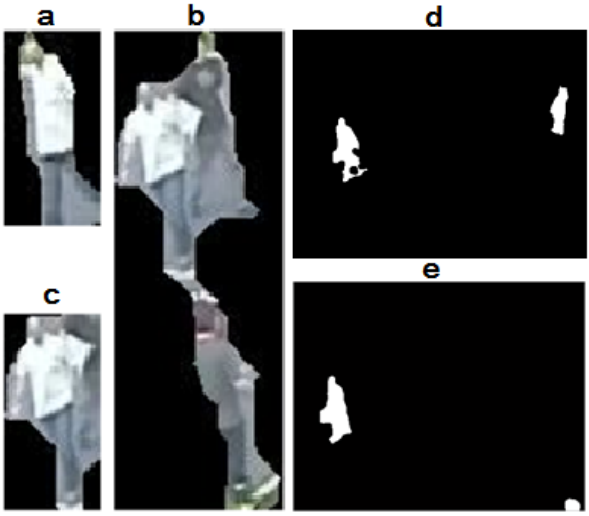

Fig. 2. (a) Appearance of the object before occlusion. (b) Appearance during merging/occlusion. (c) Detected object in the occluded/merged mask. (d) Two objects in the scene. (e) The object in the right side cannot be detected and another object enters the scene. So the number of objects in the scene stays unchanged.

2.2 Labeling of the Objects When They Are Occluded or Entering and Leaving the Scene

When an object disappears from the scene it may be occluded. The Kalman filtering is applied to estimate the new location of the object. We are able to catch the object in the estimated location if the object disappears for only a few frames. Although the idea of labeling the objects by tracking is simple, assigning the correct label in the crowded scenes is always problematic. Some objects may not be detected in the foreground not only because of the occlusion problem, but also because of the problems

related to background modeling. When another object enters the scene in the same frame, where some previously existing objects cannot be detected although they are in the scene, the number of objects in the scene stays unchanged and this may cause to think that the same objects are in the scene. For example, although the second person on the right side of the frame is detected (Fig 2.d), it cannot be detected in the following frame (Fig 2.e) but another person enters the scene in this second frame. The same problem may also happen when an object exits the scene. Similarly, an object may leave the scene and then enter some time later.

So, the objects occluded or entering or leaving the scene should be correctly labeled. If an object leaves the scene, its label and detailed information are stored to be checked in all of the future frames. When a new object enters the scene, the similarity check is performed to find out if this object was present in the video previously.

2.3 Noise in Long Term Tracking

Objects may not be detected or tracked continuously. These cases are considered as noise for tracking. In order to remove such noise, the number of frames in which an object appears is stored and it is compared with a threshold value. This value is decided heuristically but kept the same for all tests. If the object appears consistently in the scene, it is considered as a real object. On the other hand, if the object appears in a few frames and then disappears, it is considered as noise and it is neither labeled nor tracked. Also, when an object enters the scene for the first time, it is not being tracked for some number of frames. The tracking of the object starts only when the number of frames it appears is bigger than a threshold. This threshold is also decided heuristically but kept the same for all tests.

3 Experiments and Results

We used three data sets in our experiments: PETS2006, CANTATA and our data set. The summary of the video contents is given in Table 1. In this study, it is very hard to evaluate the performance of the proposed system, since we are not proposing a single algorithm to solve a single problem but we are proposing a framework which could be used for long term tracking.

We compared our system with the standard approach. We run our system and the standard approach for the same videos and showed these to 20 people and wanted them to evaluate the performances. All of the subjects agreed on the better performance of our proposed system in labeling and tracking. As another testing, we used the outputs of our proposed system and the standard approach with an event recognizer [7,16]. Event recognizer cannot detect events with the outputs obtained from the application of the standard approach but performed well with the outputs of our proposed system. As the final test, we compared the continuity of tracking. The ratio of the duration of the correctly tracked video sections to the whole video duration is used as the performance metric. The metric is computed for all videos and given in the last column of Table 1. The average is found to be 94%. This shows that the objects are detected, tracked and labeled nearly for all frames throughout the video even if they get merged, occluded or split. This could not be possible with the standard approach.

Table 1. Summary of video contents and long term tracking performances for the videos

Dataset	Event	Event Duration (s)	Percentage of correct tracking and labeling (%)
Our dataset	Meeting of two or more person	210	96
	Fighting	330	95
	Bag snatching	110	96
	Unwanted person around a car	60	90
	People leaving a package	480	98
Cantata	People leaving a bag in a parking lot	1680	96
PETS 2006	People moving around in an underground station	210	88
		Average:	94

An example run for a video piece from CANTATA database is given in Fig 3. Although packages and people get occluded, merged and split, our framework correctly labels the three main objects in the scene (two people and a package). However, the standard approach gives more than three labels and different labels for the objects.

Fig. 3. Labeling and tracking results of our proposed system (top row) and the standard method (bottom row)

Another example of our own database is shown in Fig 4 where the person and the car are tracked and labeled correctly although they get occluded, merged and split. The same label is assigned to the person all the time.

Fig. 4. Tracking and labeling for the event of unwanted man around a car

4 Conclusion

In this paper, we present a framework for longterm object tracking and labeling so that long term spatio-temporal information about every object in the scene can be provided to the event recognizer for surveillance systems. We tested our system on various data sets including various events and scenarios. With the proposed framework, better performance can be achieved and less noisy outputs can be obtained. To the best of our knowledge, we are not aware of a system which includes solutions to all possible problems in one integrated framework and which can be used generically for various events and scenarios.

Acknowledgement. This work is partially supported by the Ministry of Science, Industry and Technology of Turkey and by Havelsan Inc. under Grant SANTEZ 00896.STZ.2011-1.

References

1. Brendel, W., Fern, A., Todorovic, S.: Probabilistic Event Logic for Interval-Based Event Recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2011)

2. Ghanem, N., DeMenthon, D., Doermann, D., Davis, L.: Representation and Recognition of Events in Surveillance Video Using Petri Nets. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (2004)
3. AbdelKader, M.F., Almageed, W.A., Srivastava, A., Chellappa, R.: Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds. Computer Vision and Image Understanding, 439–455 (2011)
4. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: a review. ACM Computing Surveys 43(3) (2011)
5. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS 2005), Beijing, China, pp. 65–72 (2005)
6. Blunsden, S., Andrade, E., Fisher, R.: Non parametric classification of human interaction. In: 3rd Iberian Conf. Pattern Recog. Image Anal., pp. 347–354 (2007)
7. Kardas, K., Cicekli, N.K., Ulusoy, I.: Learning Complex Event Models Using Markov Logic Networks. In: 3rd International Workshop on Advances in Automated Multimedia Surveillance for Public Safety (AAMS-PS) ICME 2013, San Jose, USA (2013)
8. Grimson, W.E.L., Stauffer, C., Romano, R., Lee, L.: Using adaptive tracking to classify and monitor activities in a site. In: Computer Vision and Pattern Recognition (CVPR), Santa Barbara, CA, pp. 246–252 (1998)
9. Remagnino, P., Baumberg, A., Grove, T., Hogg, D.C., Tan, T., Worrall, A., Baker, K.: An integrated traffic and pedestrian model- based vision system. In: British Machine Vision Conference (BMVC), Essex, UK, pp. 380–389 (1997)
10. PETS (2006), http://www.cvg.rdg.ac.uk/PETS2006/data.html
11. CANTATA dataset, http://www.hitechprojects.com/euprojects/cantata/index.htm
12. Piccardi, M.: Background subtraction techniques: A review. The ARC Centre of Excellence for Autonomous Systems (CAS), Faculty of Engineering, UTS (2004)
13. Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L.: Background modeling and subtraction by codebook construction. In: International Conference on Image Processing, vol. 5, pp. 3061–3064 (2004)
14. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9) (2010)
15. PASCAL visual object challenge, http://pascallin.ecs.soton.ac.uk/challenges/VOC/
16. Onal, I., Kardas, K., Tabar, Y.R., Bayram, U., Cicekli, N.K., Ulusoy, I.: A Framework for Detecting Complex Events in Surveillance Videos. In: 3rd International Workshop on Advances in Automated Multimedia Surveillance for Public Safety (AAMS-PS), ICME 2013, San Jose, USA (2013)

Simulation of L2 Cache Separation Impact in CPU Performance

Erion Çano

Department of Computer Engineering,
Polytechnic University of Tirana, Albania
erion.cano8@gmail.com

Abstract. Cache memory performance is very important in the overall performance of modern CPUs. One of the many techniques used to improve it is the split of on-chip cache memory in two separate Instruction and Data caches. The current CPU organizations usually have per core separate L1 caches and unified L2 caches. This paper presents the results of simulating different CPU organizations with unified and separate L2 Instruction and Data caches using Marss-x86, a Cycle-Accurate full system simulator. The results indicate that separating the L2 cache memory provides higher overall CPU IPC. The highest improvement is 3% and is achieved in a quad-core CPU model with shared L3 cache. Analyzing the hardware costs and complications of separating L2 cache might be an interesting future work direction.

Keywords: Cache Organization, CPU Performance, L2 Cache Models, CPU IPC.

1 Introduction

The continuous improvements offered by silicon technology make it possible to place more transistors on a single chip. There are currently many commodity PCs equipped with quad-core CPUs which have considerably large L3 cache memories integrated inside. As the performance gap between CPU and main memory has been getting larger, the importance of the fast on-chip cache memories continues to rise. The first on-chip caches were small, unified and one-level only. The current chips contain large L3 caches with different organizations.

One of the simplest techniques used to improve caching and CPU overall performance is the usage of dedicated and separate cache memories for Instructions and Data. Separate L1 I and L1 D caches were a reality since more than a decade ago (since Intel® Pentium® processor models). The most important reason of using separate L1 caches is the possibility to access them in parallel, thus attaining a higher bandwidth. The drawback of having two caches is the higher miss rate in each of them as their size is lower. Other different pros and cons of the two cache organizations are discussed in Section 2.

In this article I simulate and evaluate the overall CPU performance improvements of different possible separate Instruction and Data L2 cache organizations. To have realistic results I consider only bandwidth (number of cache connections) and miss

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_10

rate (cache size) and keep the other cache specifications identical. To model and simulate the different cache organization CPUs, Marss-x86, a Cycle-Accurate full system simulator presented in [1] is used. Besides providing basic modules of processing cores, caches and memory, Marss-x86 is also extensible and permits quick integration of other simulation modules. The simulation results show that the IPC increase of splitting L2 cache into two equally sized instruction and data caches, no matter how significant, is always positive and ranges from 0.4 to 3 %.

The rest of the paper is organized as follows: Section 2 discusses pros and cons of having unified and separated Instruction and Data caches. Also it summarizes related work about size impact on cache access time and power consumption. Section 3 shortly presents Marss-x86 simulator, its structure and the features it offers. The simulations' specifications and the different machine models I have used are described in Section 4. Section 5 presents the simulation results while Section 6 concludes and shows some possible future work directions.

2 Unified vs. Split Cache Memory Organizations

Besides the traditional unified caches, the other very common design is having separate caches for instructions and data. There are certainly different pros and cons of splitting a cache memory into two smaller caches. The most important advantage of splitting the cache memory is the increase in bandwidth that results. Modern processors can read data from the instruction cache and the data cache simultaneously in a single cache memory cycle. Having two separate Instruction and Data caches and accessing them simultaneously offers the possibility to potentially double the bandwidth [2]. Also the Instruction cache does not need to manage a processor store. Having it separate from the Data cache makes possible to simplify its design. Replacement policy may also result more effective. One can be direct-mapped and the other can be highly associative.

This architecture, also known as Harvard architecture, has the drawback that it needs two complete sets of address and data lines, one to each of the caches. There are also many hardware complications and costs of having to address and index two separate caches instead of one. The most significant drawback of splitting a cache in two smaller caches is the increase in miss rate that will result in each of them as a consequence of the lower (half in case of symmetric split) capacity that will result. An important advantage of a unified cache is the balancing it offers in terms of instruction/data words. If the ratio of data to instruction words changes during the runtime it is adapted to by the replacement policy. This is also rare as the capacity is higher than in case of separate caches. In case of two separate Instruction and Data caches, having too many Instruction or Data words to cache will result in filling up one of the caches. No adaption is possible [3].

In general accessing large memories of any kind takes more time than accessing smaller memories. The circuit level higher access times of larger caches have been also analyzed at [4]. The authors report a proportional increase in access time as a function of cache size for both direct mapped and set associative caches. The authors attribute the higher delays to the circuit comparisons and tag matching. In [5] the authors show that in general it takes less time to access direct-mapped caches than set

associative caches. What is more important is the quasi-linear rise of access time with the increase of cache size. They also conclude that there is a considerable decrease in energy consumption if the cache is partitioned into several banks.

Other articles like [6] and [7] evaluate the performance of different Temporal/Spatial split of cache models. They report performance improvements due to the better exploitation of the locality patterns in the code of different applications. In [8] the authors propose an algorithm to reduce cache interference among different simultaneous processes by dynamically (and of course logically) partitioning the last level cache among those competing processes. They report significant cache performance improvements with minimal hardware overhead for the modifications. Another advantage of having smaller separate caches is the reduction in power consumption. In [9] the author presents a scheme named cooperative partitioning to logically divide the LLC ways between the competing cores of the CMP. This scheme uses a shadow tag to monitor the cache requirements of each application and makes a proportional partition of the cache blocks between the cores running the applications. He reports a reduction of 67% and 25 % reduction in dynamic and static energy consumption for dual-core systems.

The purpose of this work is to assess any overall IPC improvement gained from the split of L2 cache into two equally sized instruction and data caches. One large L2 cache provides low bandwidth because it cannot be accessed in parallel by Instructions L1 and Data L1. However it has low miss rate as it is large and can hold many blocks. Two smaller L2 caches (Instructions L2 and Data L2) provide higher bandwidth as they are accessed in parallel. However the miss rate is higher as they have smaller capacities. Being the dominant factors that influence cache performance, bandwidth and size are the two parameters I consider. The others cache characteristics are kept identical in every set of comparative simulations.

3 Brief Introduction of Marss-x86

Marss-x86 is an open source simulator for Cycle-Accurate simulations of multicore CPU configurations. From the different functionalities and features it provides the following are the most important:

- It makes use of different Cycle-Accurate simulation models for out-of-order and in-order single core and multicore CPUs implementing the x86 ISA.
- It supports switching between the Cycle-Accurate simulation mode and the x86 emulation mode of QEMU, an emerging emulator.
- Being based on QEMU, Marss-x86 can boot and execute unmodified operating systems, applications (i.e. benchmarks) and library binaries.
- It includes models of memory hierarchies for single-core and multicore chips and realizes 200 – 400 kilo instructions per second simulations in Cycle-Accurate simulation mode.

Marss-x86 reuses many components of QEMU like emulated IO devices, user interfaces etc [10]. It also provides a MIMO based interface which allows communications between the VM's software and the simulated/emulated hardware components.

Using this interface programs running in VM can send control signals to the simulator. Marss-x86 allows users to simulate different hybrid CPU configurations consisting of in-order and out-of-order cores of the same chip. It also implements Cycle-Accurate simulation for superscalar pipelines. Marss-x86 framework is extensible and permits quick integration of other simulation modules like DRAMsim, DiskSim and FlashSim. This provides the possibility for accurate and overall system simulations.

The configuration files are written in YAML, a human friendly data serialization language [11]. A typical YAML configuration (modeling) file contains three types of modules: Cores, caches and memory controllers. For each type of module Marss-x86 provides at least one basic module and gives the possibility to create custom modules based on the desired specifications. The YAML configuration file of a dual-core machine with L3 cache is given below:

```
machine:
  dual_l3_1:
    description: Dual Core CPU with L3 cache - configuration 1
    min_contexts: 2
    max_contexts: 2
    cores:
      - type: ooo
        name_prefix: ooo_
        option:
            threads: 1
    caches:
      - type: l1_mesi_32K
        name_prefix: L1_I_
        insts: $NUMCORES
        option:
            private: true
      - type: l1_mesi_32K
        name_prefix: L1_D_
        insts: $NUMCORES
        option:
            private: true
      - type: l2_mesi_256K
        name_prefix: L2_
        option:
            private: true
            last_private: true
        insts: $NUMCORES
      - type: l3_wb_4M
        name_prefix: L3_
        insts: 1
    memory:
      - type: dram_cont
        name_prefix: MEM_
        insts: 1 # Single DRAM controller
        option:
            latency: 90 # In nano seconds
    interconnects:
      - type: p2p
        connections:
            - core_$: I
              L1_I_$: UPPER
            - core_$: D
              L1_D_$: UPPER
            - L1_I_$: LOWER
              L2_$: UPPER
            - L1_D_$: LOWER
              L2_$: UPPER2
            - L3_0: LOWER
              MEM_0: UPPER
      - type: split_bus
        connections:
            - L2_*: LOWER
              L3_0: UPPER
```

This machine model is described in Section 4.3. After executing a simulation Marss-x86 gives basic results such as IPC. Many other specific simulation results can be obtained by running different statistic collection Python scripts that are provided.

4 Simulation Models

To have a reliable assessment of L2 cache memory organization impact in the overall IPC of the machine I used different CPU models such as single core with L1 and L2 caches, dual-core with L1 and L2 caches, dual-core with L1, L2 and L3 caches, quad-core with L1 and L2 caches and quad-core with L1, L2 and L3 caches. Marss-x86 reads the configuration files of these models which are written in YAML format. It generates and compiles the corresponding C++ code which is than executed. The following subsections present the simulation environment, specifications and details of the CPU models that are used.

4.1 Simulation Environment and Specifications

I used a quad-core 2.79 GHz Intel® Xeon® E5-1603 CPU equipped physical machine running Ubuntu 13.04 with kernel version 3.8.0-35-generic. I installed Marss-x86 which uses QEMU to boot and run a disk image over the simulated machines. The emulated system and application consists of Linux kernel 2.6.31.4 and Radix Sort C-implemented algorithm which is used to exercise the CPU models by sorting millions of randomly generated integer numbers. In every simulation model I used different number (1, 2 or 4) of the default Marss-x86 Out Of Order CPU cores each of which has the following specifications:

Table 1. Emulated CPU Parameters

Property	Value
freq:	2793000000
threads:	1
phys_reg_files:	4
phys_reg_file_int_size:	256
phys_reg_file_fp_size:	256
dispatch_width:	4
issue_width:	4
writeback_width:	4
commit_width:	4

Cache memory specifications depend on cache level and size. For L1 I used write-back caches for the single core model and MESI coherent caches for the multicore models. For L2 I used write-back caches if there is not a L3 cache and MESI coherent L2 if there is a shared L3 cache. L3 caches are shared and write-back in every model.

Table 2. Memory Specifications

Property	Value
RAM_size	2147483648 B (2 GB)
number_of_banks	64
latency	260 cycles
latency_ns	90 ns

The common cache specifications in every model are 64 B for line size, 2 read ports and 2 write ports. The rest of cache specifications are given in the following sections. I also used the following memory specifications in every simulation:

4.2 Models of Single-Core CPUs

In this simulation two organizations of a single core CPU with 2 levels of on-chip cache memory are compared. The first model is a very common single-core organization with shared L2 cache (i.e. Intel® Pentium M® 740 family has the same basic structure). The second model is uncommon having separate L2 I and L2 D second level caches. In the first model there are two p2p connections between the cache memories, specifically L1 I ⟷ L2 and L1 D ⟷ L2, and a single p2p connection between L2 and the main memory. In the second model the cache memory connections are L1 I ⟷ L2 I and L1 D ⟷ L2 D. There are also 2 p2p connections between the 2 L2 caches and the main memory. The goal is to assess the IPC of splitting the L2 cache in two halves, L2 I and L2 D and having the possibility to make parallel accesses between level 1 and level 2 caches. Level 1 cache memories are all identical in both models having 32 KB size, 64 sets, 8-way associativity and 2 cycles latency. Level 2 cache memories differ only in size having 2 MB vs. 1 MB L2 I + 1 MB L2 D, 4096 sets vs. 2048 + 2048 sets, 8-way associativity and 18 cycles latency. Cache memories are all write back.

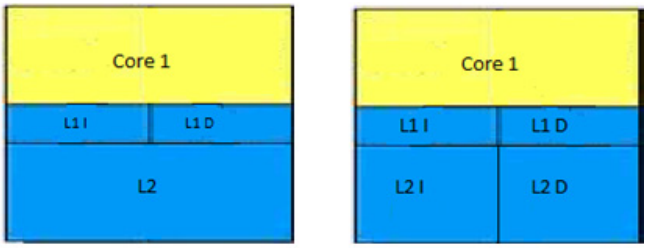

Fig. 1. Single unified L2 vs. Single separate L2

4.3 Models of Dual-Core CPUs

In this simulation I compare two organizations of a dual-core CPU with 2 levels of on-chip cache memory. The first model is a very common dual-core organization with shared L2 cache (i.e. Intel® Core Duo® L2500 family has the same basic structure).

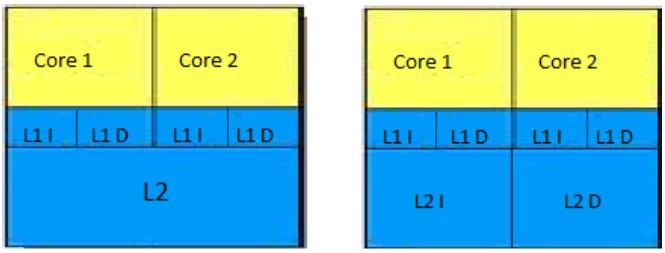

Fig. 2. Dual unified L2 vs. Dual separate L2

The second model is uncommon having shared (not per core) separate L2 I and L2 D second level caches. In the first model there are 4 split bus connections between the four L1 cache memories and L2. There is also and a p2p connection between L2 and the main memory. In the second model there are 4 split bus connections, specifically Core 1 L1 I ⟷ L2 I, Core 2 L1 ⟷ L2 I, Core 1 L1 D ⟷ L2 D and Core 2 L1 D ⟷ L2 D. There are also 2 p2p connections between the two L2 caches and the main memory. The goal is the same, the evaluation of the IPC improvement of splitting the L2 cache in two halves, L2 I and L2 D. L1 cache memories are MESI coherent and have identical specifications in both models (and in the models that follow). They have 32 KB size, 64 sets, 8-way associativity and 2 cycles latency. Level 2 cache memories differ only in size having 2 MB vs. 1 MB L2 I + 1 MB L2 D, 4096 sets vs. 2048 + 2048 sets, 8-way associativity and 18 cycles latency.

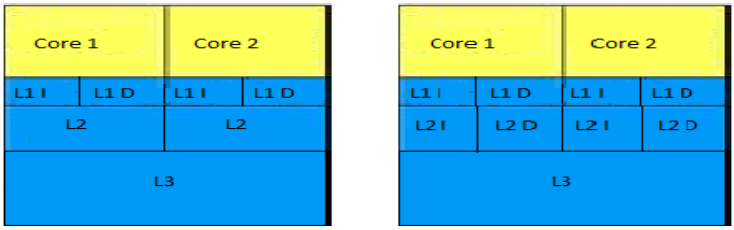

Fig. 3. Dual unified L2 with L3 vs. Dual separate L2 with L3

In this simulation I compare two organizations of a dual-core CPU with 3 levels of on-chip cache memory. The first model is a very common dual-core organization with per core L2 caches and shared L3 cache (i.e. Intel® Xeon® W3505 family). The second model is uncommon with separate per core L2 I and L2 D caches and shared L3. In the first model there are 4 p2p connections between the 4 L1 caches and the two L2 caches (Core 1 L1 I ⟷ Core 1 L2, Core 1 L1 D ⟷ Core 1 L2 , Core 2 L1 I ⟷ Core 2 L2, Core 2 L1 D ⟷ Core 2 L2). There are also 2 split bus connections, Core 1 L2 ⟷ L3 and Core 2 L2 ⟷ L3 and the p2p connection between L3 cache and the main memory. In the second model there are 4 p2p connections between the corresponding L1 and L2 caches. There are also 4 split bus connections between all level 2 caches and the level 3 cache and of course the p2p connection between L3 and the main memory. The YAML configuration file of this machine was presented in

section 3. Level 2 cache memories differ only in size having 256 KB L2 vs. 128 KB L2 I + 128 KB L2 D, 512 vs. 256 + 256 sets, 8-way associativity and 10 cycles latency. The shared L3 caches are identical in both models having 4 MB size, 4096 sets, 16-way associativity and 32 cycles latency.

4.4 Models of Quad-Core CPUs

In this simulation two organizations of a quad-core CPU with two levels of on-chip cache memory are compared. The first model is a common quad-core organization with shared L2 cache (i.e. Intel® Core 2 Extreme® family). The second model is uncommon having shared (not per core) separate L2 I and L2 D second level caches. In the first model there are 8 split bus connections between the level 1 cache memories and the shared L2. There is also and a p2p connection between L2 and the main memory. In the second model there are 4 split bus connections between level 1 instruction caches of the different cores and L2 I. There are also 4 split bus connections between level 1 data caches of the different cores and L2 D. There are of course 2 other p2p connections between the two L2 caches and the main memory. Level 2 cache memories differ only in size having 12 MB vs. 6 MB L2 I + 6 MB L2 D and 12288 sets vs. 6144 + 6144 sets, 16-way associativity and 22 cycles latency.

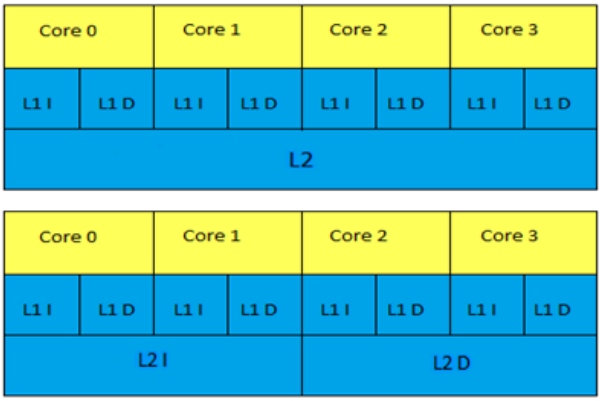

Fig. 4. Quad unified L2 vs. Quad separate L2

In this last simulation I compare 2 organizations of a quad-core CPU with 3 levels of on-chip cache memory. The first model is a quad-core organization with per core L2 caches and shared L3 cache (i.e. Intel® Core® i5 760). The second model has separate per core L2 I and L2 D and shared L3 cache. In the first model there are 8 p2p connections between the L1 caches of the different cores and the 4 L2 caches. There are also 4 split bus connections between the L2 caches and the shared L3. There is also the p2p connection between L3 and the main memory. In the second model there are 8 p2p connections between the corresponding L1 and L2 caches. There are also 8 split bus connections between all L2 caches and the L3 cache and the

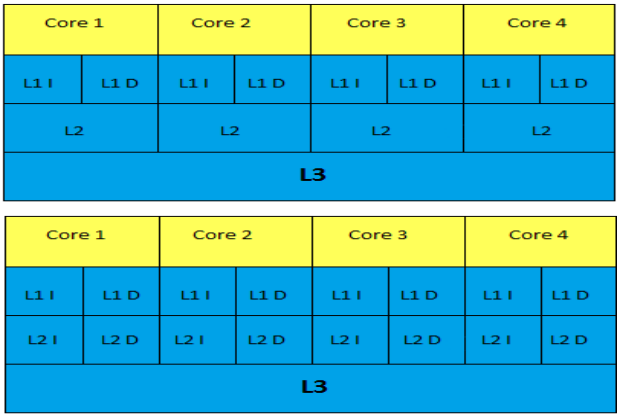

Fig. 5. Quad unified L2 with L3 vs. Quad separate L2 with L3

p2p connection between L3 and the main memory. L2 cache memories differ only in size having 256 KB L2 vs. 128 KB L2 I + 128 KB L2 D, 512 vs. 256 + 256 sets, 8-way associativity and 10 cycles latency. The shared L3 caches are identical in both models having 8 MB size, 8192 sets, 16-way associativity and 34 cycles latency.

5 Evaluation of Results

I ran each simulation (model) 5 times for 30 million cycles each. The average IPCs reported by Marss-x86 were computed and compared. Table 3 presents the IPC difference in % between the average IPCs of the two compared models in every simulation set. First thing to note is the fact that this difference is always positive. There is a slight improvement of 0.4 % from the first Single core model. Dual-Core L2 model also reveals a low improvement of 0.5 %. Splitting L2 cache when it is the last level cache doesn't seem to be beneficial, probably because of the higher miss rates of the smaller L2 I and L2 D. As there is no L3 to serve these requests, they have to go to the main memory which is much slower. The 4th simulation of the quad-core L2 model gives an improvement of 1.2 % which is higher than the first two L2 cache simulations. In this simulations the L2 caches are larger (lower miss rates) and the aggregate bandwidth demand of the 4 cores is higher. Apparently higher bandwidth prevails over higher miss rates.

Table 3. IPC difference between the compared CPU models

Compared Models	IPC Increase (%)
Single unified L2 vs. Single separate L2	0.4
Dual unified L2 vs. Dual separate L2	0.5
Dual unified L2 with L3 vs. Dual separate L2 with L3	2.4
Quad unified L2 vs. Quad separate L2	1.2
Quad unified L2 with L3 vs. Quad separate L2 with L3	3

The 3rd and 5th simulations give encouraging results. Higher miss rate delays of splitting L2 cache are minimized by the larger L3 cache. The higher bandwidth of the parallel p2p connections between the corresponding L1 and L2 caches of each core yields a considerable IPC improvement. The improvement difference between 1st, 2nd and 4th simulation against 3rd and 5th sets suggest that splitting the L2 cache doesn't pay off when there is no L3 cache to compensate the higher L2 miss rates. However it gives considerable improvement when the L3 cache is present. The improvement differences between 2nd and 4th simulation sets and also between 3rd and 5th simulation sets suggests the IPC increase is higher in CMPs with higher number of cores as there is higher bandwidth demand from L2 caches. Being aware of the many hardware costs and complications that the separation of L2 cache implies (which need to be analyzed), having separate per core Instruction and Data L2 caches may be a reality in the imminent many-core CPUs with L4 off-chip caches.

6 Conclusions and Future Work

In this paper I presented a set of simulations for different CPU L2 cache organizations comparing the overall IPC of unified vs. separate Instruction and Data L2 caches. The results indicate that splitting L2 cache into two equally sized instruction and data caches provides a not always considerable but higher IPC. This improvement is merely 0.4 % in a single core L2 organization. It is 0.5 and 1.2 in dual-core and quad-core organizations. The highest improvements are attained in dual-core and quad-core CPUs with a shared L3, respectively 2.4 % and 3 %. The results suggest that a L2 cache split in L2 Instruction cache and L2 Data cache makes sense in many core (i.e. at least four cores) CPUs with at least a L3 cache present on-chip.

Even though the results of the last simulation may seem encouraging there are different hardware costs and complications of having per core separate L2 caches. In [12] the author proposes a logical split of L1 data cache based on run-time data locality analysis. He presents an interesting evaluation of the hardware cost this of organization and concludes that the major problem is the extra space required for storing the extra tags of the two caches. A similar analysis of the extra complications and hardware costs of having separate L2 per core Instruction and Data caches is a tough undertaking and a possible future direction.

References

1. Patel, A., Afram, F., Chen, S., Ghose, K.: MARSS: A Full System Simulator for Multicore x86 CPUs. In: Design Automation Conference (2011)
2. Handy, J.: The Cache Memory Book, p. 63
3. Flynn, J.M.: Computer Architecture: Pipelined and Parallel Processor Design, p. 294
4. Wilton, J.E.S., Jouppi, P.N.: CACTI: An Enhanced Cache Access and Cycle Time Model. IEEE Journal of Solid-state Circuits 31(5), 677–688 (1996)

5. Su, C.L., Despain, M.A.: Cache Design Trade-offs for Power and Performance Optimization: A Case Study. In: ISLPED 1995 Proceedings of the 1995 International Symposium on Low Power Design, pp. 63–68 (1995)
6. Prvulovic, M., Marinov, D., Dimitrijevic, Z., Milutinovic, V.: Split Temporal/Spatial Cache: A Survey and Reevaluation of Performance. IEEE TCCA Newsletters (1999)
7. Naz, A., Rezaei, M., Kavi, K., Sweany, P.: Improving Data Cache Performance with Integrated Use of Split Cache, Victim Cache and Stream Buffers. SIGARCH Computer Architecture News 33(3), 41–48 (2005)
8. Suh, E., Rudolph, L., Devadas, S.: Dynamic Partitioning of Shared Cache Memory. Journal of Supercomputing Architecture (2002)
9. Sundararajan, T.K.: Energy Efficient Cache Architectures for Single, Multi and Many Core Processors. PhD Dissertation (2013)
10. Patel, A., Afram, F., Ghose, K.: MARSS-x86: A QEMU-Based Micro-Architectural and Systems Simulator for x86 Multicore Processors (2011)
11. Machine configuration, `http://marss86.org/~marss86/index.php/`
12. Samdani, G.Q.: A Split Data Cache Organization Based on Run-Time Data Locality Estimation. PhD Dissertation (2000)

Stock Market Trend Prediction Based on the LS-SVM Model Update Algorithm

Ivana Marković[1], Miloš Stojanović[2], Miloš Božić[3], and Jelena Stanković[1]

[1] Faculty of Economics, Trg Kralja Aleksandra Ujedinitelja 11, Niš, Serbia
{ivana.markovic,jelenas}@eknfak.ni.ac.rs
[2] College of Applied Technical Sciences, Aleksandra Medvedeva 20, Niš, Serbia
milosstojanovic10380@yahoo.com
[3] Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, Niš, Serbia
milos1bozic@gmail.com

Abstract. The paper proposes a trend prediction model based on an incremental training set update scheme for the BELEX15 stock market index using the Least Squares Support Vector Machines (LS-SVMs) for classification. The basic idea of this updating approach is to add the most recent data to the training set, as become available. In this way, information from new data is taken into account in model training. The test results indicate that the suggested model is suitable for short-term market trend prediction and that prediction accuracy significantly increases after the training set has been updated with new information.

Keywords: Stock market trend prediction, Least Squares Support Vector Machines (LS-SVMs), Model update.

1 Introduction

The stock market index, as a hypothetical portfolio of selected stocks, is commonly used to measure overall market or particular sector performance [1]. Recent studies [2], indicated that trading strategies guided by predictions regarding the direction of change in the prices could be more effective and could generate a greater yield in comparison to the precise predictions of the level of financial instrument prices. As a result, the world's largest financial markets are now turning to trading in stock market indices more and more often. Consequently, predicting the direction of the movement of the price of financial instruments has now become a current area of academic research.

In numerous studies, the algorithms of machine learning proved to be quite effective in predicting the direction of movement of the value of stock indices and contributed to the increase in yield and reduction in the risk involved in trading. Some of the more frequently adopted methods include the following: Artificial Neural Networks (ANNs) [3], linear and multi-linear regression (LR, MLR) [4], genetic algorithms (GAs) [4], and Support Vector Machines (SVMs) [5]. According to [1], the most widely used methods for stock market trend prediction include approaches based on

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_11

SVMs. In [6], it was further indicated that in most cases the LS-SVMs, and SVMs outperform other machine learning methods, since in theory they do not require any previous a priori assumptions regarding data properties. Moreover, they guarantee an efficient global optimal solution.

As a result of the fact that the financial market is a complex, evolving and dynamic system whose behavior is pronouncedly non-linear, non-stationary and stochastic [5], mining the stock market tendency is a challenging task. Evolving and non-stationary as characteristics imply that the distribution of financial time series changes over a period of time. Thus, to obtain systematically good predictions under such circumstances, it may be necessary to update the underlying models.

The existing stock market trend prediction systems usually focus on several aspects: feature selection, the selection of prediction model and feature evaluation. The problem of model updating, however, has so far not been studied in sufficient detail, particularly in the field of stock market trend prediction. Model updating strategies that correspond to time-evolving systems, including the stock rate index, can usually be undertaken from two perspectives: as incremental learning systems [7, 8, 9, 10], where the respective models are updated online as new instances become available during the training phase, and as batch learning systems [11, 12], where a collection of training instances can be updated prior to model re-training. In this paper, the second model update approach is considered, where the new data over a given time period are added to the initial training set and the respective model is then re-trained. In [11] and [12] similar concepts are presented, but for a different subject matter. To our knowledge, the proposed approach of model updating has so far not been used for stock market trend index prediction.

In this paper LS-SVMs will be used to create a prediction model, but any classification technique is suitable for the application of the proposed model updating algorithm. The problem of stock index trend prediction is modeled as a binary classification problem. Experimental results, benchmarking the standard and updated model, show that prediction accuracy can be increased after updating the initial training set with new available data.

The proposed algorithm offers a systematic approach for model updating based on new instances as they become available.

The rest of this paper is organized as follows: Section 2 presents the basic theory of LS-SVMs for classification. Section 3 presents the used updating methodology. Section 4 gives data set analysis and presents the experimental results. Finally, Section 5 provides the conclusions.

2 Least Squares Support Vector Machines for Classification

The Least Squares Support Vector Machines, proposed by Suykens in [13], includes a set of linear equations which are solved instead of a Quadratic Programming (QP) for classical SVMs. Therefore, LS-SVMs are more time-efficient than standard SVMs, but with lack of sparseness.

Let's study a training group of a total of N examples $T=\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$. In the learning phase, the model is formed based on the known training data ($\boldsymbol{x}_1$, y_1), ($\boldsymbol{x}_2$, y_2), …, ($\boldsymbol{x}_N$, y_N), where $\boldsymbol{x}_i$ are the input vectors, and y_i are the labels of binary classes that were assigned to them. Each input vector consists of numeric features, while $y_i \in \{-1, +1\}$.

According to [13] LS-SVMs for binary classification were defined as follows:

$$\min_{w,b,e} J_{LS}(w,b,e) = \tfrac{1}{2} w^T w + \gamma \tfrac{1}{2} \sum_{k=1}^{N} e_k^2 \tag{1}$$

with the equality conditions:

$$y_k[w^T \varphi(x_k) + b] = 1 - e_k, \quad k = 1,..., N \tag{2}$$

where φ is a non-linear function that maps input vectors in some higher dimensional feature space. The weight vector of the hyper plane is marked by a $\boldsymbol{w}$, while b is the scalar shift, that is, weight threshold. The variable e_k represents the allowed errors of classification, while the parameter γ controls the process, that is, the relationship between the complexity of the model and the accepted error of classification.

After solving the optimization problem defined by (1) and (2), a solution can be found in [13], the function of the separation of LS-SVM classifications is defined as:

$$y(x) = sign\left[\sum_{k=1}^{N} \alpha_k y_k K(x, x_k) + b\right] \tag{3}$$

where α_k represent the support vectors (Lagrange multipliers), and b is a constant. $K(x, x_k)$ represents the Kernel function, which is defined by the dot product between $\boldsymbol{x}$ and $\boldsymbol{x}_k$.

As presented in [14], on the basis of twenty different groups of data, the best general prediction rate was given by LS-SVM classifiers with a RBF (Radial basis function) kernel. In addition, according to [15] in cases where the number of examples for classification is much greater than the number of features, the use of the RBF kernel is also recommended. Accordingly, the RBF kernel was used, defined by:

$$K(x, x_k) = e^{\frac{-\|x - x_k\|^2}{\sigma^2}} \tag{4}$$

When training the LS-SVM model it is necessary to determine the value of parameter γ, as well as the parameters of the selected kernel, in this case the width σ. One of the ways to determine these parameters is the k fold Cross - Validation procedure in combination with a Grid - Search, described in more detail in next section.

3 The Model Update Algorithm

Most machine learning based models implemented for stock market trend prediction use a fixed-size training set in a learning phase. In other words, forecasts for several days, weeks, months or a year are made by a prediction model trained with the same training set known before model construction.

However, in stock market trend prediction that includes the constant input of new data, the generalization capability of the predictor is expected to improve following the completion of the learning process, i.e. with new available data. Thus, a dynamic update of the model is crucial for maintaining and improving the performance of the prediction model.

In the proposed model updating algorithm, an initial prediction model is re-trained on the basis of new incoming data: every P new example, when it becomes available, is added to the initial training set, and the model is then re-trained. The algorithm requires a parameter P to specify the number of training instances that are added to the initial training set, i.e. the time horizon of model re-training. The adaptation of the model to the time-evolving environment can be determined by the changes in the value of P, that is, the current scope of the model.

The optimal value of parameter P is dependent on the observed data time series and mainly based on heuristic methods. In general, large numbers of training instances that are added to the initial training set are preferred for stationary processes, while smaller numbers of instances are preferred in non-stationary environments.

The model update algorithm can be seen in Figure 1.

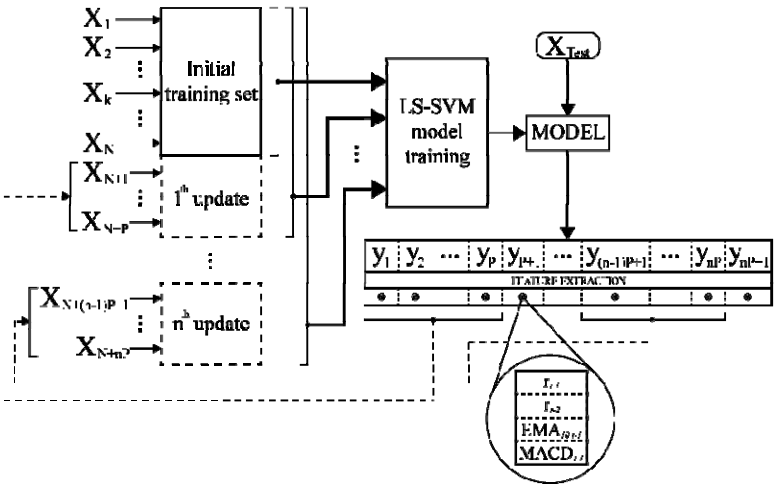

Fig. 1. Model update algorithm

The first step in the proposed algorithm is model-training based on the initial training set. The optimal (γ, σ) pair is determined on the basis of the initial training set T=($\boldsymbol{x}$, y) using a grid-search with k-fold cross validations, as mentioned in section 2. The training set is randomly subdivided into k disjoint subsets of approximately equal

size and the LS-SVM model is built k times with the current pair (γ, σ). Each time, one of the k subsets is used as the test set and the other $k - 1$ subsets are combined to form a training set. After k iterations, the average hit rate is calculated for the current pair (γ, σ). The entire process is repeated with an update of the parameters (γ, σ) until the given stopping criterion is reached, in this case the maximization of the hit rate, although other criterions can be used, depending on the nature of the classification problem. The parameters (γ, σ) are updated exponentially in the given range using predefined equidistant steps, according to the grid-search procedure. After obtaining the optimal (γ, σ) combination, the LS-SVM final forecasting model is formed according to (3) and (4).

The model is then employed for the prediction of the stock market trend for one step ahead.

The first step is the selection of a test instance $\boldsymbol{x}_t$ from the test set t=($\boldsymbol{x}_t$, y_t). It should be noted that at the moment of applying the model on the current test vector $\boldsymbol{x}_t$, the value of the associated target value y_t is unknown.

After that, based on the value of the parameter P, it is necessary to update the initial training set T=($\boldsymbol{x}$, y) for the next prediction step with P past instances from t=($\boldsymbol{x}_t$, y_t), which are known at the moment. Before the selection of the next $\boldsymbol{x}_t$ for the next step, the initial training set is updated by adding stock data from the previous P steps, which are known at that moment, and the model is re-trained. The update and re-training are performed in every *P-th* iteration of the test loop until the given number of instances in the test set is reached.

The training set in the proposed algorithm includes data which were observed after the model was initially constructed, as well as the initial data. The updating model algorithm was designed to make full use of the information, as soon as it becomes available.

4 The Experiment and Results

4.1 The Data Used in the Experiment

The value of the Belex15 index determines the price of the most liquid stocks traded on the regulated market of the Belgrade Stock Exchange. The series consists of six sizes which are determined for each day: the closing price, the change in the value of the index in relation to the previous trading day, in percentages, the opening price, highest price, lowest price and the trading volume.

The available data were divided into two groups. The first group consisted of 1811 records required for the training model, from October 26, 2005 to December 31, 2012. For the second group of data, data from January 3, 2013 to December 31, 2013 were used. A total of 253 days of trading were selected that represent whole trading year. The data from the first group were assigned to the training set, while the data from the second were used for the test set.

4.2 Feature Selection and Model Formation

For stock market trend prediction, features are usually selected from a group of technical or fundamental indicators. In this study, the technical indicators as input features were used to predict the stock market trend. In our previous study [16], we established the basis for the formation of a standard LS-SVM model for predicting the trend of the Belex15 index. There, the process of features selection was studied in more detail, along with the characteristics of the time series. The conducted analyses selected two lagged values of the logarithmic return that were statistically determined based on the values of the auto-correlational coefficients. The Exponential Moving Average (EMA), as the moving average of the closing price calculated using a smoothing factor to place a higher weight on recent closing prices, was then also selected based on its features. This indicator can be used to calculate the values backwards to an almost infinite number of steps (for example, EMA5, EMA100 or EMA200), which is an important characteristics of modeling time series. The EMA feature is consequently adjusted with respect to the time horizon, thus the selected period for calculating the EMA transformation consisted of the previous 10 days. The Moving Average Convergence-Divergence (MACD), as the indicator that measures the strength and direction of the trend and momentum, was added to the current model as it was determined in [17] to be effective in optimizing the investment strategies on emerging markets.

The detailed mathematical formulations for the applied transformations and indicators are given in Table 1.

Table 1. Input features

Features	Formula
Closing price	CP_t, t= 1,2, ... N
Logarithmic return	$r_t = \log CP_t - \log CP_{t-1}$
EMA_N	$EMA_N = r_t * k + EMA_{t-1} * (1-k)$; $k = 2/(N+1)$
MACD	$MACD = EMA_{12} - EMA_{26}$
r_{t-1}	$r_{t-1} = \log CP_{t-1} - \log CP_{t-2}$
r_{t-2}	$r_{t-2} = \log CP_{t-2} - \log CP_{t-3}$

The abovementioned transformations contribute to the stationary nature of the series, which additionally increases the effectiveness of the machine learning algorithm.

In the proposed model, the variable to be predicted is the future trend of the stock market. The feature which serves as a label for the class is a categorical variable used to indicate the movement direction of the logarithmic return on the Belex15 index over time t. If the logarithmic return over time t is larger than zero, the indicator is 1. Otherwise, the indicator is −1. Figure 2 shows the trend fluctuations. It can be determined that in reality the market price trend does not constantly follow a straight line; it is volatile, and the line fluctuates up and down repeatedly, rendering it challenging for prediction.

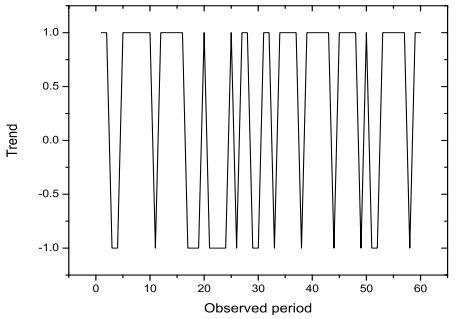

Fig. 2. Trend fluctuations

Based on the previous analysis, the following prediction model was created

$$y_t = LS-SVM(r_{t-1}, r_{t-2}, EMA_{10t-1}, MACD_{t-1}) \tag{5}$$

In order to form the LS-SVM models, LS-SVMlab [18] was used.

4.3 Experimental Results

As a general measure for the evaluation of the prediction effect, the Hit Ratio (HR) was used, which was calculated based on the number of properly classified results within the test group:

$$HR = \frac{1}{m}\sum_{i=1}^{m} PO_i \tag{6}$$

where PO_i is the prediction output of the *i-th* trading day. PO_i equals 1 if is actual value, for the *i-th* training day, otherwise, PO_i equals 0, and *m* is the number of data in the test group [19].

Table 2 shows a comparison of the hit rates obtained using the model updating algorithm (MU-LS-SVMs) with different step sizes $P = \{1, 2, 5, 10, 20\}$ with the Random walk (RW) benchmark model and the LS-SVM model without update. The RW uses the current value to predict the future value, assuming that the latter in the following period (y_{t+1}) will be equal to the current value (y_t). Step sizes are defined based on the definition of the short time stock market periods [20] and previous analyzes of the available time series [16].

The influence of the model update algorithm is clearly positive, since all updated models outperformed the model without update. It can be assumed that both MU LS-SVM_1 and MU LS-SVM_2 will outperform other models because of the observed strong autocorrelation factors in a time series for lag one and two. In addition, it can be seen that from other group of models, the best accuracy was achieved using the MU LS-SVM_{10} model, which further supports the validity of the selected parameters of the EMA features.

Table 2. Prediction accurancy of different prediction models

Model	Hit rate
RW	0.5000
LS-SVM	0.5396
MU-LS-SVM$_1$	0.5555
MU-LS-SVM$_2$	0.5555
MU-LS-SVM$_5$	0.5436
MU-LS-SVM$_{10}$	0.5476
MU-LS-SVM$_{20}$	0.5436

Furthermore, the comparison of the MU LS-SVM$_1$, RW and LS-SVM model on a temporal sequence basis which corresponds to the real frameworks of trading on the Belgrade stock exchange was studied, including the weekly, biweekly, monthly, bi-monthly, and quarterly work regime. This went on until entire trading year. The results are shown in Table 3.

Table 3. The models comparison results on the predefined time-sequence

Time Sequences	RW	LS-SVM	MU LS-SVM$_1$
0-5	0.6000	0.6000	0.6000
0-10	0.8000	0.8000	0.8000
0-20	0.7000	0.7000	0.7000
0-40	0.6000	0.6750	0.6500
0-60	0.6000	0.6167	0.6333
0-80	0.6125	0.6000	0.6125
0-100	0.6100	0.6300	0.6400
0-120	0.6083	0.6500	0.6583
0-140	0.5714	0.6286	0.6357
0-160	0.5438	0.5875	0.5938
0-180	0.5389	0.5889	0.5944
0-200	0.5200	0.5500	0.5550
0-220	0.5113	0.5520	0.5611
0-240	0.5125	0.5542	0.5625
0-252	0.5000	0.5397	0.5556

It can be noted that in the approximated first trading month, the rate of the hits is identical for all presented models. This can be explained by insufficient additional new training data and it is in favor of the previously noted strong correlation in the available data series. The longer the time period, the more dominant the prediction based on the proposed model update algorithm.

This algorithm extends computational time. The time needed to obtain the predictions increases for all the models that implement the update approach, compared to

the model trained with an initial training set (by approximately 150 seconds compared to 100 seconds). Nevertheless, an increase in computational time is compensated with an increase in the quality of the prediction results.

The results are obtained for one-day-ahead predictions using data over an extended period of time, one trading year, and exceed most of the time horizons presented in [5], [19], [21], [22], but are still in their mid-range. The results are reliable, based on all the currently available information, representing all the forms of model behavior.

5 Conclusion

A practical approach to building a dynamic model for the stock market trend prediction is proposed. Although the complexity of the calculations in the proposed algorithm is increased when compared to training only one forecasting model, it brings significant improvements in terms of stock market prediction accuracy. Every increase in precision is considered an exceptional contribution as it leads to an increase in the return and the decrease in the risk involved in trading.

As far as further research is concerned, first, in the proposed approach, prior information was not excluded. Since short periods of time were observed in the time series analyzed in this paper, the issue was not dealt with separately. In the case of the longer periods of time, the prediction model should not include all the available data. Thus, past information should be removed using a new methodology designed for that purpose.

Finally, most studies in this field deal with the prediction of market indices and the price of financial instruments on developed markets. It is important to emphasize that the prediction rate obtained in this study belongs to the stock index of emerging market of the Republic of Serbia, and that it gave competitive results.

References

1. Wang, Y., Choi, I.C.: Market Index and Stock Price Direction Prediction using Machine Learning Techniques: An empirical study on the KOSPI and HIS. ScienceDirect, 1–13 (2013)
2. Kumar, M., Thenmozhi, M.:Forecasting stock index movement: a comparison of support vector machines and random forest. In: Indian Institute of Capital Markets 9th Capital Markets Conference Paper (2006), SSRN: http://ssrn.com/abstract=876544, http://dx.doi.org/10.2139/ssrn.876544
3. Kara, Y., Boyacioglu, M., Baykan, Ö.K.: Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert Systems with Applications 38, 5311–5319 (2011)
4. Atsalakis, G.S., Valavanis, K.P.: Surveying stock market forecasting techniques – Part II: Soft computing methods. Expert Systems with Applications 36, 5932–5941 (2009)
5. Huang, W., Nakamori, Y., Wang, S.: Forecasting stock market movement direction with support vector machine. Computers & Operations Research 32, 2513–2522 (2005)
6. Phichhang, O., Wang, H.: Prediction of Stock Market Index Movement by Ten Data Mining Techniques. Modern Applied Science 3, 28–42 (2009)

7. Jiang, J., Song, C., Zhao, H., Wu, C., Liang, Y.: Adaptive and Iterative Least Squares Support Vector Regression Based on Quadratic Renyi Entropy. In: Granular Computing, GrC 2008, pp. 340–345. IEEE Press, New York (2008)
8. Read, J., Bifet, A., Pfahringer, B., Holmes, G.: Batch-Incremental versus Instance-Incremental Learning in Dynamic and Evolving Data. In: Hollmén, J., Klawonn, F., Tucker, A. (eds.) IDA 2012. LNCS, vol. 7619, pp. 313–323. Springer, Heidelberg (2012)
9. Doomretni, C., Giunnoepulos, D.: Incremental Support Vector Machine Construction. In: Data Mining, ICDM 2001, pp. 589–593. IEEE Press, New York (2001)
10. Laskov, P., Gehl, C., Kruger, S., Muller, K.: Incremental Support Vector Learning, Analysis, Implementation and Applications. Journal of Machine Learning Research 7, 1909–1936 (2006)
11. Stojanović, M., Božić, M., Stajić, Z., Milošević, M.: LS-SVM model for electrical load prediction based on incremental training set update. Przegląd Elektrotechniczn 4, 195–199 (2013)
12. Guajardo, J.A., Weber, R., Miranda, J.: A model updating strategy for predicting time series with seasonal patterns. Applied Soft Computing 10, 276–283 (2010)
13. Suykens, J., Vandewalle, J.: Least Squares Support Vector Machines. Neural Processing Letters 9, 293–300 (1999)
14. Gestel, T.V., Suykens, A.K., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., Moor, B.D., Vandewalle, J.: Benchmarking Least Squares Support Vector Machine Classifiers. Machine Learning 54, 5–32 (2004)
15. Božić, M., Stajić, Z., Stojanović, M.: Short-term load forecasting using least square support vector machines. In: Infoteh Jahorina, pp. 326–329 (2010)
16. Marković, I., Stanković, J., Stojanović, M., Božić, M.: Stock exchange trend prediction of Belex15 index with LS-SVM classifier. In: XIII International Symposium Infoteh-Jahorina, pp. 739–742 (2014)
17. Eric, D., Andjelic, G., Redzepagic, S.: Application of MACD and RVI indicators as functions of investment strategy optimization on the financial market. Proceedings of the Faculty of Economics of Rijeka 27(1), 171–196 (2009)
18. Brabanter, K.D., Karsmakers, P., Ojeda, F., Alzate, C., Brabanter, J.D., Pelckmans, M.D.K., Vandewalle, B.J., Suykens, J.A.K.: LS-SVMlab Toolbox User's Guide. Technical report, ESAT-SISTA (2011)
19. Yuling, L., Guo, H., Hu, J.: An SVM-based Approach for Stock Market Trend Prediction. In: Neural Networks (IJCNN), pp. 1–7. IEEE Press, New York (2013)
20. Bradić-Martinović, A.: Stock market prediction using technical analysis. Economic Anal. 170, 15–145 (2006)
21. Lahmiri, S.: A Comparison of PNN and SVM for Stock Market Trend Prediction using Economic and Technical Information. International Journal of Computer Applications 29, 24–30 (2011)
22. Ni, L.-P., Ni, Z.-W., Gao, Y.-Z.: Stock trend prediction based on fractal feature selection and support vector machine. Expert Systems with Applications 38, 5569–5576 (2011)

Open Financial Data from the Macedonian Stock Exchange

Bojan Najdenov, Hristijan Pejchinoski, Kristina Cieva,
Milos Jovanovik, and Dimitar Trajanov

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University,
Skopje, Republic of Macedonia
{bojan.najdenov,milos.jovanovik,dimitar.trajanov}@finki.ukim.mk,
hristijan.pejcinoski@gmail.com, kristina_cieva@yahoo.com

Abstract. The concept of Open Data, which represents the idea that public data should be published in a machine-readable format, starts to take a significant role in modern society. Public data from various fields are being transformed in open data formats and published on systems which allow easier consumption from software agents and applications, as well as the users behind them. On the other hand, people in the business world are trying for a few decades now to establishing standards for financial accounting that govern the preparation of financial reports. Financial reporting has crucial significance for companies today, since it is a record of their work which is presented to their stakeholders and represents a starting point for future business decisions and strategies. In this paper, we use data from the Macedonian Stock Exchange and data from different web sites of Macedonian companies in order to create datasets of Open Financial Data relevant for our country, thus increasing the transparency and improving the data accessibility. We describe the process of transforming the data into 4 star Open Data, and present use-case scenarios which use data from our generated datasets and from the World Bank. The datasets are published and accessible via a SPARQL endpoint, and we demonstrate how a software application can make use of them.

Keywords: Finances, Open Data, Macedonian Stock Exchange, World Bank, RDF, Ontologies.

1 Introduction

The main idea that lies behind the concept of Open Data[1] is that public data should be free and available to everyone. We live in a world where information holds great value. Having the right information at the right time, in the right way, builds modern societies, drives technologies forward, develops businesses and even saves lives. The exponential growth of datasets about people, technological artifacts and organizations brought us in position where we have on disposal vast amounts of information ready

[1] http://okfn.org/opendata/

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_12

to be rearranged and shaped in order to create additional value [1]. This implies that the structured, machine-readable, open data that is free to access and interlink, is becoming the future of startup companies and business in general.

Linked Open Data[2] is a community effort to alleviate the problem of the lack of sufficiently interlinked datasets on the Web. Through this effort, a significant number of large-scale datasets have now been published in the LOD cloud[3], which is growing constantly [2]. As we see in Fig. 1, datasets from different fields are publicly available in Linked Open Data format, thanks to the contributors to the Linked Open Data community [3].

The concept of Linked Open Data provides us with a way to connect datasets stored on different locations, by using the Semantic Web standards such as RDF, OWL and SPARQL. By using the existing Web infrastructure, data from different data silos can be successfully interconnected and the Web can be used to decrease the barriers which occur during process of linking [4]. Linked Open Data enables better data analysis by simplifying the process of combining information sources. Datasets from various industry fields can then be used in different ways and by many entities, e.g. regulatory bodies and banks, when it comes to financial data [5].

Financial accounting is primarily oriented towards creating the financial reports for the companies' work. The process of creating the reports is dependent on a huge amount of datasets. Thus, this is a field that necessarily requires different approaches for representation, storage, querying and visualizing of the data. We find this issue of big importance for today's companies and economies which motivated us to work on a practical solution using data about Macedonian companies provided by the Macedonian Stock Exchange, the World Bank and the information and data they publish on their websites.

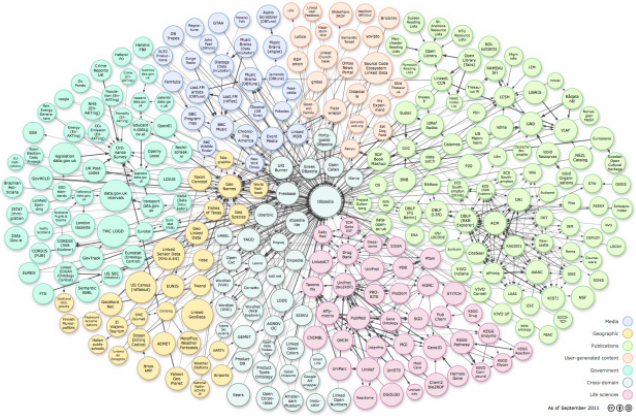

Fig. 1. The LOD Cloud, as of September 2011

[2] http://linkeddata.org/
[3] http://lod-cloud.net/

2 Related Work

Numerous projects exist which have a major target to either publish financial or corporate data in Open Data formats, or enable their annotation with the technologies of the Semantic Web, in order to leverage their value. The World Bank, as one of the most important financial institutions on a global level, puts great effort in many projects which result with creating Open Data. Other significant projects in this area are the Financial Industry Business Ontology (FIBO), the Open Corporates project and the Financial Report Ontology.

The World Bank aims towards decreasing extreme poverty in the world, through proving financial and technical assistance to developing countries. The financial support the developing countries receive is in form of low-interest loans, credits and grants, or investments in various areas like healthcare, education, infrastructure, resource management etc. The World Bank, as a global institution, supports the ideas behind the Open Data concept, and therefore shares its public data freely on their website[4].

In [6], the authors introduce an interesting project which aims towards designing new methods for extraction of data and, based on that, developing a prototype for extracting financial information from the semi-structured text. They believe that in the financial world numbers are often one main target, but they are meaningless without any semantic meta-data describing what kind of information they represent.

The Financial Industry Business Ontology[5] (FIBO) is an initiative to define and describe terms and rules for financial data. Its goal is building a representation of the information about financial instruments, market data, business entities, etc. along with the relationships between them.

Open Corporates[6] is one of the largest open databases of companies in the world, having information about 63 million companies from around the globe. They publish the data in XML, RDF or JSON format and it can be downloaded from their website. They believe that basic corporate information about all the companies in the world should be brought together in one place, making it easier to access, use and connect with other data.

The Financial Report Ontology[7] is a project developed with the idea of providing an ontology that would describe the financial reports as concepts, as well as their individual entries. The ontology aims to assist companies in the process of creating annotated financial reports.

3 Macedonian Open Financial Data

3.1 Public Data from the Macedonian Stock Exchange

The Macedonian Stock Exchange (MSE)[8] is the only financial institution in Macedonia that is authorized to organize, execute and regulate the trading of

[4] http://data.worldbank.org/
[5] http://www.omg.org/hot-topics/fibo.htm
[6] http://opencorporates.com/
[7] http://financialreportontology.wikispaces.com/
[8] http://www.mse.mk/en/

securities. It was established in 1995 as a joint stock company and the first trading occurred in March, 1996. The main purpose of MSE is to provide security and efficiency in the organized trading of securities in Macedonia.

MSE is comprised of two market segments: Official Market and Regular Market. The stock market indices are MBI10 (Macedonian Blue Chip Index), which includes the stocks of the 10 most traded companies, MBID (Macedonian Stock Exchange Index of publicly held companies), which includes the stocks of the publicly held companies listed on MSE and OMB (Bond Index), which includes issued bonds listed on MSE.

MSE publishes most of its data on their website, either as PDF files or in HTML tables. Among all of the published data, like stock prices, different indices, information about growth trends on securities, etc., our main topic of interest are the financial reports which MSE member companies publish. We gathered the financial report data from the MSE website, converted it and stored it in CSV format. We did the same process for gathering and storing the company data, which we obtained from individual companies websites.

3.2 Open Data from the World Bank

As we already noted, the World Bank published data from its projects on their website. Parts of these data are the financial data, which allow us to see what global funds the World Bank manages, visualize them or build models over them.

Many different financial datasets can be found on World Bank's website[9] in various different formats, such as CSV, JSON, PDF, RDF, RSS, XLS, XLSX and XML. Some of their datasets can be accessed via the public SPARQL endpoint which the World Bank provides[10], as part of their Linked Data project. The dataset that we are interested in contains data on commitments against contracts that were reviewed by the Bank before they were awarded (prior-reviewed Bank-funded contracts) under IDA/IBRD[11] investment projects and related Trust Funds. We downloaded this dataset in RDF format and linked its data with data published by MSE and Macedonian companies. The procedure will be described in details.

4 Ontologies for the Datasets

4.1 Ontology for the World Bank Dataset

We loaded the dataset from the World Bank data store into a local Virtuoso Universal Server[12] instance, as an RDF graph. Since all the entries in the dataset refer to a loan awarded to a company by the World Bank, a single entry in the dataset can be considered as a resource which provides all the details related to a specific loan.

[9] https://finances.worldbank.org/all-datasets
[10] http://worldbank.270a.info/sparql
[11] http://data.worldbank.org/indicator/DT.DOD.MWBG.CD
[12] http://virtuoso.openlinksw.com/

4.2 Corporate Registry Ontologies

As we already mentioned, Open Corporates holds a large publicly available dataset of information about companies as legal entities, for all around the world. Unfortunately, they do not hold any information about Macedonian companies, and therefore we cannot use their datasets in the context of Macedonian financial data.

However, we did analyze their data and the ontologies they use for semantic annotation, so we decided to reuse those ontologies and annotate our data in a similar manner. Another motivation for this was the similarity between the structures of the dataset from Open Corporate had with the data we were able to collect for Macedonian companies. The ontologies we use in describing the companies as legal entities are listed in Table 1.

Table 1. The ontologies we reused for Macedonian company data

Prefix	URI
foaf	http://xmlns.com/foaf/0.1/
vCard	http://www.w3.org/2006/vcard/ns#
adms	http://www.w3.org/ns/adms#
rov	http://www.w3.org/ns/regorg#
skos	http://www.w3.org/2004/02/skos/core#

We use the rov:RegisteredOrganization class in order to represent a legal entity or organization which is legally registered, i.e. a company that we have data about. The rest of the DataType properties we use to describe a Registered Organization can be found in Table 2.

Table 2. The DataType properties we use

Property	Description
rov:legalName	The legal name of the company.
rov:registration	The registration is a fundamental relationship between a legal entity and the authority with which it is registered and that confers legal status upon it. rov:registration is a sub property of adms:identifier which has a range of adms:Identifier.
vCard:extended-address	The address of the object.
vCard:hasTelephone	To specify the telephone number for telephony communication with the object.
skos:notation	Refined name of a company.
foaf:homepage	A homepage for some company. Every value of this property is a foaf:Document.
rdfs:label	Information about the basic activities of a company.

4.3 Financial Report Ontology

Every member of the Macedonian Stock Exchange provides annual financial reports which are the balance sheet, income statement, statement of cash flows and the statement of retained earnings. Our focus in this paper is the balance sheet of the companies in particular, which requires an ontology to be provided so that we could semantically annotate that data.

For this purpose we decided to reuse the Financial Report Ontology which, as we already described, defines the basic financial report terms.

In the ontology we find the class Fundamental Accounting Concept, which represents one full financial report. Its properties are divided into five groups: General Information properties, Balance Sheet, Income Statement, Statement of Comprehensive Income and Cash Flow Statement properties. For our local reports we will use only General Information, Balance Sheet and Income Statement properties.

Table 3. The properties in the CFRL ontology

Property	Description
cfrl:hasReport	This property connects a company i.e. instance of RegisteredOrganization class, with its financial report.
cfrl:hasLoan	This property points to the World Bank loans that are made by that company.

4.4 Corporate Financial Reports and Loans Ontology

In order to be able to successfully complete the annotation and linking process between the datasets, we developed the Corporate Financial Reports and Loans Ontology (CFRL). In it, we introduce two object properties: “hasReport” and “hasLoan”. Their main role is to provide means of interlinking the datasets. The description of these two properties can be found in Table 3.

5 Linking the Datasets

Before we begin explaining the process of interlinking the datasets, we must state that our goal is to interlink the data from our corporate registry dataset, i.e. the data we gathered from various websites of different companies, with the data we acquired from the World Bank about loans that companies were awarded and also with the financial reports data we got from the Macedonian Stock Exchange. Conceptually, the linking we wish to achieve is shown in Fig. 2.

Fig. 2. Linking the datasets

5.1 Mapping the Data from CSV to RDF

The next step of our work is mapping and transforming datasets from CSV files to RDF and to do that, we use the Virtuoso Universal Server, which provides mechanisms for data transformation management and querying using SPARQL.

The technical process of mapping and transforming the data from CSV to RDF was done using the R2RML mapping language[13], as described in [7], for the corporate registry dataset and the financial reports dataset, respectively. The dataset about loan details originating from the World Bank was already in RDF format, so we imported it directly in the Virtuoso Universal Server instance, as an RDF graph.

5.2 Interlinking the RDF Datasets

Having transformed all the datasets into RDF graphs in Virtuoso, our next step was to interlink the data, as shown in Fig. 2. For that purpose of we created two properties in our CFRL ontology: “cfrl:hasReport” and “cfrl:hasLoan”.

The “cfrl:hasReport” property links a company with its published financial reports. That means, we connect a “RegisteredOgranization” entity with its financial report i.e. “FundamentalAccountingConcept”, by matching the values of the company name. For this purpose we use the “skos:notation” property of a “RegisteredOgranization” entity, and the “fac:EntityRegistrantName” property of a “FundamentalAccountingConcept” entity.

The property “cfrl:hasLoan” interlinks a “RegisteredOgranization” entity with its loan entities. Similarly to the previous property, we create the connection by matching the names of the correspondent companies. For this purpose we use the “rov:legalName” of an “RegisteredOrganization” entity, and the “worldbank:supplier” property of a loan entry.

These interlinking processes were done using SPARQL queries over the datasets.

The resulting linked data that we generated, can be accessed through a public SPARQL endpoint[14].

[13] http://www.w3.org/TR/r2rml/
[14] http://linkeddata.finki.ukim.mk/sparql

6 Use-Cases

The main purpose of using interlinked Open Data datasets is the ability to increase the value and usability of the separate datasets, by providing advances use-case scenarios. We are going to describe two of the many possible scenarios.

6.1 Displaying Information from the World Bank

We demonstrate the use of the "hasLoan" property to retrieve information about a company which obtained a loan from the World Bank, or to be more precise, the dates when the company signed contracts for getting loans with the World Bank, the total contract amount (USD) and which sector was the loan dedicated to. For the purpose of the demonstration, we show the top 5 loans and their details. The SPARQL query is the following:

```
prefix cfrl: <http://linkeddata.finki.ukim.mk/lod/ontology/cfrl#>
prefix worldbank: <http://finances.worldbank.org/resource/>
prefix rov: <http://www.w3.org/ns/regorg#>

SELECT ?s ?csd ?tca ?ms WHERE {
   ?company rov:legalName ?s .
   ?s cfrl:hasLoan ?l .
   ?l worldbank:contract_signing_date ?csd ;
      worldbank:supplier_contract_amount_usd ?tca ;
      worlbank:major_sector ?ms .
} ORDER BY DESC (?tca) LIMIT 5
```

The result of the executed query at our Virtuoso SPARQL endpoint, are shown in Table 4.

Table 4. Results from the SPARQL query

Supplier	Contract signing date	Total contract amount	Major sector
Granit	Mar 26, 2009	$9,802,524.00	Transportation
Granit	Dec 04, 2009	$6,197,108.00	Transportation
Granit	Dec 04,2009	$5,323,028.00	Transportation
Granit	Mar 26, 2009	$4,519,095.00	Transportation
Granit	Dec 04,2009	$3,785,761.00	Transportation

6.2 Displaying Information from the Financial Reports

In this section we show how the "hasReport" property that we defined in our CFRL ontology, can be used to provide additional information about companies. One such scenario would be to retrieve information about the top 5 companies by the profit they

made in the year of 2012, in Macedonian Denars (MKD). For that purpose we can use the following SPARQL query:

```
prefix cfrl: <http://linkeddata.finki.ukim.mk/lod/ontology/cfrl#>
prefix fac:
<http://www.xbrlsite.com/2013/FinancialReportOntology/Prototype04/FundamentalAc
countingConcepts.xml#>
prefix rov: <http://www.w3.org/ns/regorg#/>

SELECT ?name ?profit ?period WHERE {
   ?cmp cfrl:hasReport ?rep ; rov:legalName ?name .
   ?rep fac:GrossProfit ?profit ; fac:FiscalPeriod ?period .
   FILTER (?period = 2012)
} ORDER BY ?profit LIMIT 5
```

The result of this query, showing the name of such companies and the profit they made in the year of 2012, can be seen in Table 5.

Table 5. Results from the SPARQL query

Name	Profit (MKD)	Period
ALKALOID AD SKOPJE	3,291,423	2012
Stopanska Banka AD Skopje	2,376,477	2012
Tikvesh AD Skopje	339,049	2012
GD GRANIT AD - Skopje	291,238	2012
Vitaminka AD Prilep	102,378	2012

7 Conclusion and Future Work

Data, information and knowledge management are key activities in modern economies and considerable efforts and resources are devoted for research in these areas, by different organizations in the world. Having data structured and interlinked provides a whole new area of opportunities for data usage and management. This provides huge benefits in the information dissemination processes and provides mechanisms so that information can be shared easily between bank divisions, institutions and distributed to all stakeholders.

In this paper we gave an overview of the process of transforming the one-star and two-star data about companies into four-star Open Data and connected it with a dataset from the World Bank. We also provided use-case scenarios which gave examples of how our local data and how the data from the World Bank can be used in order to provide information which is not available when the datasets are isolated. With this, we hope our work contributes to the goals of the Open Data Initiative[15] in Macedonia.

[15] http://opendata.gov.mk/

In the future, we plan to continue our work in these fields, increase the amount of datasets, connect our data with other remote resources and transform these datasets further to five-star data, interlinked with financial data published on the LOD cloud. This would improve the quality of the use-cases we provide and also create new opportunities for development of creative applications and analysis. We hope our work serves as a motivation to companies, financial institutions, organizations around the world, to recognize the benefits of open financial data and publish their public data on the Web in raw and machine-readable format.

Acknowledgment. The work in this paper was partially financed by the Faculty of Computer Science and Engineering, at the Ss. Cyril and Methodius University in Skopje, as part of the research project "Semantic Sky 2.0: Enterprise Knowledge Management".

References

1. Cardoso, J., Pedrinaci, C., Leidig, T., Rupino, P., De Leenheer, P.: Open semantic service networks. In: International Symposium on Services Science (ISSS), Leipzig, Germany (2012)
2. Möller, K., Hausenblas, M., Cyganiak, R., Handschuh, S., Grimnes, G.: Learning from Linked Open Data Usage: Patterns & Metrics. In: Web Science Conference (WSC) (2010)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), 1–22 (2009)
4. Kundra, V.: Digital Fuel of the 21st Century: Innovation through Open Data and the Network Effect. Joan Shorenstein Center on the Press, Politics and Public Policy, Harvard College (2012)
5. Radzimski, M., Sánchez-Cervantes, J.L., Rodríguez-González, A., Gómez-Berbís, J.M., García-Crespo, A.: FLORA –Publishing Unstructured Financial Information in the Linked Open Data Cloud. In: First International Workshop on Finance and Economics on the Semantic Web (FEOSW) (2012)
6. Bjoraa, E.: Ontology guided financial knowledge extraction from semi-structured information sources. Master Thesis in Information and Communication Technology, Agder University Colledge, Grimstad (May 2003)
7. Jovanovik, M., Najdenov, B., Trajanov, D.: Linked Open Drug Data from the Health Insurance Fund of Macedonia. In: 10th International Conference for Informatics and Information Technology (2013)

Pseudo Random Sequence Generators Based on the Parastrophic Quasigroup Transformation

Verica Bakeva, Vesna Dimitrova, and Mile Kostadinoski

Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University, Skopje, Macedonia
{verica.bakeva,vesna.dimitrova}@finki.ukim.mk,
mile.kostadinoski.mk@gmail.com

Abstract. Pseudo random sequence generators (PRSG) produce sequences of elements that imitate natural random behavior and they have extensive applications in many fields like cryptography, authentication and cryptanalysis. Using quasigroup string transformations, a PRSG is introduced in [1]. Here, we propose a new design of PRSG using parastrophic quasigroup transformation defined in [2]. This generator is called Parastrophic Quasigroup Pseudo Random Sequence Generator (PQPRSG). We investigate the goodness of quasigroups of order 4 for designing of PQPRSQG using classifications given in [3] and linearity of quasigroups defined in [4]. At the end, we give experimental results about the period of the generator.

Keywords: pseudo random sequence generators, quasigroup, quasigroup transformation, parastrophe, period.

1 Introduction

A pseudo random sequence generators (PRSG) is a deterministic algorithm that produce almost random sequences of elements. A PRSG starts with some random sequence of elements, which is usually a short random sequence known as seed and returns output that is much longer pseudo random sequence of elements. Since PRSGs are deterministic algorithms, there is no guaranty that a theoretically ideal random sequence can be produced. They produce only pseudo random sequences. Every sequence, produced with PRSG has a certain period. A period of pseudo random sequence represents the minimal distance between two sequential appearances of the same pseudo random subsequence [5,6].

In this paper we propose a new design of PRSG using parastrophic quasigroup transformations. Quasigroups and quasigroup transformations are very useful for designing of cryptographic primitives, error detecting and error correcting codes. The reasons for that are the structure of quasigroups, their large number, the properties of quasigroup transformations, etc. The quasigroup string transformations and their properties were considered in several papers ([7] and other).

A quasigroup $(Q, *)$ is a groupoid (i.e. an algebra with one binary operation $*$ on the finite set Q) satisfying the following property:

$$(\forall u, v \in Q)(\exists! x, y \in Q) \ \ (x * u = v \ \& \ u * y = v) \tag{1}$$

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_13

According to (1) a groupoid $(Q, *)$ is a quasigroup if and only if the equations $x*u = v$ and $u*y = v$ have unique solutions x and y for each given $u, v \in Q$. Every quasigroup $(Q, *)$ has a set of five quasigroups, called *parastrophes*, denoted by $/, \backslash, \cdot, /\!\!/, \backslash\!\!\backslash$ which are defined in Table 1.

Table 1. Parastrophes of quasigroup operations $*$

Parastrophes operation
$x\backslash y = z \iff x * z = y$
$x/y = z \iff z * y = x$
$x \cdot y = z \iff y * x = z$
$x/\!\!/y = z \iff y/x = z \iff z * x = y$
$x\backslash\!\!\backslash y = z \iff y\backslash x = z \iff y * z = x$

In this paper we use the following notation for parastrophic operations:

$$f_1(x, y) = x * y,\ f_2(x, y) = x\backslash y,\ f_3(x, y) = x/y,$$
$$f_4(x, y) = x \cdot y,\ f_5(x, y) = x/\!\!/y,\ f_6(x, y) = x\backslash\!\!\backslash y.$$

Let $A = \{1, \ldots, s\}$ $(s \geq 2)$ be an alphabet and denote by $A^+ = \{x_1 \ldots x_k |\ x_i \in A,\ k \geq 1\}$ the set of all nonempty finite strings over A.

Note that $A^+ = \bigcup_{k \geq 1} A^k$, where $A^k = \{x_1 \ldots x_k |\ x_i \in A\}$. Assuming that (A, f_i) is a given quasigroup, for a fixed letter $l \in A$ (called leader) a transformation $E = E_{f_i,l} : A^+ \to A^+$ (see [7]) can be defined by

$$E_{f_i,l}(x_1 \ldots x_k) = y_1 \ldots y_k \Leftrightarrow \begin{cases} y_1 = f_i(l, x_1), \\ y_j = f_i(y_{j-1}, x_j), \quad j = 2, \ldots, k. \end{cases} \tag{2}$$

Next, let describe briefly a modified quasigroup transformation called parastrophic quasigroup transformation, defined in [2], which later will be used in the implementation of the PRSG. Let p be a positive integer and $x_1 x_2 \ldots x_n$ be an input message. Using previous transformation E, a parastrophic quasigroup transformation $PE = PE_{l,p} : A^+ \to A^+$ can be defined as follows.

At first, let $d_1 = p$, $q_1 = d_1$, $s_1 = (d_1 \bmod 6) + 1$ and $A_1 = x_1 x_2 \ldots x_{q_1}$. Applying the transformation $E_{f_{s_1},l}$ on the block A_1, we obtain the encrypted block

$$B_1 = y_1 y_2 \ldots y_{q_1} = E_{f_{s_1},l}(x_1 x_2 \ldots x_{q_1}).$$

Further on, for given i, let the encrypted blocks $B_1, \ldots, B_{i-1}$ be obtained and d_i be calculated using the last two symbols in B_{i-1}, i.e,. $d_i = 4y_{q_{i-1}-1} + y_{q_{i-1}}$. Let $q_i = q_{i-1} + d_i$, $s_i = (d_i \bmod 6) + 1$ and $A_i = x_{q_{i-1}+1} \ldots x_{q_i - 1} x_{q_i}$. We apply the transformation $E_{f_{s_i}, y_{q_{i-1}}}$ on the block A_i and obtain the encrypted block

$$B_i = E_{f_{s_i}, y_{q_{i-1}}}(x_{q_{i-1}+1} \ldots x_{q_i}).$$

Now, the parastrophic transformation is defined as

$$PE_{l,p}(x_1x_2\dots x_n) = B_1||B_2||\dots||B_r. \tag{3}$$

Note that the length of the last block A_r may be shorter than d_r (depends on the number of letters in input message).

For arbitrary quasigroup on a set A and for given $l_1,\dots l_n$ and $p_1,\dots,p_n$, we define mappings PE_1, PE_2, …, PE_n as in (3) such that PE_i is corresponding to p_i and l_i. Using them, we define the transformation $PE^{(n)}$ as follows:

$$PE^{(n)} = PE^{(n)}_{(l_n,p_n),\dots,(l_1,p_1)} = PE_n \circ PE_{n-1} \circ \dots \circ PE_1,$$

where $\circ$ is the usual composition of mappings.

An important property of one transformation for application in cryptography is the uniform distribution of the substrings in the output message. This property is given in [8] with the following theorem.

Theorem 1. *Let $\alpha \in A^+$ be an arbitrary string and $\beta = PE^{(n)}(\alpha)$. Then the m-tuples in β are uniformly distributed for $m \leq n$.*

Further on, this theorem is reason for using PE-transfomation in design of PRSG based on the parastrophic quasigroup transformation.

2 Classifications of Quasigroups of Order 4 Useful in Cryptography

Many classifications of quasigroups of order 4 are given in several papers. These classifications are made in order to distinguish the quasigroups useful in cryptography from the ones that are not. In this part we will consider several classifications of the quasigroups of order 4 and we will derive a conclusion which class of quasigroups is most suitable for using in cryptography.

2.1 Classification by Number of Different Parastrophes

The first classification of quasigroups of order 4 that we consider is given in [3] and it is based on the number of different parastrophes of a quasigroup. For each of the 576 quasigroups of order 4, the number of different parastrophes is found and according to this number, the set of quasigroups of order 4 is divided into 4 classes.

In Table 2 we give a classification of quasigroups by the number of different parastrophes and by the fractality of the quasigroups. The fractality of quasigroups is introduced in [9] and based on it, quasigroups are divided into fractal and non-fractal. According to the number of different parastrophes, the fractal and non-fractal quasigroups of order 4 are classified separately.

Table 2. Cardinality of classes of quasigroups by number of different parastrophes and cardinality

No. parastrophes	Total No. quasigroups	No. fractal quasigroups	No. non-fractal
1	16	16	0
2	2	2	0
3	240	96	144
6	318	78	240
Total	576	192	384

2.2 Classification Using *PE*-Transformation

Using PE- transformation instead of E-transformation, in [3], authors proposed a similar classification based on fractality of quasigroups. According to this classification the class of fractal quasigroups is divided in 2 subclasses: parastrophic-fractal and fractal parastrophic-non-fractal quasigroups. The cardinality of subclasses of parastrophic fractal and fractal parastrophic-non-fractal quasigroups for different number of parastrophes are given in Table 3.

Table 3. Cardinality of subclass of fractal quasigroups by number of different parastrophes

No. parastrophes	No. parastrophic fractal	No. fractal parastrophic non-fractal
1	16	0
2	0	2
3	72	24
6	0	78
Total	88	104

2.3 Classification Using Boolean Representation of Quasigroups

In [4], the authors gave a representation of a quasigroup as vector valued Boolean functions. Namely, a quasigroup $(Q, *)$ of order 2^n can be represented by a vector valued Boolean function $f : \{0,1\}^{2n} \to \{0,1\}^n$. Let represent an arbitrary x of the quasigroup as binary vector $x = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$. Then, for each $x, y \in Q$

$$x * y \equiv f(x_1, \ldots, x_{2n}) = (f_1(x_1, \ldots, x_{2n}), \ldots, f_n(x_1, \ldots, x_{2n}))$$

where we take

$$x = (x_1, x_2, \ldots, x_n),\ y = (x_{n+1}, x_{n+2}, \ldots, x_{2n})$$

and

$$f_i : \{0,1\}^{2n} \to \{0,1\}$$

are the corresponding components of f.

Using this Boolean representation, in [4], the quasigroups are divided into two classes: *linear* quasigroups and $non - linear$ quasigroups by Boolean representation. There are 144 linear and 432 non-linear quasigroups of order 4.

Analysing all previous classifications of quasigroups of order 4, using exhaustive verification, we prove the following new theorem.

Theorem 2. *Parastrophes of linear quasigroups by Boolean representation of order 4 are linear as well.*

Directly from the theorem, the following corollary holds.

Corollary 1. *All non-linear quasigroups by Boolean representation of order* 4 *have non-linear parastrophes.*

The properties of Theorem 4 and Corollary 1 are used for determination which quasigroup are suitable for design of proposed generator.

3 Pseudo Random Sequence Generator Using Parastrophic Quasigroup Transformation

As we mentioned before, quasigroups and quasigroup transformations can be used for construction of PRSG. The reason for this can be found in the structure of the quasigroups and their large number, but also in the properties of the quasigroup transformations. In [1], the authors introduced an implementation of PRSG using quasigroup transformation. This generator is highly scalable, fairly unpredictable and cryptographically secure if the quasigroup (used to build the generator) remains unknown.

In this paper we propose a similar design of PRSG using PE-transformations. This generator is called *Parastrophic Quasigroup Pseudo Random Sequence Generator* (PQPRSG). Before we show the implementation of PQPRSG, let briefly discuss the reasons for using PE-transformation in this design.

One of the main reasons is the uniform distribution of the output sequence. According to Theorem 1, the m-tuples in the output sequence, after n applications of the PE-transformation, are uniformly distributed for $m \leq n$, which provides a natural behavior of the pseudo random subsequences of length not greater than n. This means, that we can apply the PE-transformation sufficiently many times on a sequence and we can expect to obtain sequence with enough large period. Notice that we must have enough large sequence in order to done a relevant statistical analysis and to obtain uniform distribution of m-tuples after n applications of PE-transformation.

Also, using PE-transformation, we increase the number of quasigroups useful in cryptography. Fractal quasigroups are not good for designing cryptographic primitives since they produce regular structures. But, with the classification

using PE-transformation, some of the fractal quasigroups can still be used for cryptographic purposes (since they become parastrophic-non-fractal).

Now lets present the design of PQPRSG. Let $(Q, *)$ be a given quasigroup of arbitrary order, $l \in Q$ be a given leader, p be a positive integer and $\alpha = xxx\ldots x$ be an input sequence of length k, where $x \in Q$. The first sequence $\alpha_1 = x_1^{(1)}x_2^{(1)}x_3^{(1)}\ldots x_k^{(1)}$ is obtained as $PE_{p,l}(\alpha)$. The r^{th} sequence $\alpha_r = x_1^{(r)}x_2^{(r)}x_3^{(r)}\ldots x_k^{(r)}$ is obtained as

$$PE_{p,l}^{(r)}(\alpha) = PE_{p,l}(\alpha_{r-1}), \quad r = 1, 2, \ldots, n.$$

The output sequence of the PQPRSG for given quasigroup $(Q, *)$, p, l and α is the sequence $\alpha_n = x_1^{(n)}x_2^{(n)}x_3^{(n)}\ldots x_k^{(n)}$ which is obtained after n applications of PE-transformation.

It is reasonable to expect that not all quasigroups provide the same period of the sequences produced by PQPRSG. In other words, some quasigroups will produce sequences with greater period than others, depending on their structure and properties. Since the classifications of quasigroups are based on the structure and the properties of the quasigroups, we expect that some classes of quasigroups are more suitable for using in the PQPRSG than others. Therefore, an analysis of the classifications of quasigroup is needed to determine which quasigroups will give best results about the period of the generator.

In the next section we present experimental results about the period of pseudo random sequences produced with PQPRSG using quasigroups of order 4 with different parastrophes.

4 Experimental Results about the Period of the PQPRSG

We made many experiments in order to see how the period grows with each application of the PE-transformation and how the choice of quasigroup and the leader affects the period of the output sequences of PQPRSG. In our experiments, we considered quasigroups of order 4, with different parastrophes. After analyzing the results of the experiments, we conclude that every output sequence produced with the PQPRSG has non-periodical subsequence. This non-periodical subsequence appears at the beginning of the sequences (called *non-periodical part*) and after it the elements of the sequence started periodically to repeat. The non-periodical part is greater or equal to the periodical part and it grows with each application of the PE-transformation. This means that we can obtain sequences with arbitrary long non-periodical part after enough applications of PE-transformation. In Fig. 1, we show the growth of the non-periodical part of the sequence, for 2 quasigroups (randomly chosen) and leaders.

As we expected, the experiments show that the choice of the quasigroup has great effect on the performances of the PQPRSG. Also, the experiments confirm that the choice of a leader is also important, since we obtain different results

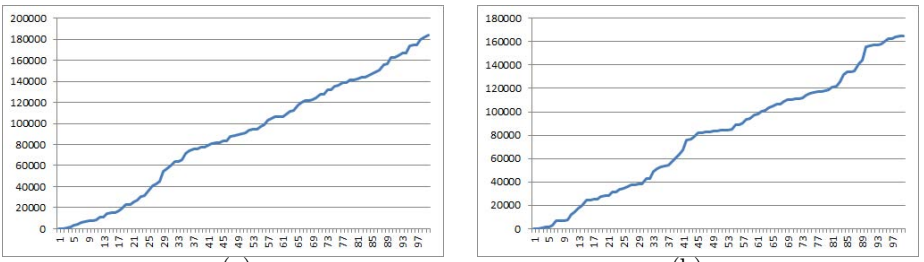

Fig. 1. Growth of the non-periodical part of the sequence produced with (a) quasigroup#300, leader 1 and (b) quasigroup#333, leader 3

when we use the same quasigroup and different leader. This effect can be seen from the results of the experiment showed in Table 4.

In Table 4 we give the length of the non-periodical parts of the output sequences produced with the PQPRSG, after 40 applications of the PE- transformation using the quasigroups with lexicographic number 150, 244, 351 and 420 for all of the leaders. As we can see from the table, the non-periodical parts produced with the same quasigroup differ from each other when different leader is used.

Table 4. Length of the non-periodical part of the sequences produced by PQPRSG using appropriate quasigroups and leaders

leader l	quasigroup#150	quasigroup#244	quasigroup#351	quasigroup#420
1	26532	32340	48552	36840
2	42369	37054	37876	39429
3	24382	35957	39317	60859
4	41920	35245	41950	49414

Since the PE-transformation is based on the parastrophes of a given quasigroup, it is reasonable to assume that the number of parastrophes of a quasigroup (used in the PQPRSG) has effect on the produced sequences. The results of the experiments show that when a quasigroup with more different parastrophes is used then we obtain a greater length of the non-periodical part of the produced sequences. This is showed in Fig. 2, where the length of the non-periodical part of the sequences is presented for quasigroups with 3 parastrophes and 6 parastrophes, after 100 applications of the PE-transformation. There all quasigroups were randomly chosen and the leaders are those which give best results for the suitable quasigroup.

As we can see in Fig. 2, the length of all non-periodical parts of the sequences produced with a quasigroup with 6 parastrophes is greater than suitable non-periodical parts when quasigroups with 3 parastrophes are used. The reason for

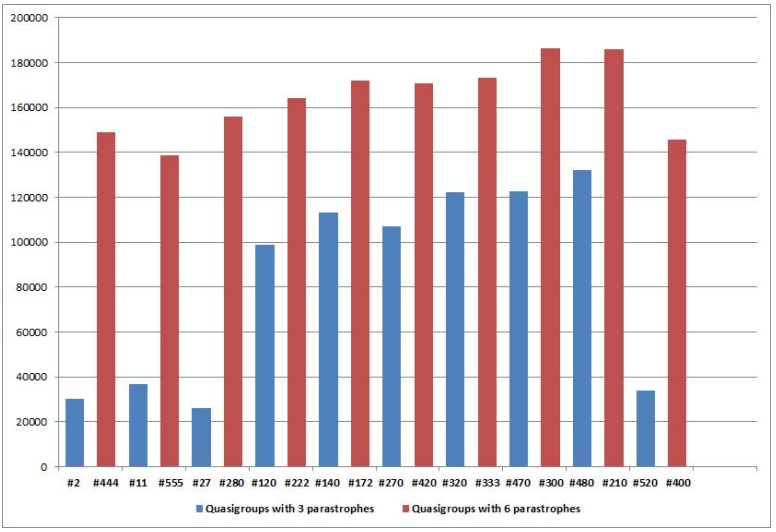

Fig. 2. Length of the non-periodical part of the sequence produced with quasigroups with 3 parastrophes and 6 parastrophes

this is the structure of the PE-transformation which represents the core of the PQPRSG.

As we mention before, fractal quasigroups are not good for cryptographic purposes since they produce regular structures. Therefore, when a fractal quasigroup is used, it is expected that the length of the non-periodical part of the produced sequences is smaller than the suitable length when a non-fractal quasigroup is used. Since PQPRSG uses PE-transformation for obtaining the sequences, we expect to obtain better results for the fractal parastrophic-non-fractal class of quasigroups, similar as the non-fractal quasigroups. We investigated the difference between the length of the non-periodical part of the sequences produced with fractal, non-fractal and fractal parastrophic-non-fractal quasigroups. The results are showed in Fig. 3.

In Fig. 3 we give the length of the sequences produced with a fractal, a fractal parastrophic-non-fractal and a non-fractal quasigroup, after 100 applications of the PE-transformation. Similarly as previous, all quasigroups were randomly chosen and the leaders are those which give best results for the suitable quasigroup.

We can see that the length of the non-periodical part of the sequences obtained with a fractal parastrophic-non-fractal quasigroup is approximately same as the length of the non-periodical part obtained with non-fractal quasigroup. This means that fractal parastrophic-non-fractal quasigroups can be used in producing sequences with PQPRSG. Also, we can see that the fractal quasigroups give relatively small non-periodical subsequences and therefore they are not good for using in the PQPRSG.

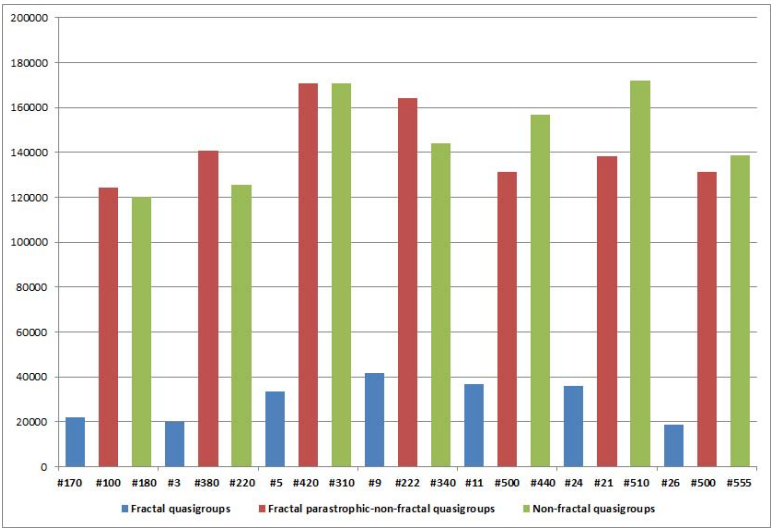

Fig. 3. Length of the non-periodical part of the sequence produced with arbitrary fractal, fractal parastrophic-non-fractal and non-fractal quasigroups

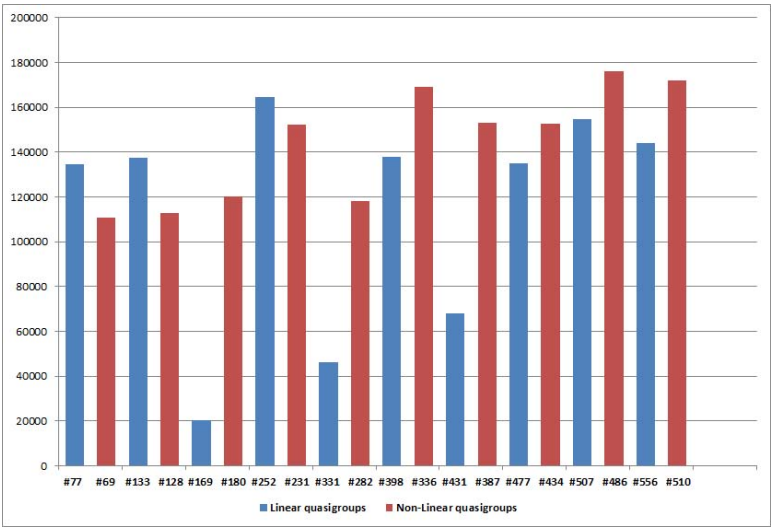

Fig. 4. Length of the non-periodical part of the sequence produced with arbitrary linear and non-linear quasigroups

We finish the statistical analysis of the PQPRSG with focus on the linear and non-linear quasigroups. The results from the experiments with linear and non-linear quasigroups show that almost always the sequences obtained with non-linear quasigroups have greater non-periodical part than the sequences obtained with linear quasigroups. This is showed in Fig. 4.

In Fig. 4 we show the length of the sequences produced with a linear and non-linear quasigroups, after 100 applications of the PE-transformation. There all quasigroups were randomly chosen and the leaders are those which give best results for suitable quasigroup.

From the results showed in Fig. 4 we conclude that the non-linear quasigroups are better for using in the PQPRSG than the linear quasigroups.

5 Conclusion

In this paper we introduced a design of PRSG, called PQPRSG, which is based on the parastrophic quasigroup transformation. We made many experiments in order to determine which quasigroups and leaders are best for designing of PQPRSG. The results obtained from the experiments showed that the choice of the quasigroup and leader has great effect on the sequences produced with the PQPRSG. We conclude that the quasigroups with 6 parastrophes give better results for non-periodical part of the produced sequences than the quasigroups with 3 parastrophes. From the analysis of the results we also concluded that the PQPRSG gave best results when fractal parastrophic-non-fractal or non-fractal quasigroups were used for the producing of the sequences. The same conclusion can be made for using the class of non-linear quasigroups. Using exhaustive verification, we proved that all linear quasigroups by Boolean representation of order 4 have linear parastrophes and non-linear quasigroups by Boolean representation of order 4 have non-linear parastrophes.

References

1. Dimitrova, V., Markovski, J.: On Quasigroup Pseudo Random Sequence Generators. In: Proceedings of the 1st Balkan Conference in Informatics, Thessaloniki, Greece, pp. 393–401 (2003)
2. Bakeva, V., Dimitrova, V., Popovska-Mitrovikj, A.: Parastrophic Quasigroup String Processing. In: Proceedings of the 8th Conference on Informatics and Information Technology with International Participants, Bitola, Macedonia, pp. 19–21 (2011)
3. Dimitrova, V., Bakeva, V., Popovska-Mitrovikj, A., Krapež, A.: Cryptographic Properties of Parastrophic Quasigroup Transformation. In: Markovski, S., Gushev, M. (eds.) ICT Innovations 2012. AISC, vol. 207, pp. 235–243. Springer, Heidelberg (2013)
4. Gligoroski, D., Dimitrova, V., Markovski, S.: Quasigroups as Boolean functions, their equation systems and Groebner bases. In: Groebner Bases, Coding, and Cryptography, pp. 415–420. Springer (2009) ISBN 978-3-540-93805-7
5. Knuth, D.: The Art of Computer Programming, vol. 2. Addison-Wesley (1977)
6. Stinson, R.D.: Cryptography -Theory and Practice. CRC press (1995)
7. Markovski, S., Gligoroski, D., Bakeva, V.: Quasigrouop string processing: Part 1. In: Contributions, Sec. Math. Tech. Sci., MANU, XX 1-2 13–28 (1999)
8. Bakeva, V., Popovska-Mitrovikj, A., Dimitrova, V.: Resistance of Statistical Attacks of Parastrophic Quasigroup Transformation, arXiv: 1404.0781v1 (cs.CR) (2014), `http://arxiv.org/pdf/1404.0781v1.pdf`
9. Dimitrova, V., Markovski, S.: Classification of quasigroups by image patterns. In: Proceedings of the 5th International Conference for Informatics and Information Technology, Bitola, Macedonia, pp. 152–160 (2007)

Optimizing ELARS Algorithms Using NVIDIA CUDA Heterogeneous Parallel Programming Platform

Vedran Miletić, Martina Holenko Dlab, and Nataša Hoić-Božić

University of Rijeka, Department of Informatics,
Radmile Matejčić 2, 51000 Rijeka, Croatia
{vmiletic,mholenko,natasah}@inf.uniri.hr

Abstract. Scalability is an important property of every large-scale recommender system. In order to ensure smooth user experience, recommendation algorithms should be optimized to work with large amounts of user data. This paper presents the optimization approach used in the development of the E-learning activities recommender system (ELARS). The recommendations for students and groups in ELARS include four different types of items: Web 2.0 tools, collaborators (colleague students), optional e-learning activities, and advice. Since implemented recommendation algorithms depend on prediction of students' preferences, algorithm that computes predictions was offloaded to graphics processing unit using NVIDIA CUDA heterogeneous parallel programming platform. This offload increases performance significantly, especially with large number of students using the system.

Keywords: e-learning, recommender system, ELARS, algorithm optimization, heterogeneous paralell programming, NVIDIA CUDA.

1 Introduction

Recommender systems support users in identifying services in information-rich environments. These systems can provide a solution for the information overload problem by recommending items that are potentially useful for the target user or that are within the scope of his/her interests [1]. In addition, recommender systems ensure personalization since recommendations are generated according to user's characteristics. Therefore, recommender systems are often used across different domains like entertainment industry, e-commerce and e-learning [4].

The usefulness of items (utility) that can be recommended is expressed as a numerical value (rating). This value is determined by the user or it can be predicted. Accordingly, the recommendation problem comes down to the prediction of the unknown utility values in order to recommend item or items with the highest utility to the target user. Prediction algorithms are usually based on two commonly used methods: collaborative filtering and content-based recommendations [1,4]. Performance of the recommender system depends on the accuracy or precision of the prediction algorithm. Therefore, appropriate algorithm should

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_14

be chosen based on experiment in which a few algorithms are compared using some evaluation metric. Another important property of the recommender system that may affect user experience is scalability. The system should be able to scale well, ideally both horizontally and vertically, to keep up with increasing number of users navigating through ever-enlarging collections of items [18], producing large amount of data to be analyzed. ELARS, described below, is an example of such a system, and our research described in this paper is on scaling ELARS for a large number of users.

Horizontal scaling, or scaling out, implies adding more nodes to a distributed system, in this case adding more servers to a cluster running the application [17]. Vertical scaling, or scaling up, means adding resources to a single node in the system, in this case adding additional central processing units (CPUs), graphics processing units (GPUs), memory etc. When required to scale, one can opt for horizontal or vertical scaling, or combine both. There are advantages to each of two approaches. From the programming standpoint, vertical scaling is simpler, but tends to be limited by hardware specifications (i. e. one can only fit a limited number of CPUs and GPUs or limited amount of memory in a single node). Horizontal scaling enables addition of more resources. However, it requires a more complex programming model that may or may not fit a particular application. Also, larger numbers of nodes in a distributed system implies increased management complexity.

This paper presents our approach to performance optimization of ELARS algorithms using NVIDIA CUDA heterogeneous parallel programming platform enabling code to run on both GPU(s) and CPU(s). We found the preference prediction to be particularly demanding in terms of computation time. Suitable parts of preference prediction algorithms are moved to GPU for execution to improve overall performance. Meanwhile, CPU executes the parts not suitable for the GPU. Our approach utilizes a single node and allows vertical scaling with introduction of faster CPUs and GPUs. Future server systems running web application will be increasingly heterogeneous, with its computation power divided over a number of different processors [12]. Our approach contributes to promotion of heterogeneous parallel programming usage in web application development. To the best of our knowledge, this is also first application of NVIDIA CUDA in domain of e-learning.

The paper is organized as follows: first we present ELARS, then we introduce heterogeneous parallel programming approach with focus on NVIDA CUDA, then we describe our approach to algorithm parallelization. We do performance benchmarks, and conclude along with possible directions for future work.

2 E-Learning Activities Recommender System

E-learning Activities Recommender System – ELARS [10] supports collaborative e-learning activities in an online learning environment that consists of a learning management system (LMS) and 10 different Web 2.0 tools [3]. In such environment, students use LMS to study the learning content, read the instructions,

solve online tests, communicate with others and similar. They use Web 2.0 tools for realization of e-learning activities like seminar writing, mind-mapping or WebQuests. Students use recommender system in parallel with other components of the learning environment to choose between recommended items.

2.1 Recommendation Algorithms in ELARS

The system provides personalization by recommending optional Web 2.0 tools, collaborators, optional e-learning activities (e-tivities), and offering advice to students and groups. Recommendations are based on several students characteristics [11]: preferences of Web 2.0 tools, preferences of learning styles, knowledge level and activity level. Preferences of Web 2.0 tools and learning styles are collected via questionnaires at the beginning of the course. Knowledge level is determined based on student's results on online tests and activity level, which represents quantity and continuity of student's (group's) contributions in e-tivities, is calculated based on activity traces collected from Web 2.0 tools.

From the scalability point of view, the most challenging task of the recommendation process is to predict not known Web 2.0 tools preferences that are used for calculating utility of potential collaborators, offered Web 2.0 tools or optional e-tivities. Web 2.0 tools preferences are predicted using hybrid approach [16]. Recommender switches between collaborative filtering and content-based recommendations based on the number of known preferences in the system's database. Using collaborative filtering technique preference of the target student for target tool is predicted based on preferences of similar students (nearest neighbours) for target tool [1]. n case nearest neighbours cannot be found, content-based recommendations technique [16] is used and preference is predicted according to target student's preferences for other tools.

2.2 System Performance Bottlenecks

Since the number of system's users is much bigger than the number of tools included in the learning environment, performance bottleneck was found in case of collaborative filtering. This technique is performed in two phases which can both be addressed using GPU: neighbourhood selection and target tool predicted preference computation.

Neighbourhood Selection. When performing collaborative filtering the recommender system uses knowledge about similar users' preferences regarding target tool [1]. Therefore, similarity of the target student with students for which target tool preference exists in the system database is calculated. Similarity between students is determined based on known preferences or learning styles preferences according to VARK model [6]. In that process we calculate cosine similarity [15] which is commonly used metric for collaborative filtering. To form the neighbourhood, k students who are the most similar to the target student are selected. The configuration parameter k is set to 20 using cross-validation.

Predicted Preference Computation. Preference value pp is predicted based on normalized preference values for nearest neighbours tp'_i [1] using formula 1. Normalization of neighbours' preferences using mean-centering method is performed to unify the criteria on the basis of which neighbours expressed their preference. Besides normalization, to increase the accuracy of prediction algorithm weighting factors w_i are used. Values w_i represent the similarity of neighbour i with target student. By using such weights the influence of preference from certain neighbour to the result value is bigger if he/she is more similar to the target student. This effect is further improved by introducing amplification factor $\alpha = 2$. The resulting value is normalized using expression in the denominator and added to the target student's mean preference $\overline{tp}$.

$$pp = \overline{tp} + \frac{\sum_{i=1}^{k} w_i^{\alpha} tp'_i}{\sum_{i=1}^{k} w_i^{\alpha}} \tag{1}$$

3 Heterogeneous Parallel Programming

Usage of graphics processors for general-purpose computing started with programmable shaders on the NVIDIA GeForce FX and AMD Radeon series of graphics cards in early 2000s [20]. Shaders were programmed using either High-level shading language (HLSL) from Microsoft DirectX, OpenGL Shading Language (GLSL), or NVIDA Cg. Despite the requirement to significantly change the algorithms to adapt them for graphics processing unit (GPU), programming non-graphics problems using shaders became popular soon afterwards.

NVIDIA recognized the potential for the utilization of the GPU for general purpose calculations, and with GeForce 8 series opened up the GPU using a custom application programming interface (API) named Compute Unified Device Architecture, or CUDA for short [13]. CUDA has been made available to the public in February 2007, and is supported by all NVIDIA graphics processors released since.

Despite the appearance of the open standard named OpenCL which has the same purpose as (proprietary) CUDA, CUDA and therefore NVIDIA continues to dominate the market. Beside being first to appear, it is also due to greater amount of literature available and better programming tools. While both standards are very similar, they are not compatible [5].

3.1 Related Research Efforts

GPUs have so far been used to solve problems in bioinformatics, chemistry, physics, mathematics, medicine, mechanical engineering, electrical engineering, computer science, and other science and engineering disciplines. Garland et. al. survey applications of CUDA to a diverse set of data parallel problems, finding varying speedups of GPU-enabled algorithms vs CPU-only versions depending on the algorithm properties [8]. Gregg and Hazelwood, as well stress that any benchmarks that lead to conclusions on speedup should provide information on

where the data is assumed to be, because copying data from CPU memory to GPU memory and back can take a significant amount of time [9]. It also is worth noting applications of CUDA that could be used in recommender systems; Garcia, Debreuve, and Barlaud implement brute force k nearest neighbours selection using CUDA C and conclude that GPU speedup can be up to 120 times compared to equivalent CPU code implemented in C [7], including data copies from CPU to GPU and back in computation time.

3.2 GPU Architecture and CUDA Programming Model

Single instruction, multiple data (SIMD) is a class of parallel processors that have a larger number of processing elements that can do the same operation on multiple data simultaneously. This feature exploits data-level parallelism. GPUs, unlike most central processing units (CPUs), are SIMD processors, which allows acceleration of suitable algorithms. Performance gains vary greatly, and can be anything from a couple of percent to one or even two orders of magnitude.

From now on, we focus solely on programming NVIDIA GPUs using CUDA programming language. CUDA began as an extension of programming languages C/C++ and Fortran. Special directives were added to both languages that allowed to offload parts of computation to GPU.

CUDA API exposes threads, which are grouped in blocks of threads, which are again grouped in grid of blocks. Each block allows indexing in three dimensions, while the number of dimensions for grid is limited to two. This programming model is intended to fit multidimensional arrays. Functions written in CUDA intended to run on the GPU are named kernels. On each kernel call, the number of blocks and threads on which the kernel will be executed is specified. Therefore each kernel can be written for data of varying shape and size.

3.3 CUDA Libraries

CUDA ecosystem offers a number of libraries that simplify programming and even provide highly optimized versions of frequently used algorithms, for example reductions and sorting. In this work we use Thrust [2] and PyCUDA [14] which we describe below.

Thrust. Thrust is a C++ template library for CUDA based on C++ Standard Template Library. Thrust offers a number of data parallel primitives such as scan, sort, and reduce. These primitives can be used along with existing C++ code to ease offloading parts of code onto the GPU and enable rapid prototyping of CUDA applications.

PyCUDA. PyCUDA is a Python module that enables programmers to access CUDA API. Due to Python's clean syntax, it is very suitable for prototyping software. With PyCUDA, it becomes possible to also for prototype software that uses CUDA. In addition, PyCUDA enables access to existing CUDA C/C++ libraries such as Thrust.

4 Algorithm Parallelization Approach

We found Python and PyCUDA to be very suitable for rapid prototyping and comparison of different approaches to parallelization. We ported preference prediction code from existing C# implementation to Python, utilizing NumPy [19]. NumPy is a Python module providing high-level interface to C-like arrays for efficient numerical computation. Large parts of NumPy are implemented in C and Fortran; this in and of itself resulted in significant performance improvement in implementation of our algorithms done in Python and NumPy compared to implementation done in C#. During prototyping stage we also simplified the resulting program by caching data which is normally retrieved from database.

4.1 Neighbourhood Selection Parallelization

Since number of potential neighbours grows with increase in number of students – system's users, neighbourhood selection is a performance bottleneck. For example, if the ELARS was deployed at University of Rijeka which has nearly 20000 students, a popular Web 2.0 tool could easily have 10000 or even 15000 entered preferences.

Brute force search is used to select nearest neighbours. Thrust function `thrust::sort_by_key()`, which sorts key-value pairs, was a good fit for GPU version of the algorithm. Somewhat counter-intuitively, key set is similarity expressed as floating-point number, while the value set consists of student IDs. Since neither hashing nor numerical operations are done using these floating-point numbers, usage of floats here does not lead to problems. In each iteration key and value arrays are copied to the GPU, and after sorting both arrays are retrieved from the GPU.

4.2 Predicted Preference Computation Parallelization

To optimize predicted preference computation, we opted to compute all the predictions in a single kernel execution. To achieve it, learning similarities matrix is copied to the GPU on program initialization. After that, pairs of target students and tools are collected and stored in a 2D array, as well as normalized target tool preferences for all pairs. Normalized target tool preferences are computed on the CPU. Since it is required to select only non-null values from the array, we expect such selection done on GPU would slow down computation significantly and compensate for potential benefits of parallelization.

Both arrays are copied to the GPU, and computation of sum elements for nominator and denominator of formula 1 is done. Sum reduction can then be done on either CPU or GPU. If reduction is done on the CPU, whole array is copied back; if it is done on GPU, only a single resulting value is copied back.

5 Performance Measurements

A system with AMD FX-8150 8-core CPU and NVIDIA GeForce GTX 660 GPU was used for testing and benchmarking. We should emphasize that neither 64-bit

floating point precision nor large amounts of GPU memory are required in this domain, so commodity GeForce GPUs can be used as well as more expensive ones from Tesla and Quadro series.

We measure the computation time required to get student's tool preference for all students for all tools, that is, to predict all unknown preference values. The dataset we used for testing has 57,5% unknown preferences. Dataset is loaded from flat text files. We measure computation time for 10 tools for 640, 1280, 2560, 5120, 10240, 20480 students. Computation time for each number of students for CPU-only and CUDA-enabled GPU and CPU code is shown in Figure 1. CPU code in both cases uses no parallelization and runs on a single CPU core.

On the GPU side, we should note that both PyCUDA GPU Array module and Thrust library dynamically determine number of blocks and threads to be used for computation depending on data and GPU used, without needing to be manually specified by user.

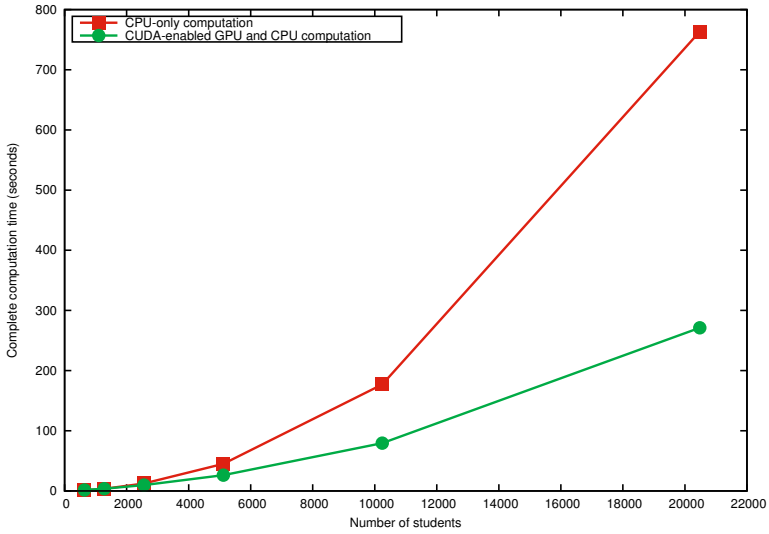

Fig. 1. Performance measurements for entire algorithm

We can observe that for smaller number of students (640, 1280, and 2560) CPU-only and CUDA-enabled code are very close in terms of computation time. At 5120 students they start to visibly diverge, and difference becomes even greater in cases of 10240 and 20480 students. We can see that the gap widens with increasing number of students, arriving at nearly 3 times the speedup in favor of GPU at 20480 students.

Total execution time of sorting in k nearest neighbours selection, including copying of data from CPU to GPU memory and back, is shown in Figure 2. We can observe the exponential increase in computation time for the CPU, while the

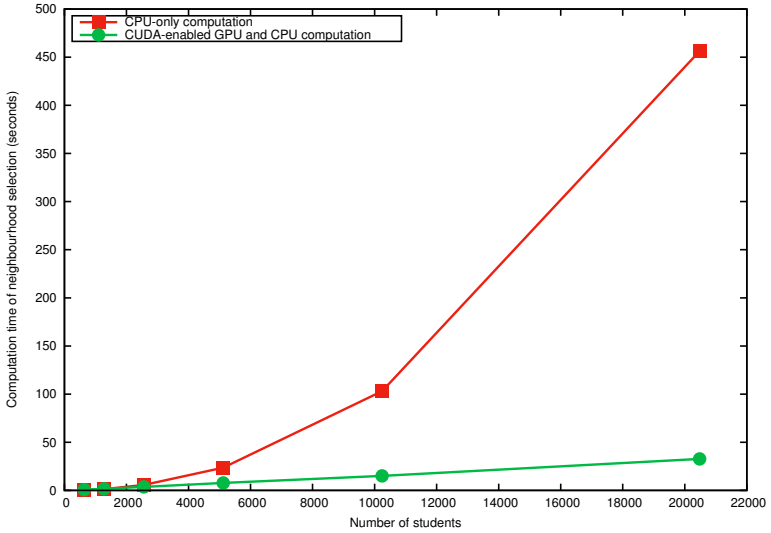

Fig. 2. Performance measurements for sorting in k nearest neighbours selection

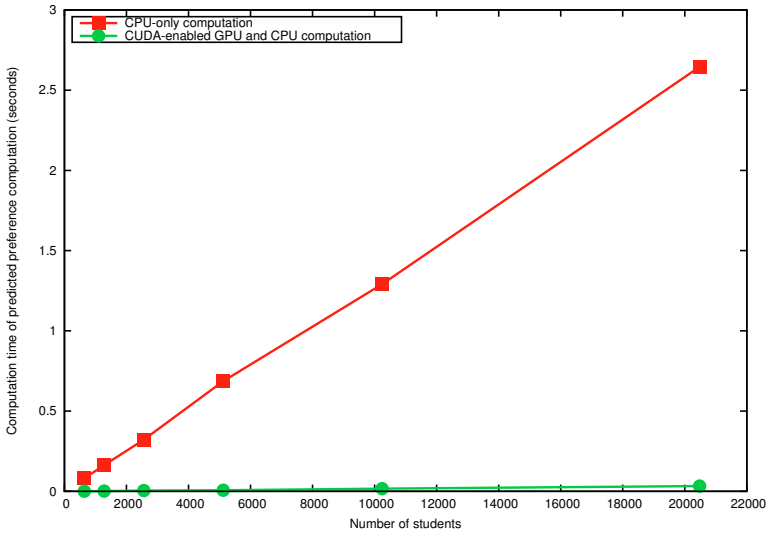

Fig. 3. Performance measurements for target tool predicted preference computation

for the GPU the increase remains linear for the number of students we measured. Difference in computation time starts to be apparent with 3 times GPU speedup over CPU at 5120 students, increase to nearly 7 times at 10240, and finally ends up being over 14 times at 20480 students. It is also apparent that k nearest neighbours dominates the computation time of the entire preference prediction algorithm in larger cases, taking nearly over half of computation time.

Total execution time of target tool predicted preference computation, again including copying of data from CPU to GPU memory and back, is shown in Figure 3. We can see how both CPU and GPU computation time increases linearly with the number of students, but GPU computation time consistently remains around two orders of magnitude smaller. In other words, we get GPU speedup over CPU between 75 and 100 times.

6 Conclusions, Discussion and Future Work

We presented an approach to optimization of recommender system performance by offloading parts of algorithms to GPUs. We find this approach to be reasonable considering the expectation that future machines will be increasingly heterogeneous, and their computing power will be divided across a range of chips with different characteristics targeting certain kinds of problems. We found the optimizations we used to improve performance significantly, and allow scaling for a larger number of users.

While CUDA dominates the market at present, it is realistic to expect that in the future OpenCL play a significant role. Intel and AMD, both supporting OpenCL, already ship most of their CPUs with integrated GPU, and recently with AMD Berlin APU making into the Opteron line of processors this trend moved this in the server domain as well. Both performance and power consumption of GPUs and APUs has the potential to make them an attractive choice in the server environment, even outside the usual scientific computing applications.

With these technology developments in mind, our other future plans include offloading even larger amounts of computation on the GPU, porting of CUDA-enabled code from Python to C# to ease integration within ELARS, and finally deployment in production at University of Rijeka.

Acknowledgments. The research has been conducted under the project "E-learning Recommender System" (reference number 13.13.1.3.05) supported by University of Rijeka (Croatia). Performance benchmarks were done at University CUDA Teaching Center provided by NVIDIA.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (2005)
2. Bell, N., Hoberock, J.: Thrust: A productivity-oriented library for cuda. GPU Computing Gems 7 (2011)
3. Dlab, M.H., Hoić-Božić, N.: An approach to adaptivity and collaboration support in a web-based learning environment. International Journal of Emerging Technologies in Learning (2009)
4. Drachsler, H., Hummel, H., Koper, R.: Identifying the goal, user model and conditions of recommender systems for formal and informal learning. Social Information Retrieval for Technology Enhanced Learning 10(2), 4–24 (2009)

5. Du, P., Weber, R., Luszczek, P., Tomov, S., Peterson, G., Dongarra, J.: From cuda to opencl: Towards a performance-portable solution for multi-platform gpu programming. Parallel Computing 38(8), 391–407 (2012)
6. Fleming, N.D.: I'm different; not dumb. modes of presentation (vark) in the tertiary classroom. In: Research and Development in Higher Education, Proceedings of the 1995 Annual Conference of the Higher Education and Research Development Society of Australasia (HERDSA), vol. 18, pp. 308–313 (1995)
7. Garcia, V., Debreuve, E., Barlaud, M.: Fast k nearest neighbor search using gpu. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2008, pp. 1–6. IEEE (2008)
8. Garland, M., Grand, S., Nickolls, J., Anderson, J., Hardwick, J., Morton, S., Phillips, E., Zhang, Y., Volkov, V.: Parallel computing experiences with cuda. IEEE Micro Magazine 28(4), 13–27 (2008)
9. Gregg, C., Hazelwood, K.: Where is the data? why you cannot debate cpu vs. gpu performance without the answer. In: 2011 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), pp. 134–144. IEEE (2011)
10. Holenko Dlab, M., Hoić-Božić, N.: Recommender system for web 2.0 supported elearning. In: 2014 IEEE Global Engineering Education Conference (EDUCON) Proceedings (2014)
11. Holenko Dlab, M., Hoić-Božić, N., Mezak, J.: Personalizing e-learning 2.0 using recommendations. In: The Proceedings of MIS4TEL Conference (2014)
12. Keckler, S.W., Dally, W.J., Khailany, B., Garland, M., Glasco, D.: Gpus and the future of parallel computing. IEEE Micro 31(5), 7–17 (2011)
13. Kirk, D.: Nvidia cuda software and gpu parallel computing architecture. In: ISMM, vol. 7, pp. 103–104 (2007)
14. Klöckner, A., Pinto, N., Lee, Y., Catanzaro, B., Ivanov, P., Fasih, A.: Pycuda and pyopencl: A scripting-based approach to gpu run-time code generation. Parallel Computing 38(3), 157–174 (2012)
15. Lee, L.: Measures of distributional similarity. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 25–32. Association for Computational Linguistics (1999)
16. Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., Koper, R.: Recommender systems in technology enhanced learning. In: Recommender Systems Handbook, pp. 387–415. Springer (2011)
17. Michael, M., Moreira, J.E., Shiloach, D., Wisniewski, R.W.: Scale-up x scale-out: A case study using nutch/lucene. In: IEEE International Parallel and Distributed Processing Symposium, IPDPS 2007, pp. 1–8. IEEE (2007)
18. Shani, G., Gunawardana, A.: Evaluating recommendation systems. In: Recommender Systems Handbook, pp. 257–297. Springer (2011)
19. Van Der Walt, S., Colbert, S.C., Varoquaux, G.: The numpy array: a structure for efficient numerical computation. Computing in Science & Engineering 13(2), 22–30 (2011)
20. Wu, E., Liu, Y.: Emerging technology about gpgpu. In: IEEE Asia Pacific Conference on Circuits and Systems, APCCAS 2008, pp. 618–622. IEEE (2008)

Automated Synthesis of Initial Conceptual Database Model Based on Collaborative Business Process Model

Drazen Brdjanin, Goran Banjac, and Slavko Maric

University of Banja Luka, Faculty of Electrical Engineering,
Patre 5, 78000 Banja Luka, Bosnia and Herzegovina
{bdrazen,goran.banjac,ms}@etfbl.net

Abstract. This paper presents an approach to automated design of the initial conceptual database model. The source model is a collaborative business process model represented by BPMN, while the target model is represented by a UML class diagram. Automated synthesis of the target model is driven by typical business process patterns and includes automatic extraction of data objects, message flows and business process participants, as well as automatic generation of corresponding classes and their associations. Application of the implemented ATL-based generator is illustrated on a real business process model.

Keywords: ATL, BPMN, Collaborative Business Process Model, Conceptual Database Model, Model-driven, UML.

1 Introduction

The data model constitutes one of the most important artifacts in the information system design process, as well as the crucial component of software system models. Consequently, the automatization of data model design has been the subject of research for many years. The idea of MDSDM[1] is more than 25 years old [1]. Although the first papers reporting the model-driven tools for (semi)automatic synthesis of the data model were published in the mid-1990s, the fully automatic MDSDM is still the subject of intensive research. In the existing literature there are only a small number of papers presenting the implementation of the automatic model-driven generator of the target data model with the corresponding evaluation results, while the great majority of papers only present modest achievements in (semi)automated, or even manual, data model synthesis.

In this paper we are focussed on automated synthesis of the initial CDM[2], i.e. semantic data model containing abstractions of persistent entities and their relationships in the entire system. We consider a collaborative BPM[3], represented

[1] Model-Driven Synthesis of Data Model.
[2] Conceptual Database Model.
[3] Business Process Model.

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_15

by BPMN[4], as the starting point for automated generation of the target model, which is represented by UML[5] class diagram.

The paper is structured as follows. After the introduction, the second section presents the related work. The third section gives an overview of source and target models. The fourth section presents an analysis of the semantic capacity of typical business process patterns in collaborative BPMs, and provides the rules for automated generation of the initial CDM. Application of the implemented generator is illustrated in the fifth section. Finally, we conclude the paper and highlight the directions for further research.

2 Related Work

The survey [1] shows that current MDSDM approaches, depending on the source notation, can be classified as: *function-oriented*, *process-oriented*, *communication-oriented*, and *goal-oriented*. POMs[6] constitute the largest category of models being used as a starting point for the MDSDM. The boom of these approaches was influenced by the appearance of metamodel-based notations, particularly the UML activity diagram and BPMN, as well as the ATL[7] and QVT[8] transformation languages. The survey [1] shows that the semantic capacity of POMs has not yet been sufficiently identified to enable the automatic synthesis of the complete data model, since the existing approaches still do not have significant *precision* (the percentage of correct automatically generated concepts) and *recall* (automatically generated percentage of the target model) in the automated generation of some types of associations and class members.

The BPMN, as the starting base for MDSDM, is used in [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]. There are two QVT-based proposals [10, 13], but with modest achievements in the automated generation of the analysis level class diagram, as well as several proposals [8, 9, 12, 16, 18] for the semiautomated generation. All proposals are based on incomplete source model (i.e. BPM containing a single diagram, although a real model contains a finite set of diagrams representing all business processes in a domain). An overview of BPMN-based MDSDM approaches is given in Fig. 1.

Rungworawut and Senivongse in [7] propose some heuristic guidelines for extraction of classes and archetype patterns for synthesis of associations. However, these guidelines are not suitable for automated target model synthesis. Some of them are implemented in [8].

Brambilla et al. in [9, 12] assume that domain classes are already identified, and they extend such initial domain model by adding generic metamodels of users and processes as well as corresponding classes for activities (subclasses of corresponding metaclass from the process metamodel) in the source BPM. They

[4] Business Process Model and Notation [2].
[5] Unified Modeling Language [3, 4].
[6] Process-oriented models.
[7] ATLAS Transformation Language [5].
[8] Query/View/Transformation [6].

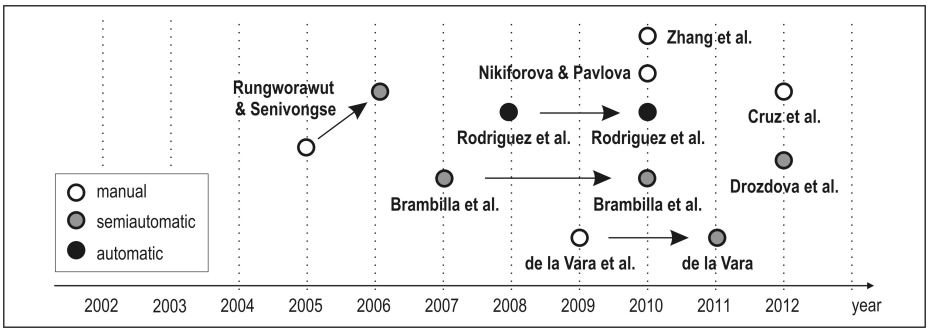

Fig. 1. Overview of BPMN-based MDSDM approaches

do not propose rules for synthesis of associations between domain classes and classes that represent activities. Since the structure of the generated model is not optimal, they propose some additional mechanisms for the reduction of classes, and present the corresponding semiautomatic tool in [12].

Rodriguez et al. in [10, 13] also take a BPMN-represented source BPM, but propose its mapping into the corresponding UML activity diagram and target data model synthesis based on UML activity diagrams.

De la Vara et al. in [11] propose some guidelines for class diagram synthesis based on BPMN and additional textual specifications of information flows in BPM. Based on these guidelines, de la Vara in [16] proposes and partially implements informal rules for synthesis of classes and associations. However, the proposed set of rules enable only a semiautomatic data model synthesis.

Cruz et al. in [17] propose: (i) mapping of data stores and participants into the corresponding classes, (ii) synthesis of *participant-object* associations with appropriate cardinalities (1:* and *:*) between the corresponding classes, and (iii) identification of class members based on the additional textual specifications of the corresponding data stores. The proposed approach is not implemented.

A BPMN-based BPM, as the starting base for MDSDM, is also considered in [14, 15, 18], but without implementation.

3 Overview of Source and Target Models

3.1 Source Model

In this paper we consider a collaborative BPMN diagram (collaborative BPM) as a source model. Related excerpt from the metamodel [2] is shown in Fig. 2.

A collaborative BPM contains two or more pools (`Participant`). A pool can contain lanes (`Lane`) which are often used for modeling internal roles. Let P be a set of all pools and lanes (in the rest of the paper, shortly referred to as *participants*) in a source BPM, where P_P subset contains all pools, and P_L subset contains all lanes. A `MessageFlow` (message exchange between participants) connects two pools (or objects within the pool). Let M be a set of all message flows in a source BPM.

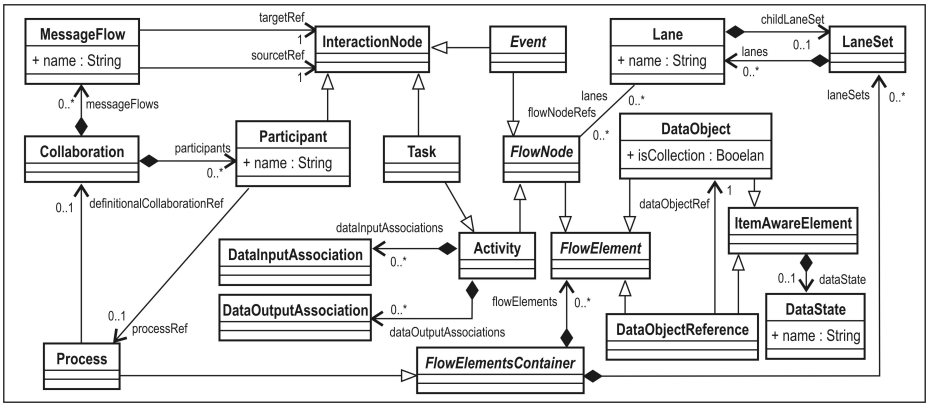

Fig. 2. BPMN metamodel [2] excerpt for collaborative BPM representation

Participants perform tasks (`Task`). Let T be a set of all tasks in a source BPM. A task can have inputs and outputs. Inputs can be represented by (process) `DataInput`, `DataObject` or `DataStore` elements. Outputs can be represented by `DataOutput`, `DataObject` or `DataStore` elements. Let O be a set of all objects in a source BPM. Input objects and tasks are connected via data input associations (`DataInputAssociation`). Tasks and output objects are connected via data output associations (`DataOutputAssociation`). `DataInput`, `DataOutput`

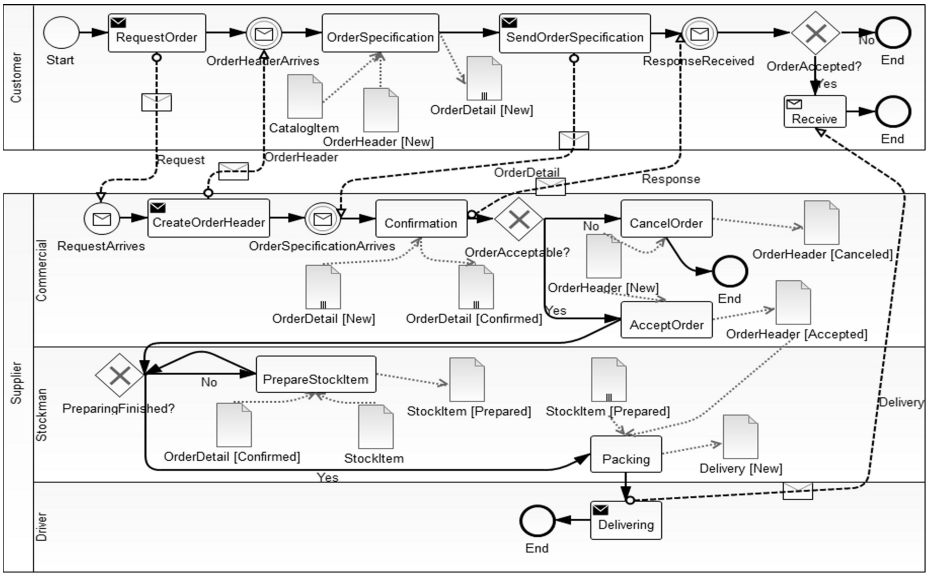

Fig. 3. Sample collaborative BPM used in the paper as the source model

and `DataObject` elements can represent a single object or a collection of objects. `DataStore` can have unlimited capacity, or exact (specified) capacity.

Figure 3 depicts the sample BPM that will be used in this paper as the source model. As a real model of order processing, it is sufficiently illustrative to cover the most important concepts and basic rules for automated CDM synthesis.

3.2 Target Model

We use the UML class diagram to represent the CDM (related excerpt from the UML infrastructure [3] is shown in Fig. 4). Let E and R be sets of *classes* and their *associations* in the target CDM. Since we are currently focused on automated generation of proper structure of the target model: (i) each generated class $e \in E$ will (if necessary) contain only one `ownedAttribute` named `id`, which represents a primary key, and (ii) each generated association $r \in R$ will be a binary association, whose two `memberEnd` attributes represent `source` and `target` association ends with the appropriate multiplicities.

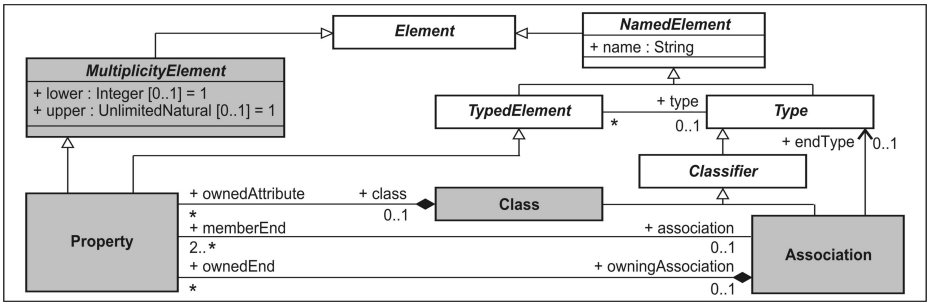

Fig. 4. UML metamodel [3] excerpt for CDM representation

4 Automated CDM Synthesis

The survey [1] shows that the semantic capacity of BPM has not yet been sufficiently identified to enable the automatic synthesis of the complete target data model. In this paper we follow the approach [19] for automated CDM synthesis based on BPM represented by a UML activity diagram.

4.1 Automated Generation of Classes

There are three important bases for automated generation of classes in the target CDM: (i) *participants*, (ii) *objects*, and (iii) *activations of existing objects*.

Participants. According to [19], each *participant* from the source BPM is to be mapped into the corresponding class of the same name in the target CDM. In this way, a set E_{PP} of classes generated for all pools $p_p \in P_P$ is as follows:

$$E_{PP} = \Big\{ e_P \in E \;\Big|\; e_P = T_{PP}(p_p),\ p_p \in P_P \Big\},$$

where the T_{PP} rule, that maps pool p_p into the corresponding class e_P, is:

$$T_{PP}(p_p) \stackrel{def}{=} e_P \mid \big(name(e_P) = name(p_p)\big).$$

Lanes are to be mapped into the corresponding classes in the target CDM, as well. Since some lanes that belong to different pools may have the same names, the naming of the generated classes should be different, e.g. concatenation of the parent pool name and the given lane name. In this way, a set E_{PL} of classes generated for all lanes $p_l \in P_L$ is as follows:

$$E_{PL} = \Big\{ e_L \in E \mid e_L = T_{PL}(p_l),\ p_l \in P_L \Big\},$$

where the T_{PL} rule, that maps lane p_l (belonging the pool p_p) into class e_L, is:

$$T_{PL}(p) \stackrel{def}{=} e_L \mid \big(name(e_L) = concat(name(p_p), name(p_l))\big).$$

A set E_P of classes generated for all participants is $E_P = E_{PP} \cup E_{PL}$.

Objects. During the execution of a business process, participants perform tasks. Each task may have a number of input and/or output objects that can be in different states. According to [19], each different type of objects from the source BPM is to be mapped into the corresponding class of the same name in the target CDM. In a collaborative business process, participants exchange messages, as well. Due to the similar semantics of objects and messages, the same rule is to be applied for both objects and message flows. In this way, a set E_O of classes generated for all *objects* $o \in O \cup M$ is as follows:

$$E_O = \Big\{ e_O \in E \mid e_O = T_O(o),\ o \in O \cup M \Big\},$$

where the T_O rule, that maps object o into class e_O, is:

$$T_O(o) \stackrel{def}{=} e_O \mid \big(name(e_O) = name(o)\big).$$

Activations of Existing Objects. An *activation* represents the fact that some *existing*[9] object(s) constitute(s) the input in some task that changes its/their state. In the sample BPM, the `PrepareStockItem` task represents the activation of the `StockItem` object(s). As suggested in [20], class representing activated objects are to be named by concatenation of the object name and state name.

Let X_A be a set of *activations* $\langle p, t, o \rangle$ in the source BPM, where $p \in P$ is participant performing task $t \in T$ on existing object $o \in O$. Total set E_A of classes generated for all activations $\langle p, t, o \rangle \in X_A$ is:

$$E_A = \Big\{ e_A \in E \mid e_A = T_A(\langle p, t, o \rangle),\ \langle p, t, o \rangle \in X_A \Big\},$$

where the T_A rule, that maps activation $\langle p, t, o \rangle$ into class e_A, is:

$$T_A\big(\langle p, t, o \rangle\big) \stackrel{def}{=} e_A \mid \big(name(e_A) = concat(name(o), state(o))\big).$$

[9] Existing objects are objects that are not created in the given business process, but in some other business process. In the sample BPM, objects of the `StockItem` type, as well as objects of the `CatalogItem` type, are existing objects.

Finally, total set E of classes in the target CDM is $E = E_P \cup E_O \cup E_A$.[10]

4.2 Automated Generation of Associations

We distinguish three different kind of class associations in the target CDM: (i) *participant-participant* associations, (ii) *participant-object* associations, and (ii) *object-object* associations.

***Participant-Participant* Associations.** These associations are originated by the fact that some pool may contain several lanes representing different business roles. Each role is performed by a particular participant. With the course of time, the same role may/will be performed by many different participants. Hence, each role is an abstraction of a number of entities of the same type, and all of them belong to the same pool. This implies that class representing a pool has associations with classes representing all its lanes.

Let P_{PL} be a set of pairs $\langle p, l\rangle$, where pair $\langle p, l\rangle \in P_{PL}$ represents the fact that pool $p \in P_P$ contain a lane $l \in P_L$, in the source BPM[11]. Total set R_{PP} of *participant-participant* associations for the given business process is:

$$R_{PP} = \Big\{ r_{PP} \in R \;\Big|\; r_{PP} = T_{PP}(\langle p, l\rangle),\ \langle p, l\rangle \in P_{PL} \Big\},$$

where the T_{PP} rule, that maps pair $\langle p, l\rangle$ into association r_{PP} between classes e_P and e_L corresponding to the given pool and lane, respectively, is:

$$\begin{aligned} T_{PP}\big(\langle p, l\rangle\big) \stackrel{def}{=} r_{PP} \;\Big|\; \Big(& name(r_{PP}) = concat(name(e_P), name(e_L)) \wedge \\ & \big(memberEnd(r_{PP}) = \{source, target\} \mid \\ & \quad type(source) = e_P \wedge multiplicity(source) = 1 \wedge \\ & \quad type(target) = e_L \wedge multiplicity(target) = *\big)\Big). \end{aligned}$$

***Participant-Object* Associations.** There are several typical bases for automated generation of *participant-object* associations that are related to: (i) *creation and subsequent usage of generated objects*, (ii) *exchange of messages*, and (iii) *activation of existing objects and subsequent usage of activated objects*.

Creation and Subsequent Usage of Generated Objects. Let G_C be a set of triplets $\langle p, t, o\rangle$, where triplet $\langle p, t, o\rangle$ represents the fact that task $t \in T$, performed by participant $p \in P$, creates object $o \in O$ in the source BPM[12]. Let G_U be a set of triplets $\langle p, t, o\rangle$, where triplet $\langle p, t, o\rangle$ represents the fact that generated object $o \in O$ constitutes the input object in task $t \in T$, performed by participant $p \in P$, in the BPM[13]. Total set G of triplets $\langle p, t, o\rangle$ representing the facts of creation and subsequent usages of generated objects is $G = G_C \cup G_U$,

[10] For the sample BPM: $E_{PP} = \{$`Customer`, `Supplier`$\}$, $E_{PL} = \{$`Supplier_Commercial`, `Supplier_Stockman`, `Supplier_Driver`$\}$, $E_O = \{$`Request`, `Response`, `OrderHeader`, `CatalogItem`, `OrderDetail`, `StockItem`, `Delivery`$\}$, $E_A = \{$`StockItem_Prepared`$\}$.

[11] e.g. $\langle$`Supplier`,`Supplier_Commercial`$\rangle$.

[12] e.g. $\langle$`Supplier_Stockman`,`Packing`,`Delivery`$\rangle$.

[13] e.g. $\langle$`Customer`,`OrderSpecification`,`OrderHeader`$\rangle$.

and total set R_{PG} of *participant-object* associations representing facts of creation and subsequent usages of generated objects for the source BPM is:

$$R_{PG} = \Big\{ r_{PG} \in R \;\Big|\; r_{PG} = T_{PG}(\langle p,t,o\rangle),\ \langle p,t,o\rangle \in G \Big\},$$

where the T_{PG} rule, that maps triplet $\langle p,t,o\rangle$ into association r_{PG} between classes e_P (participant) and e_G (generated object), is:

$$\begin{aligned} T_{PG}\big(\langle p,t,o\rangle\big) \overset{def}{=}\ r_{PG} \mid \Big(& name(r_{PG}) = name(t) \wedge \\ & \big(memberEnd(r_{PG}) = \{source, target\} \mid \\ & \quad type(source) = e_P \wedge multiplicity(source) = 1 \wedge \\ & \quad type(target) = e_G \wedge multiplicity(target) = *\big)\Big). \end{aligned}$$

Exchange of Messages. Let M_E be a set of pairs $\langle p,m\rangle$, where pair $\langle p,m\rangle$ represents the fact that participant $p \in P$ sends or receives message flow $m \in M$ in the source BPM[14]. Total set R_{ME} of *participant-object* associations corresponding to the exchange of messages between participants is:

$$R_{ME} = \Big\{ r_{ME} \in R \;\Big|\; r_{ME} = T_{ME}(\langle p,m\rangle),\ \langle p,m\rangle \in M_E \Big\},$$

where the T_{ME} rule, that maps pair $\langle p,m\rangle$ into association r_{ME} between classes e_P (participant) and e_M (message flow), is:

$$\begin{aligned} T_{ME}\big(\langle p,m\rangle\big) \overset{def}{=}\ r_{ME} \mid \Big(& name(r_{ME}) = concat(name(p), name(m)) \wedge \\ & \big(memberEnd(r_{ME}) = \{source, target\} \mid \\ & \quad type(source) = e_P \wedge multiplicity(source) = 1 \wedge \\ & \quad type(target) = e_M \wedge multiplicity(target) = *\big)\Big). \end{aligned}$$

Activation and Subsequent Usage of Activated Objects. Total set X_A of activations has already been defined. Let X_U be a set of triplets $\langle p,t,o\rangle$, where triplet $\langle p,t,o\rangle$ represents the fact that activated existing object o is used in task $t \in T$, performed by participant $p \in P$.[15] Total set X of triplets $\langle p,t,o\rangle$ representing the facts of activation and subsequent usages of activated existing objects is $X = X_A \cup X_U$, and total set R_{PA} of *participant-object* associations representing facts of activation and subsequent usages of activated objects is:

$$R_{PA} = \Big\{ r_{PA} \in R \;\Big|\; r_{PA} = T_{PA}(\langle p,t,o\rangle),\ \langle p,t,o\rangle \in X \Big\},$$

where the T_{PA} rule, that maps triplet $\langle p,t,o\rangle$ into association r_{PA} between classes e_P (participant) and e_A (activation) is:

$$\begin{aligned} T_{PA}\big(\langle p,t,o\rangle\big) \overset{def}{=}\ r_{PA} \mid \Big(& name(r_{PA}) = name(t) \wedge \\ & \big(memberEnd(r_{PA}) = \{source, target\} \mid \\ & \quad type(source) = e_P \wedge multiplicity(source) = 1 \wedge \\ & \quad type(target) = e_A \wedge multiplicity(target) = *\big)\Big). \end{aligned}$$

[14] e.g. ⟨`Customer,Request`⟩ and ⟨`Supplier_Commercial,Request`⟩.
[15] e.g. ⟨`Supplier_Stockman,Packing,StockItem_Prepared`⟩.

***Object-Object* Associations.** There are two typical bases for automated generation of *object-object* associations [19] that are related to: (i) *activation of existing objects*, and (ii) *actions having input and output objects*.

Activation of Existing Objects. Besides the association between classes corresponding to the participant and activation of existing object, one more association is to be generated for each activation (between classes corresponding to the existing object and its activation). Total set R_{EA} of *object-object* associations between classes corresponding to the existing objects and their activations for the source BPM is:

$$R_{EA} = \Big\{ r_{EA} \in R \;\Big|\; r_{EA} = T_{EA}(\langle p,t,o\rangle),\ \langle p,t,o\rangle \in X_A \Big\},$$

where the T_{EA} rule, that maps triplet $\langle p,a,o\rangle$ into association r_{EA} between classes e_E (existing object) and e_A (activation) is:

$$\begin{aligned} T_{EA}\big(\langle p,t,o\rangle\big) \overset{def}{=}\ r_{EA} \;\Big|\; \Big(& name(r_{EA}) = name(t) \wedge \\ & \big(memberEnd(r_{EA}) = \{source, target\} \mid \\ & \quad type(source) = e_E \wedge multiplicity(source) = 1 \wedge \\ & \quad type(target) = e_A \wedge multiplicity(target) = *\big)\Big). \end{aligned}$$

Actions Having Input and Output Objects. Each task $t \in T$ having $u \in \mathbb{N}$ different types of input objects $io_1, ..., io_u \in O_I$ and $v \in \mathbb{N}$ different types of output objects $oo_1, ..., oo_v \in O_O$ can be considered as a set $S(t) = \{\langle io_j, t, oo_k\rangle, 1 \le j \le u, 1 \le k \le v\}$ of $u \times v$ SISO (single input – single output) tuples. Total set $R_{OO}(t)$ of *object-object* associations for the given task $t \in T$ is:

$$R^{OO}(t) = \Big\{ r_{OO} \in R \;\Big|\; r_{OO} = T_{OO}\big(\langle io,t,oo\rangle\big),\ \langle io,t,oo\rangle \in S(t) \Big\},$$

where the basic T_{OO} rule, that maps a SISO tuple $\langle io,t,oo\rangle$ into binary association r_{OO} between corresponding classes, is given with:

$$\begin{aligned} T_{OO}\big(\langle io,t,oo\rangle\big) \overset{def}{=}\ r_{OO} \;\Big|\; \Big(& name(r_{OO}) = name(a) \wedge \\ & \big(memberEnd(r_{OO}) = \{source, target\} \mid \\ & \quad type(source) = e_{IO} \wedge multiplicity(source) = m_s \wedge \\ & \quad type(target) = e_{OO} \wedge multiplicity(target) = m_t\big)\Big). \end{aligned}$$

The corresponding source and target classes e_{IO} and e_{OO} are given with:

$$e_{IO} = \begin{cases} e_G,\ io \in O_G \\ e_X,\ io \in O_{Xn} \\ e_A,\ io \in O_{Xa} \end{cases} \qquad e_{OO} = \begin{cases} e_G,\ oo \in O_G \\ e_A,\ oo \in O_{Xa} \end{cases},$$

where $O_G \subseteq O_I$ represents a set of generated input objects, $O_{Xa} \subseteq O_I$ represents a set of activated existing input objects, while $O_{Xn} \subseteq O_I$ represents a set of existing input objects that are not activated. The corresponding source and target association end multiplicities are:

$$m_s = \begin{cases} *,\ io \text{ is a collection} \\ 1,\ io \text{ is a single object} \end{cases} \qquad m_t = \begin{cases} *,\ io \in O_{Xn} \vee oo \text{ is a collection} \\ 1, \text{otherwise} \end{cases}.$$

Total set R_{OO} of *object-object* associations, generated for all tasks having input and output objects, is $R_{OO} = \bigcup_{t \in T} R_{OO}(t)$.

Finally, total set R of associations in the target CDM is:

$$R = R_{PP} \cup R_{PG} \cup R_{ME} \cup R_{PA} \cup R_{EA} \cup R_{OO}.$$

5 Experimental Results

The previously defined rules, due to their declarative nature, are suitable for the implementation of the target generator using standard transformation languages, such as ATL and QVT. We have implemented an ATL-based generator and applied it to the sample BPM in the Eclipse-Topcased development platform. The visualization result of the automatically generated CDM is shown in Fig. 5. Due to the space limitations, we are not able to provide complete evaluation of the obtained CDM. However, from the database designer's point of view, the generated CDM has a very high *recall* and *precision* (both over 90%), since all classes and the large majority of generated associations could be retained in the target CDM. Only several associations have been generated with partly incorrect cardinalities (e.g. `CancelOrder` and `AcceptOrder`). This flaw is related to some control patterns that are presently not covered by the transformation rules.

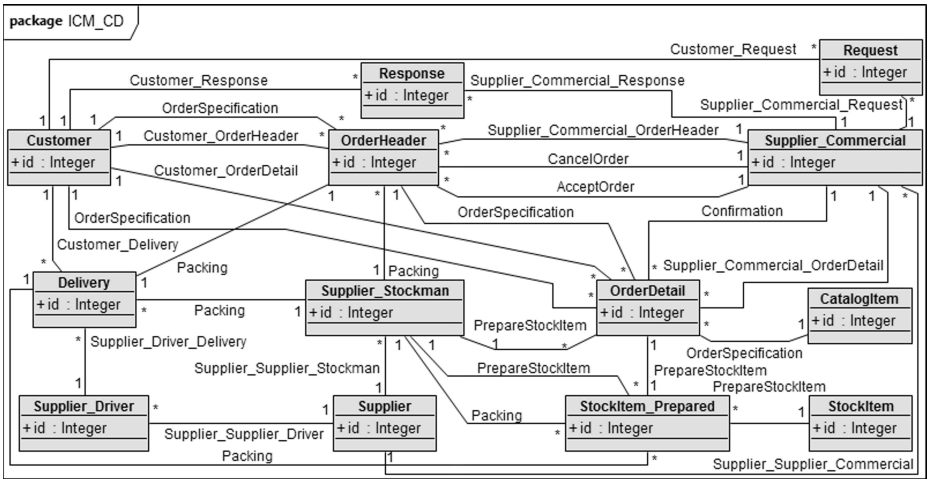

Fig. 5. Automatically generated CDM based on source BPM

6 Conclusion

In this paper we have presented an approach to the automated CDM design, that is based on collaborative BPM represented by BPMN. We have identified the semantic capacity of typical business process patterns and defined formal rules

for automated CDM design. Based on those formal rules we have implemented ATL-based automatic CDM generator and applied it to a real business model.

The evaluation of automatically generated CDM implies that the generator is able to generate a very high percentage of the target CDM with a very high precision, higher than it could be obtained by other existing approaches. An extensive evaluation of the approach, based on statistically reliable number of models, will be part of future work, as well as the further identification of the semantic capacity of BPM for automated CDM design.

References

1. Brdjanin, D., Maric, S.: Model-driven Techniques for Data Model Synthesis. Electronics 17(2), 130–136 (2013)
2. OMG: Business Process Model and Notation (BPMN), v2.0. OMG (2011)
3. OMG: Unified Modeling Language: Infrastructure, v2.4.1. OMG (2011)
4. OMG: Unified Modeling Language: Superstructure, v2.4.1. OMG (2011)
5. Jouault, F., Allilaire, F., Bezivin, J., Kurtev, I.: ATL: A model transformation tool. Science of Computer Programming 72(1-2), 31–39 (2008)
6. OMG: MOF 2.0 Query/View/Transformation Specification, v1.0. OMG (2008)
7. Rungworawut, W., Senivongse, T.: From business world to software world: Deriving class diagrams from business process models. In: Proc. of the 5th WSEAS Int. Conf. on Aplied Informatics and Communications, pp. 233–238. WSEAS (2005)
8. Rungworawut, W., Senivongse, T.: Using ontology search in the design of class diagram from business process model. PWASET 12, 165–170 (2006)
9. Brambilla, M., Cabot, J., Comai, S.: Automatic generation of workflow-extended domain models. In: Engels, G., Opdyke, B., Schmidt, D.C., Weil, F. (eds.) MODELS 2007. LNCS, vol. 4735, pp. 375–389. Springer, Heidelberg (2007)
10. Rodríguez, A., Fernández-Medina, E., Piattini, M.: Towards obtaining analysis-level class and use case diagrams from business process models. In: Song, I.-Y., et al. (eds.) ER Workshops 2008. LNCS, vol. 5232, pp. 103–112. Springer, Heidelberg (2008)
11. de la Vara, J.L., Fortuna, M.H., Sánchez, J., Werner, C.M.L., Borges, M.R.S.: A requirements enegineering approach for data modelling of process-aware information systems. In: Abramowicz, W. (ed.) Business Information Systems. LNBIP, vol. 21, pp. 133–144. Springer, Heidelberg (2009)
12. Brambilla, M., Cabot, J., Comai, S.: Extending conceptual schemas with business process information. Advances in Soft. Eng. 2010, Article ID 525121 (2010)
13. Rodriguez, A., Garcia-Rodriguez de Guzman, I., Fernandez-Medina, E., Piattini, M.: Semi-formal transformation of secure business processes into analysis class and use case models: An MDA approach. Inf. and Soft. Techn. 52(9), 945–971 (2010)
14. Zhang, J., Feng, P., Wu, Z., Yu, D., Chen, K.: Activity based CIM modeling and transformation for business process systems. Int.J. of SE and KE 20(3), 289–309 (2010)
15. Nikiforova, O., Pavlova, N.: Application of BPMN instead of GRAPES for two-hemisphere model driven approach. In: Grundspenkis, J., Kirikova, M., Manolopoulos, Y., Novickis, L. (eds.) ADBIS 2009. LNCS, vol. 5968, pp. 185–192. Springer, Heidelberg (2010)

16. de la Vara, J.L.: Business process-based requirements specification and object-oriented conceptual modelling of information systems. PhD Thesis, Valencia Polytechnic University (2011)
17. Cruz, E.F., Machado, R.J., Santos, M.Y.: From business process modeling to data model: A systematic approach. In: Proc. of QUATIC 2012, pp. 205–210. IEEE (2012)
18. Drozdová, M., Mokryš, M., Kardoš, M., Kurillová, Z., Papán, J.: Change of paradigm for development of software support for elearning. In: Proc. of ICETA 2012, pp. 81–84. IEEE (2012)
19. Brdjanin, D., Maric, S.: An Approach to Automated Conceptual Database Design Based on the UML Activity Diagram. Computer Science and Information Systems 9(1), 249–283 (2012)
20. Brdjanin, D., Maric, S.: Towards the automated business model-driven conceptual database design. In: Morzy, T., Härder, T., Wrembel, R. (eds.) Advances in Databases and Information Systems. AISC, vol. 186, pp. 31–43. Springer, Heidelberg (2013)

Computer-Aided Diagnosis of Malign and Benign Brain Tumors on MR Images

Emre Dandıl[1,3,*], Murat Çakıroğlu[2], and Ziya Ekşi[3]

[1] Vocational High School, Dept. of Computer Tech., Bilecik Seyh Edebali University, Turkey
emre.dandil@bilecik.edu.tr
[2] Faculty of Technology, Department of Mechatronics Engineering, Sakarya University, Turkey
muratc@sakarya.edu.tr
[3] Faculty of Technology, Department of Com. Engineering, Sakarya University, Turkey
ziyae@sakarya.edu.tr

Abstract. Determination of the most suitable type of treatment for brain cancer depends on the accurate detection of the type, location, size and borders of the tumor. Computer-aided diagnosis(CAD) systems help to physician in order to facilitate realizing of these aims. In this study, a CAD system was designed to detect brain tumors with computer assistance using T1 and T2 weighted MR images. The designed system segments brain tumor region of MR image using spatial-Fuzzy C-Means(FCM) method. Also, features of tumor region are extracted with image processing methods. Subsequently, support vector machine(SVM) is used for classification of benign and malign tumors in the CAD system. According to detailed test results, the proposed CAD system recognizes brain tumors with 91.49% accuracy, 90.79% sensitivity and 94.74% specificity.

Keywords: Brain tumor, computer aided diagnosis, magnetic resonance imaging, classification, SVM.

1 Introduction

Cancer is spreading widely due to various reasons such as increased use of technological tools, consumption of unhealthy foods and rise in stressful events. It is estimated that cancer, second cause of death after cardiovascular diseases today, will become the main reason for deaths in the coming years. Based on data obtained from international scientific institutions such as American Cancer Society (ASCO) [1], rate of death caused by cancer is increasing rapidly in the whole world.

There are two types of brain tumors as benign and malignant. Benign tumors grow slowly and do not spread to neighboring tissues whereas malignant tumors grow rapidly, are aggressive and may spread to other adjacent organs [2]. Brain tumors are among the leading type of cancers in recent years with seriously high contracting rates. Surgical operations, chemotherapy, and radiotherapy are used in the treatment of brain tumors [3]. The best type of treatment depends on physician's accurate detection of the type, location, size and borders of the tumor. Therefore, it is crucial

* Corresponding author.

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_16

to detect the tumor as early and accurately as possible to administer appropriate treatment. Physicians detect the tumor by interpreting these MR images. However, it is highly problematic for physicians to discriminate, make decisions and provide diagnosis in some cases. Misdiagnosis and wrong treatment methods create heavy financial burdens for the patient, reduce patient comfort and result in irremediable situations. CAD systems that will contribute to physicians' decision making processes are needed to minimize these obstacles.

Nowadays, CAD systems have been used as supplementary systems in the diagnosis of several diseases. CAD systems which will particularly support the diagnosis of cancer have been very popular in recent years. By using MR images, Fletcher-Heath et.al. [4] proposed a method that allows automatic segmentation of brain tumors through the use of fuzzy c-means clustering method (FCM). Juang et.al.[5] suggested the use of k-means method for the segmentation of brain tumors. On the other hand, some studies mention the limitations of k-means method and emphasize that more successful results can be obtained through the use of fuzzy c-means clustering techniques (FCM) [6,7]. Artificial intelligence methods are also used in the detection of brain tumors. Reddick et.al. [8] proposed a two-phase system approach to discriminate among the gray matter, white matter and other parts of the brain with Self-Organizing Maps (SOM) method by using MR images. SOM method was also utilized by Vijayakumar et.al. [9] to segment and rank brain tumors by using MR images. In their study, Sachdeva et.al. [10] proposed a segmentation, feature extraction and classification based ANN system that detects brain tumors. Machine learning methods are used in the detection, segmentation and classification of brain tumors as well. In their study, Zacharaki et.al. [11] implemented the classification of brain tumors and identification of their phases with the support vector machine learning by using MR images. Arakiri and Reddy [12] proposed a computer aided system to detect and classify the brain tumors with the help MR images.

Examination of literature about computer-aided diagnosis of brain tumors shows that in general, studies focus on identification of brain tumors. However it is evident that a holistic CAD system to differentiate between benign and malignant tumors, rank tumors, to monitor changes in tumors and implement patient management phases is absent. CAD software that can differentiate between benign and malignant tumors has been designed in this study to meet this need. Proposed CAD software can systematically detect benign/malignant tumors with high success rate.

2 Designed CAD System for Brain Tumors

The CAD system designed to detect brain tumors at early stages and classify them as benign and malignant is composed following phases. These phases includes (i) image pre-processing and enhancement, (ii) brain skull stripping and (iii) segmentation of the brain tumor, (iv) feature extraction (v) feature selection and (vi) classification of benign and malignant tumors. Block design of the CAD system is presented in Fig. 1.

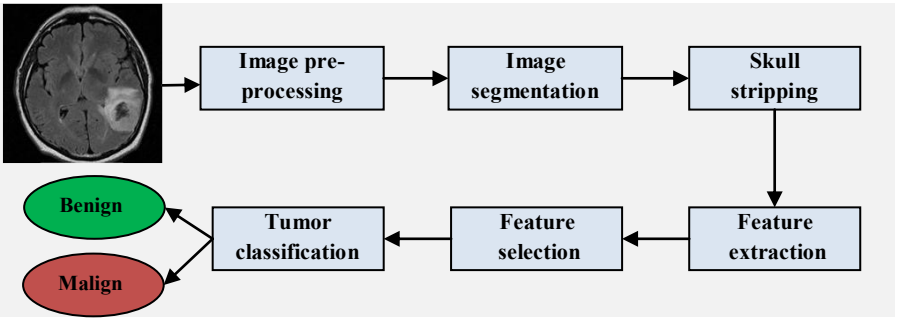

Fig. 1. Block design of the CAD system designed to detect brain tumors

2.1 Image Pre-processing

In this phase of the CAD system, the images are pre-processed to enhance image quality and remove noise. The main goal of pre-processing process is to prevent misleading results that can occur in segmentation and classification processes. First, 3x3 median filter is applied to remove very small noise and parasites. Subsequently, histogram equalization is used to remove roughness on MR images. Fig. 2 shows unprocessed brain MR images and MR images after pre-processing.

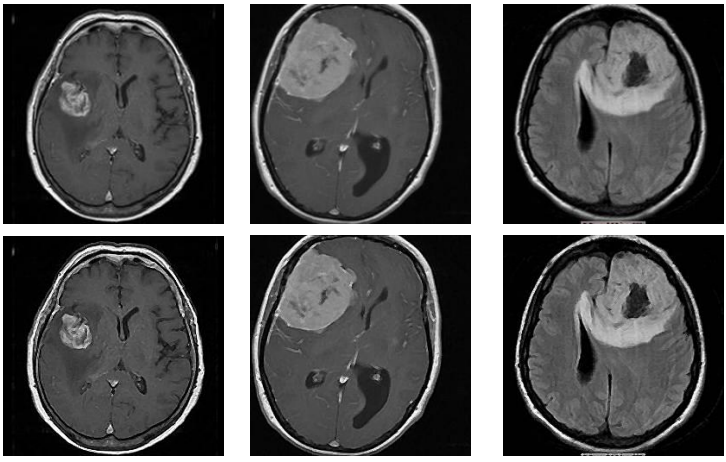

Fig. 2. Pre-processed brain images with brain tumor on MR images

2.2 Skull Stripping

This phase aims to strip the skull completely from the full brain image to eliminate the unnecessary regions [13]. In this phase, a modified novel thresholding approach (MNTA) is used to strip the skull. Skull section of the pre-processed MR image was found to have an average value between 0.2 (low) and 0.7(high) with this method. Algorithm 1 presents the steps of the thresholding method.

Algorithm 1. Double Thresholding Algorithm(MNTA)

Input: I*(input image),* ***lt****(low threshold value),* ***ht****(high threshold value)*

Output: I9*(skull stripped image)*

```
Procedure MNTA( I, lt, ht )
    1:  I2=convert I to double image
    2:  For i=1 : row length of I
            For j=1 : column   length of I
                if (I(i,j)>lt and I(i,j)<ht))
                    I3(i,j)=1;
                Else
                    I3(i,j)=0;
                End If
            End For
        End For
    3: I4 = convert I3 to binary image
    4: I5 = erode I4 image
    5: I6 = dilate I5 image
    6: I7 = fill holes in I6 image
    7: I8 = convert I7 to grayscale image
    8: I9 = I*I8
    9: Return I9
End Procedure
```

Fig. 3 presents some samples for skull stripping of pre-processed MR images by MNTA. It's seen that the proposed method successfully strips the skull of brain MR images.

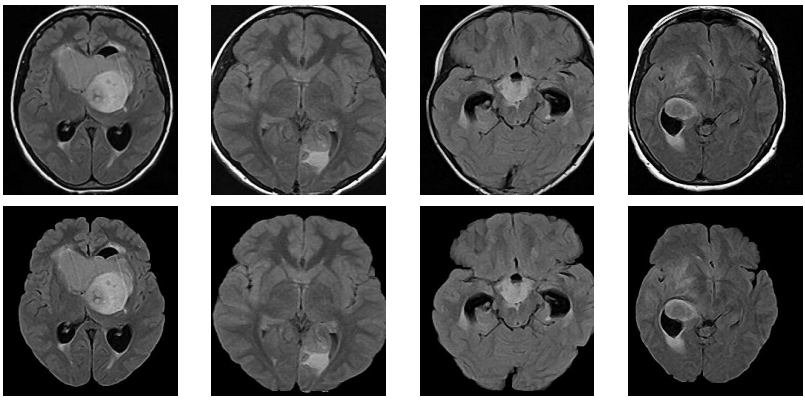

Fig. 3. Process of skull stripping in brain images

2.3 Image Segmentation

Image segmentation phase aims to segment the tumors on MR images. This study proposed spatial-FCM method to segment the brain tumors [14]. FCM feature analysis is an

unsupervised method used in image processing techniques such as clustering, medical imaging and target recognition [15]. Spatial information of the membership functions is used in spatial-FCM compared to traditional FCM method. Spatial function is defined as the sum total of membership functions of each neighboring pixel. Therefore, the FCM clustering method is made more sensitive. Fig. 4 presents the data on the brain images in the current study segmented with s-FCM.

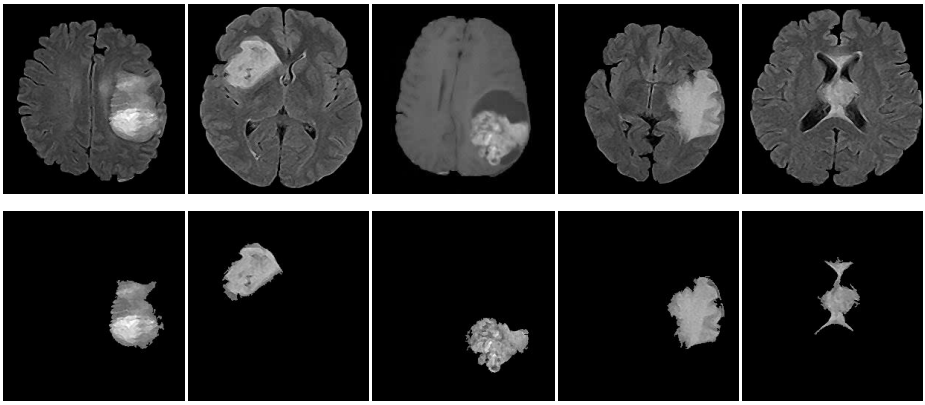

Fig. 4. Segmentation of T1 and T2 weighted brain MR images using s-FCM

2.4 Feature Extraction

Following the segmentation of tumor regions, it is necessary to identify whether the tumor is benign or malignant. Although the decision is made with the help of classifiers, some identifying features on the tumor help the decision process. Therefore, tumor features on MR images that will present the characteristics of benign or malignant tumors should be extracted. In this study, feature extraction techniques with

Table 1. Feature extraction methods, extracted features and their numbers

Feature ext. method	**Extracted features**	**Num. of ext. features**
First statistical features(FIF)	Standart deviation, Entropy, Mean, Skewness, Kurtosis,Variance	6
Shape features(SF)	Circularity, Eccentricity, Area, BoundigBox, Centroid, FilledArea, ConvexArea, EquivDiameter, EulerNumber, Extent, Perimeter, Orientation, Solidity	16
Gray-level co-occurrence matrix(GLCM)	Angular Second Moment, Entropy, Dissimilarity, Contrast, Inverse Difference, Correlation, Homogeneity, Autocorrelation, Cluster Shade, Cluster Prominence, Maximum probability, Sum of Squares, Sum Average, Sum Variance, Sum Entropy, Difference Variance, Difference Entropy, Information measures of correlation-1, Information measures of correlation2, Maximal correlation co-efficient, Inverse difference normalized, Inverse difference moment normalized	4x22=88
	Number of total extracted features	**110**

methods such as first statistical features (FIF), gray-level co-occurrence matrix (GLCM) features and shape features(SF) were used to extract tumor features. FIF is used to obtain standard statistical information such as gray-level pixel value means and standard deviation in an image. SF is used to extract many geometric information such as the roundness, area and periphery of the shape [16]. GLCM is utilized to obtain detailed statistical information on a gray-level image [17,18,19]. 6 statistical, 16 shape and 88 GLCM features were extracted in the proposed CAD system. As a result, a total of 110 different features were extracted from the ROI. Table 1 presents the feature extraction methods, extracted features and their numbers.

2.5 Feature Selection

Since 110 features extracted in this study are copious for purposes of classification, they may have negative effects on the decision period. It is crucial to select the most appropriate features from the original feature vector to increased classification accuracy. Therefore, Principal Component Analysis (PCA) was used in the current study as a feature reduction method to reduce dimensionality. PCA is a statistical and general use feature reduction method used to reduce dimensionality of complex data entries composed of large pieces of information [20,21]. In this study, the most appropriate 6 features were selected from 110 features obtained through feature extraction by reducing the dimensionality with PCA.

2.6 Tumor Classification

SVM is used in the classification phase of the proposed CAD system to differentiate between benign and malign tumors. SVM is an effective learning method used in classification problems [22]. A SVM classifier helps obtain the equation of the best linear classifier to separate the two categories. SVM uses kernel functions in separating classes with large data. SVM provides better results in applications with less data with bigger dimensionality [23]. A SVM classifier structure that can discriminate between benign-malignant tumors were used in this study.

3 Experimental Results

An image database was created for the CAD system used in the study by collecting a total of 376 T1 and T2 based MR images (248 malignant and 128 benign) based on pathological results of 67 different patients. Tumor sizes changed between 3 and 45 mm. Images in the database were collected from 17 female and 50 male voluntary patients between the ages of 27 and 79. MR images were obtained from the MR equipment in Dr. Nafiz Körez Sincan State Hospital and Medpix, Department Of Radiology and Radiological Sciences [24]. All applications in the proposed CAD system were realized with MATLAB software. All experimental studies were undertaken by using a personal computer with 3.4 GHz i7 processor, 8 GB memory and Windows 7 operating system.

3.1 Detection Rates

50% (188) of the data in the CAD system were used for training purposes whereas 50% (188) were used for testing. A confusion matrix was obtained to assess benign/malignant tumor classification performance. This matrix is composed of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) tumors. In designed CAD system, Table 2 presents the confusion matrix obtained from test data based on experimental results using support vector machine classifier.

Table 2. Obtained confusion matrix for test data in classification

CAD System	*Positive*	*Negative*	**Total**
Positive(malign)	138	14	152
Negative(benign)	2	34	36
Total	140	48	**188**

As Table 2 shows, while 138 of the 152 malignant brain tumors were identified to be malignant (TP) and 14 tumors were misclassified as benign (FN). On the other hand, 34 of the 36 benign tumors were identified as benign (TN) and only 2 tumors were misclassified as malignant (FP). Accuracy=(TP+TN)/(TP+TN+FN+FP), Sensitivity=TP/(TP+FN) and Specificity=TN/(FP+TN) criteria were calculated to obtain detection performance of the proposed CAD system as a percentage. Fig. 5 presents the performance results According to these results, the software successfully diagnoses approximately 91 cases out of 100.

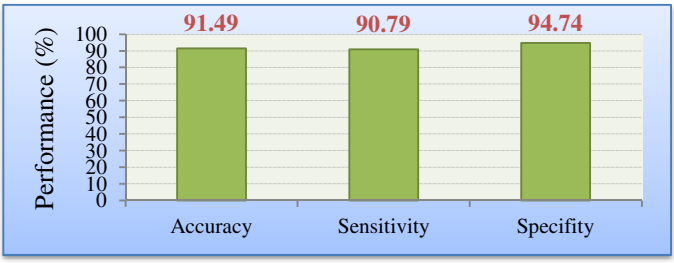

Fig. 5. Evaluation of performance measurement criteria

Fig. 6 shows the ROC(Receiver Operating Characteristic) curve for the classification accuracy of the designed CAD system. ROC curve is a preferred method to define and reliably compare the accuracy of diagnostic tests. The higher the area under the ROC curve, the bigger the success rate of the classification system [25,26]. The obtained ROC curve points to high accuracy.

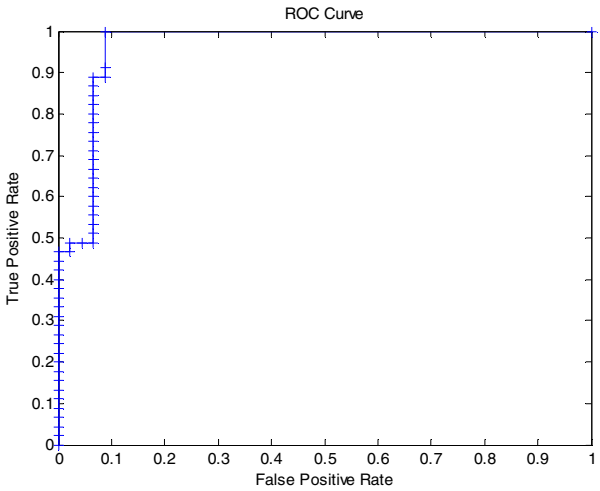

Fig. 6. ROC curve for the designed CAD system

3.2 Detection Duration

Fig. 7 displays computation times for each step in the CAD system. Longer times in segmentation may be related with the use of s-FCM method which is an iterative technique. The reason for large times in feature extraction and selection phases is based on the large amounts of necessary mathematical operations. However, the proposed CAD system classifies a brain MR image as benign/malignant in a time frame of 1-2 seconds. This is a reasonable time frame when it is compared with the time it needs for a radiologist to make decisions regarding the malignancy of the tumor.

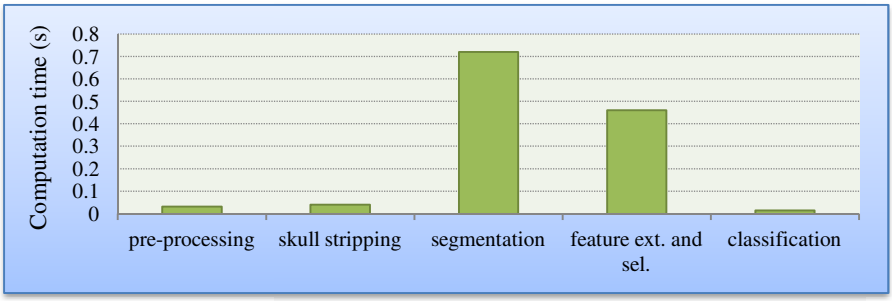

Fig. 7. Computation times in the phases of the CAD system

4 Conclusion and Future Works

This study proposes an automatic CAD system that successfully classifies brain tumors as benign/malignant using MR images. Main image processing techniques were used on MRI images in image pre-processing phase. Additionally, brain tumors were

segmented with the help of segmentation phase using s-FCM. FIF, SF and GLCM techniques were used for feature extraction and PCA method was used for feature selection. SVM method was used in the classification phase. The rates for accuracy, sensitivity and specificity were found as 91.49%, 90.79% and 94.74% respectively. Increasing the training database may provide better detection performance.

This study is a single step in an integrated CAD system that allows detection of tumors, differentiation of benign and malignant tumors, tumor ranking, monitoring the changes that occur in tumors and implementing patient management stages. In future studies, we will focus on integrating other features and design an integrated CAD platform. It is also planned to code the future CAD systems with open source code platforms such as ITK [27] instead of MATLAB to ensure reduction in the decision period and provide more effective operations.

Acknowledgments. We would like to thank the authorities in Dr. Nafiz Körez Sincan State Hospital and Medpix[24] for their significant contributions to our image database by providing MR data. This work was funded by Sakarya University BAPK (No: 2014-50-02-015).

References

1. American Society of Clinical Oncology (ASCO), http://www.asco.org/
2. Pauline, J.: Brain Tumor Classification Using Wavelet and Texture Based Neural Network. International Journal of Scientific & Engineering Research 3(10), 1–7 (2012)
3. Huo, J., Okada, K., Kim, H.J., Pope, W.B., Goldin, J.G., Alger, J., Brown, M.S.: CADrx for GBM brain tumors: predicting treatment response from changes in diffusion weighted MRI. Algorithms 2(4), 1350–1367 (2009)
4. Fletcher-Health, L.M., Hall, L., Goldgof, D.B., Murtagh, F.: Automatic segmentation of non-enhancing brain tumors in magnetic resonance images. Artificial Intelligence Medicine 21, 43–63 (2001)
5. Juang, L.H., Wu, M.: MRI brain lesion image detection based on color-converted K-means clustering segmentation. Measurement 43(7), 941–949 (2010)
6. Kolen, J.F., Hutcheson, T.: Reducing the time complexity of the fuzzy c-means algorithm. IEEE Trans. Fuzy Syst. 10(2), 263–267 (2002)
7. Murugavalli, S., Rajamani, V.: A high speed parallel fuzzy c-mean algorithm for brain tumor segmentation. Bioinform. Med. Eng. 6(1), 29–34 (2006)
8. Reddcik, W.E., Glass, J.O., Cook, E.N., Elkin, T., Deaton, R.: Automated segmentation and classifcation of multispectral magnetic resonance images of brain using artificial neural networks. IEEE Trans. Med. Imaging. 16(6), 911–918 (1997)
9. Vijayakumar, C., Damayanti, G., Pant, R., Sreedhar, C.M.: Segmentation and grading of brain tumors on apparent diffusion coefficient images using self-organizing maps. Computerized Medical Imaging and Graphics 31, 473–484 (2007)
10. Sachdeva, J., Kumar, V., Cupta, I., Khandelwal, N., Ahuja, C.K.: Segmentation, Feature Extraction, and Multiclass Brain Tumor Classification. J. Digit Imaging (2013), doi:10.1007/s10278-013-9600

11. Zacharaki, E.I., Wang, S., Chawla, S., Yoo, D.D., Wolf, R., Melhem, E.R., Davatzikos, C.: Classification of Brain Tumor Type and Grade Using MRI Texture and Shape in a Machine Learning Scheme. Magnetic Resonance in Medicine 62, 1609–1618 (2009)
12. Arakiri, M.P., Reddy, G.R.M.: Computer-aided diagnosis system for tissue characterizationbof brain tumor on magnetic resonance images. Signal, Image and Video Processing (2013), doi:10.1007/s11760-013-0456-z
13. Gambino, O., Daidone, E., Sciortino, M., Pirrone, R., Ardizzone, E.: Automatic Skull Stripping in MRI based on Morphological Filters and Fuzzy C-means Segmentation. In: Annual International Conference of the IEEE EMBS, Boston, Massachusetts, USA, August 30-September 3 (2011)
14. Chuang, K.S., Tzeng, H.L., Chen, S., Wu, J., Chen, T.J.: Fuzzy c-means clustering with spatial information for image segmentation. Computerized Medical Imaging and Graphics 30, 9–15 (2006)
15. Lye, N.S., Kandel, A., Schneider, M.: Feature-based fuzzy classification for interpretation of mammograms. Fuzzy Sets Syst. 114, 271–280 (2002)
16. Mingqiang, Y., Kidiyo, K., Joseph, R.: A survey of shape feature extraction techniques. In: Pattern Recognition Techniques, Technology and Applications, pp. 43–90. Intech (2008)
17. Akilandeswari, U., Nithya, R., Santhi, B.: Reviewon feature extraction methods in pattern classification. Euro. J. Sci. Res. 71(2), 265–272 (2012)
18. Haralick, R.M., Shanmugam, K., Dinstein, I.: Texture features for image classification. IEEE Trans. Syst. Man Cybern. 3(6), 610–621 (1973)
19. Clausi, D.A.: An analysis of co-occurrence texture statistics as a function of grey level quantization. Can. J. Remote Sensing. 28(1), 45–62 (2002)
20. Camdevyren, H., Kanik, A., Keskyn, S.: Use of principal components cores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs. Ecological Modelling 181, 581–589 (2005)
21. Chen, L.H., Chang, S.: An adaptive learning algorithm for principal component analysis. IEEE Transactions on Neural Networks 6(5), 1255–1263 (1995)
22. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
23. Shen, J., Pei, Z., Lee, E.: Support Vector Regression in the Analysis of Soft-Pad Grinding of Wire-Sawn Silicon Wafers. CITSA 2004/ISAS (2004)
24. Medpix, Department of Radiology and Radiological Sciences, `http://rad.usuhs.edu/medpix`
25. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27(8), 861–874 (2006)
26. Bradley Andrew, P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30(11), 45–59 (1997)
27. Insight Segmentation and Registration Toolkit (ITK), `http://www.itk.org/`

Novel Gene Ontology Based Distance Metric for Function Prediction via Clustering in Protein Interaction Networks

Kire Trivodaliev[1], Ilinka Ivanoska[1], Slobodan Kalajdziski[1], and Ljupco Kocarev[1,2]

[1] Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, Intelligent Systems Department, Rugjer Boshkovikj 16, 1000 Skopje, Macedonia
[2] Macedonian Academy of Sciences and Arts, Bul. Krste Misirkov 2, 1000 Skopje, Macedonia
{kire.trivodaliev,ilinka.ivanoska,slobodan.kalajdziski, ljupco.kocarev}@finki.ukim.mk

Abstract. The increased availability of large-scale protein-protein interaction (PPI) data has made it possible to have a network level understanding of the basic components and organization of the cell machinery. A significant number of proteins in protein interaction networks (PIN) remain uncharacterized and predicting their function remains a major challenge. We propose a novel distance metric for PIN clustering. First we augment the graph representing the PIN with weights derived from Gene Ontology (GO) semantic similarity and we use this augmented representation in a random walk with restarts (RWR) process. The distance between a pair of proteins is calculated from the steady state distribution of the RWR. We validate our approach by function prediction via clustering in a purified and reliable Saccharomyces cerevisiae PIN. We show that the rise of function prediction performance when using the novel distance metric is significant, as compared to traditional approaches.

Keywords: Distance metric, Graph clustering, Protein interaction network, Protein function prediction.

1 Introduction

Proteins within cells rarely act as single isolated units when they perform their functions. Even further, proteins involved in the same cellular processes are often involved in some type of interaction [1], which makes the protein-protein interactions (PPI) fundamental to almost all biological processes [2]. The advent of high-throughput technologies has allowed the construction of protein interaction networks [3] providing us with an initial global picture of protein interactions on a genomic scale but also helping us understand the basic components and organization of cell machinery from a global network level. The rapid improvement of data acquisition techniques creates a big gap between the protein data produced and experimentally characterized proteins, prompting the development of effective means to analyze and

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_17

endow high-throughput data with functional meaning. Thus, the computational function prediction is one of the most challenging problems of the post genomic era.

PPI data has the nature of networks. There is more information in a protein interaction network (PIN) compared to sequence or structure alone. A protein in a PIN is annotated with one or more functional terms. A major effort in protein annotation is the creation of the Gene Ontology (GO) [4], the most well known bio-ontology; structured and controlled vocabulary for describing gene and protein products. GO term sets associated with interacting proteins within a PIN can be used in performing semantic driven protein comparison. This comparison is called semantic similarity. There is no single best way to calculate semantic similarity considering the current bio-ontologies, but several metrics have been proposed to calculate protein semantic similarity in the context of the GO [5].

We can now define the process of protein function prediction in the context of PIN as the process of understanding the protein's interaction neighborhood and the functions associated within. The simplest view of a PIN is an unweighted, undirected graph in which every protein is represented with a single node and interactions between two proteins are represented with edges between the corresponding pair of nodes. The grouping of nodes in a graph representing the PIN i.e. the clustering of the PIN is proven to be an effective approach towards protein function prediction [6]. Traditional graph-based agglomerative methods employ a variety of similarity measures between nodes to partition PPI networks, but they often result in a poor clustering arrangement that contains one or a few giant core clusters with many tiny ones[7]. To improve the clustering results, PPI networks were weighted based on topological properties[8-11] such as shortest path length, clustering coefficients, node degree, or the degree of experimental validity. In [12] the edge-betweenness and its modified version, using weights generated from micro array expression profiles, have been used as a method to find functional modules in the PIN.

PINs can also be analyzed in regards of extraction of densely connected subgraphs or protein complexes. Molecular Complex Detection (MCODE) [13] is based on node weighting by local neighborhood density and outward traversal from a locally dense seed protein to isolate densely connected regions. Restricted Neighborhood Search Clustering (RNSC) [14]), is a cost-based local search algorithm that explores the solution space to minimize a cost function, calculated according to the numbers of intra-cluster and inter-cluster edges. The Markov Cluster algorithm (MCL) [15] simulates a flow on the graph by calculating successive powers of the associated adjacency matrix. More recent approaches exploit semantic similarity measures based on GO between pairs of proteins within the PIN. PROCOMOSS [16] uses a multi-objective evolutionary approach in which graphical properties as well as biological properties based on GO semantic similarity measure are considered as objective functions for detecting protein complexes in a PIN. CSO [17] performs clustering based on network structure and ontology attributes similarity on GO attributed PINs. Both of these algorithms achieve state-of-the-art performance and are another proof that the PIN needs to be augmented and that GO provides the necessary resources for that aim.

In this paper we address the problem of clustering in PINs in a twofold manner. First, we add weights to the graph representation for the PIN and second, using these weights we get proximity estimates which are incorporated in the distance between nodes of the graph. We validate our approach by function prediction via clustering in a purified and reliable Saccharomyces cerevisiae PIN. We show that the rise of function prediction performance when using the novel distance metric is significant, as compared to traditional approaches.

2 Research Methods

A general architecture we use for predicting protein function via clustering of a PIN using a similarity based distance metric is illustrated in Fig. 1. The following sections explain each of the steps.

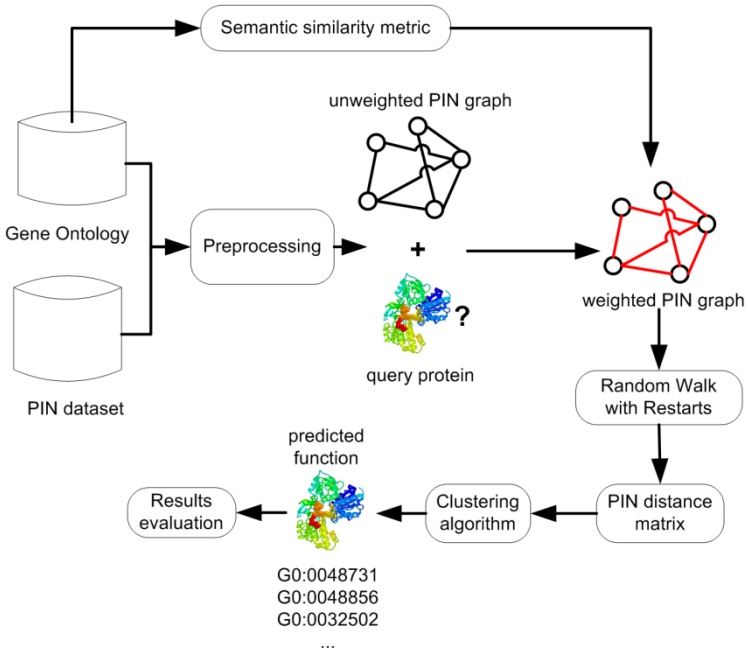

Fig. 1. A general architecture for predicting protein function via clustering of a PIN using a semantic similarity based distance metric

2.1 Protein-Protein Interaction Data

We conduct our experiments on Saccharomyces cerevisiae PPI data which are compiled from a number of established datasets used in previous research on PPI. In the first step we merge the PPI datasets of Uetz [18], Ito [19], Ho [20], Krogan [21], and Gavin [22]. Next, from these interactions we leave only those which have more than one supporting experimental evidence found in public databases like: DIP,

MIPS, MINT, BIND or BioGRID. The functional terms for each protein are taken from the SGD database [23], and are unified with the GO terminology. This data is further purified as follows. First, the trivial functional terms are erased. Next, additional terms are calculated for each protein by the policy of transitive closure derived from the GO. The extremely frequent terms (appearing in more than 300 proteins) are also excluded, because they are very general and do not carry significant information. The final dataset is highly reliable and consists of 2502 proteins with 6354 interactions between them and has a total of 888 functional terms.

2.2 Semantic Similarity Measure

Semantic similarity in an ontology with a directed acyclic graph structure, such as GO, can be quantified in two ways. Edge-based approaches rely on counting the number of edges in the graph path between two terms. Node-based approaches are essentially comparing the properties of the terms involved, which can be related to the terms themselves, their ancestors, or their descendants. The most commonly used concept here is information content (IC), which gives a measure of how specific and informative a term is. The IC of term t is defined as negative log likelihood $-log\ p(t)$, where $p(t)$ is the probability of occurrence of t in a specific knowledgebase. There are also hybrid methods that combine the two types of methods for semantic similarity, and they give weights to the GO nodes or edges according to their type.

We have compared a number of semantic similarity metrics [24] and our results showed that we get best results, when used in function prediction via clustering, for Resnik's [25] semantic similarity metric. This metric was originally developed for WordNet and later applied to GO and it defines that the similarity between two terms is the IC of their most informative common ancestor (MICA):

$$sim_{Resnik}(t_1, t_2) = max_{t \in A(t_1, t_2)}(-log\ p(t)) \tag{1}$$

where $A(t_1, t_2)$ is the set of common ancestors for terms t_1 and t_2. To define the larity between two proteins p_1 and p_2, each having annotation set $T_1 = \{ t_{11}, t_{12}, ..., t_{1m} \}$ and $T_2 = \{ t_{21}, t_{22}, ..., t_{2n} \}$, respectfully, we first define a similarity matrix for proteins p_1 and p_2 as $SIM = [sim_{ij} = sim_{Resnik}(t_{1i}, t_{2j})]$ with size $m \times n$. We can now define the similarity between p_1 and p_2 as:

$$sim(p_1, p_2) = max\left(\frac{1}{m} \sum_{i=1}^{m} max_{1 \le j \le n}\ sim_{ij}, \frac{1}{n} \sum_{j=1}^{n} max_{1 \le i \le m}\ sim_{ij} \right) \tag{2}$$

2.3 PIN Graph Augmentation

The simplest way to enrich a simple unweighted graph representation is to add weights to edges and thus define a weighted graph $G=(V,E,W)$ for the PIN, where W is a matrix whose elements w_{ij} are the weights of the edges $(i, j) \in E$, $i,j \in V$. We de-

fine weight as being consisted of two parts: a content-based weight and a structure-based weight.

A content-based weight calculation is one that assigns weight w_{ij}^1 to the edge (*i, j*) by looking at the terms ("contents") associated with nodes *i* and *j*, not taking their environment (the graph structure) into account. We compute w_{ij}^1 as follows:

$$w_{ij}^1 = \frac{max([\,sim(i,j)\,]) - sim(i,j)}{max([\,sim(i,j)\,]) - min([\,sim(i,j)\,])} \tag{3}$$

with $sim(i,j)$ calculated according to (2); max and min are the maximum and minimum value over all possible protein pairs similarities.

A structure-based weight calculation is one that takes the context of the nodes *i* and *j* into account, but not the contents of the nodes themselves, when calculating weight w_{ij}^2 for the edge (*i, j*). The structural information of the graph G_2 is naturally encoded in its adjacency matrix $A=[a_{ij}]$ so we can define the weight matrix $W^2 = [w_{ij}^2]$ as follows:

$$W^2 = W^1 \times A + A \times W^1 \tag{4}$$

where $W^1 = [w_{ij}^1]$ is the content-based weight matrix. Since $a_{ij} = 0$, $\forall (i,j) \notin E$, for each w_{ij}^2 the first part of equation (4) gives the sum of all content-based weights of edges between node *i* and all neighbors of *j*, while the second part is the sum of all content-based weights between node *j* and all neighbors of *i*. PINs are known to have proteins that interact with many other which gives rise to hubs, which means (4) will be biased, so we average and normalize the values to overcome this unwanted effect and get equation (5).

$$W^2 = \left(W^1 \times A^1 + A^2 \times W^1\right) \Big/ 2 \tag{5}$$

where $A^1 = [a_{ij} / (\sum_{n=1}^{N} a_{nj})]$, $A^2 = [a_{ij} / (\sum_{n=1}^{N} a_{in})]$, and $N=|V|$.

The final weight of an edge combines both content-based and structure-based weights; a natural way of combining them is taking the average of the two:

$$W = \left(W^1 + W^2\right) \Big/ 2 \tag{6}$$

The two parts of the weight calculation are essential in the validation phase. When we choose a protein to be a query we remove all the annotations associated with it. The importance of this type of calculation lies in the fact that if we don't have the structure part the query protein would be isolated and its interaction context lost and any algorithm we apply would lead to low performance.

2.4 Distance Metric

To define the distance between two nodes (proteins) in the PIN graph we first define a Random Walk with Restarts (RWR) on the graph. A RWR starting at node *i* is defined with:

$$P_t^i = (1-c)W^T P_{t-1} + cr \tag{7}$$

where $P_t^i(j)$ is the probability that at time *t* the random walker would end up at node *j*, *c* is the restart (return to start node) probability, and *r* is a column vector for which only the *i*-th element equals 1 and all other are 0. The restart probability defines the diameter around the start node which the walker would traverse before returning to the start node, with c=1 traversing only the immediate neighborhood, and c=0 the whole graph. In our experiments we got best results using c=0.5 and that is the default setting in the results section. The steady state distribution, i.e. $P_{t+1}^i = P_t^i$, is a measure of affinity or closeness of node *i* to other nodes in the network. We can now define an intermediate distance from node *i* to node *j* as:

$$d'(i,j) = 1 - P_t^i(j) \tag{8}$$

Note that since the PIN graph is irregular we have $d'(i,j) \neq d'(j,i)$. The distance between nodes *i* and *j*, based on (8), is now defined as:

$$D(i,j) = \left[d'(i,j) + d'(j,i)\right]/2 \tag{9}$$

2.5 Function Prediction

This distance metric can now be used in the clustering of the PIN graph. We use two different clustering algorithms: *k*-medoids and agglomerative hierarchical clustering. We apply the k-medoids algorithm as defined in [26] (with k = 150) and a single-linkage hierarchical clustering with cut-off made according to [27]. Once we obtain the clusters we can predict the functions of a query protein by scoring the functions within its cluster, *K*. Each term is ranked by its frequency of appearance as a term assigned to nodes within the cluster:

$$s(j)_{j \in T_K} = \sum\nolimits_{i \in K} z_{ij} \tag{10}$$

where T_K is the set of terms present in the cluster *K*, and

$$z_{ij} = \begin{cases} 1, & \text{if i-th node from K is assigned with the j-th term from } T_K \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

3 Results and Discussion

The effective evaluation of protein functional annotation is challenging. The lack of agreed measures and benchmarks used for assessment of the methods performance makes this task difficult. In our work we used the leave-one-out method when only one protein at time plays the role of a query protein and is considered as unannotated. Once the clustering algorithm has been applied, for each term present in the query cluster (i.e. the cluster of the query protein) we calculate its rank according to (11), and all ranks are then normalized to [0,1]. The query protein is annotated with all functions with rank above a previously determined threshold ω. We change the threshold in the [0,1] range and compute the numbers for the four possible different classes which can occur during the assignment process: True Positive (TP) - when annotation is assigned and is part of the true annotation set, True Negative (TN) - when annotation is not assigned to the protein and is not part of the true annotation set, False Positive (FP) - when annotation is assigned but is not part of the true annotation set, False Negative (FN) - when annotation is not assigned but is part of the true annotation set.

Each annotation is assigned to one of the four classes. Using the number of annotations in each class we can calculate the following statistical measures:

$$Sensitivity\ (TruePositiveRate) = \frac{TP}{TP + FN}, FalsePositiveRate = \frac{FP}{FP + TN} \tag{12}$$

Graphed as coordinate pairs, the Sensitivity and the FalsePositiveRate form the receiver operating characteristic (ROC) curve. The Area Under Curve (AUC) of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

We performed our experiments using the setup as explained in the previous section. We compared the results of the novel metric with the performance we get when using a standard distance metric. We also made experiments in which instead of using GO similarity metric for augmenting the PIN graph we used a non-semantic metric, i.e. normalized Jaccard index, which weighs an edge based on direct comparison of the interacting nodes. The difference between the semantic and non-semantic approach depicts the benefits of using GO in the function prediction process.

Table 1. Results for function prediction using k-medoids and different distance metrics

metric	ω =	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	AUC
standard	sens	0.822	0.672	0.536	0.453	0.380	0.312	0.258	0.201	0.163	0.132	0.748
	fpr	0.551	0.298	0.139	0.081	0.057	0.032	0.018	0.011	0.006	0.002	
jaccard	sens	0.884	0.689	0.547	0.426	0.364	0.284	0.220	0.168	0.125	0.092	0.822
	fpr	0.351	0.216	0.118	0.075	0.041	0.028	0.016	0.009	0.005	0.001	
novel GO	sens	0.912	0.710	0.552	0.461	0.370	0.295	0.236	0.174	0.129	0.092	0.851
	fpr	0.317	0.202	0.093	0.058	0.033	0.020	0.015	0.008	0.005	0.001	

Table 2. Results for function prediction using single-linkage and different distance metrics

metric	ω =	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	AUC
standard	sens	0.817	0.628	0.489	0.399	0.312	0.230	0.188	0.124	0.112	0.055	0.791
	fpr	0.374	0.163	0.091	0.052	0.038	0.017	0.013	0.006	0.004	0.002	
jaccard	sens	0.889	0.718	0.535	0.418	0.320	0.251	0.192	0.137	0.116	0.091	0.856
	fpr	0.279	0.139	0.089	0.049	0.037	0.020	0.011	0.005	0.003	0.001	
novel GO	sens	0.918	0.739	0.539	0.453	0.327	0.262	0.207	0.142	0.119	0.091	0.886
	fpr	0.246	0.125	0.064	0.032	0.030	0.012	0.010	0.005	0.003	0.001	

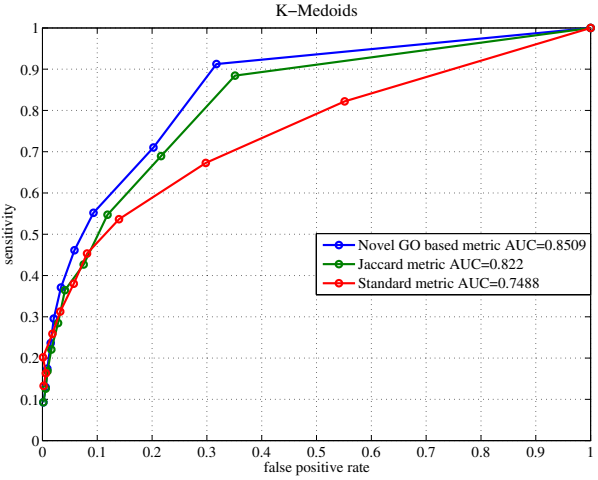

Fig. 2. ROC curves for function prediction using k-medoids and different distance metrics

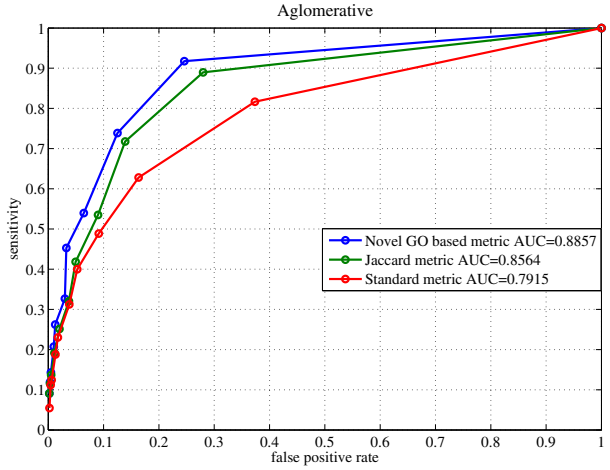

Fig. 3. ROC curves for function prediction using single-linkage and different distance metrics

Table 1 and Table 2 show the results for the function prediction when k-medoids and single-linkage agglomerative clustering are applied using the different distance metrics. We can see that the overall performance, i.e. the AUC value, is significantly improved when we use the novel GO based metric as compared to the standard distance metric. We can also see that the semantic metric outperforms the non-semantic metric. Furthermore, the drop in the false positive rate values is notable and again the novel GO based distance metric outperforms the other two metrics. Fig.1 and Fig.2 show the corresponding ROC curves for Table 1 and Table 2. These results confirm all our previous assumptions that the novel GO based metric is by far better than the standard distance metric and also outperforms the non-semantic one.

4 Conclusion

A novel distance metric for PIN clustering was presented. The construction of the metric is performed in two stages. First, we augment the graph representation for the PIN using GO semantic similarity as to encode the annotations in the graph within weights of the graph edges. Second, we compute the steady state distribution of a RWR on the augmented graph and we use this as a proximity estimate in the calculation of distance between nodes of the graph. We validated our approach by function prediction via clustering in a purified and reliable Saccharomyces cerevisiae PIN. Experiments revealed that the function prediction performance when the novel GO based distance metric is used outperforms standard and non-semantic distance metric, regardless of the clustering algorithm used. These results are proof that when using PINs in computational function prediction their graph representation needs to augmented and the GO is a suitable resource for that aim.

References

1. von Mering, C., Krause, R., Sne, B., et al.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417(6887), 399–403 (2002)
2. Hakes, L., Lovell, S.C., Oliver, S.G., et al.: Specificity in protein interactions and its relationship with sequence diversity and coevolution. PNAS 104(19), 7999–8004 (2007)
3. Harwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W.: From molecular to modular cell bi-ology. Nature 402, c47–c52 (1999)
4. The gene ontology consortium: Gene ontology: Tool for the unification of biology. Nature Genetics 25(1), 25–29 (2000)
5. Pesquita, C., Faria, D., Bastos, H., Ferreira, A., Falcão, A.O., Couto, F.M.: Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatic 9(5), S4 (2008)
6. Brohée, S., van Helden, J.: Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7, 48 (2006)
7. Barabasi, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113 (2004)
8. Arnau, V., Mars, S., Marin, I.: Iterative cluster analysis of protein interaction data. Bioinformatics 21, 364–378 (2005)

9. Rives, A.W., Galitski, T.: Modular organization of cellular networks. PNAS 100, 1128–1133 (2003)
10. Friedel, C.C., Zimmer, R.: Inferring topology from clustering coefficients in protein-protein interaction networks. BMC Bioinformatics 7, 519 (2006)
11. Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A.: Detection of functional modules from protein interaction networks. Proteins 54, 49–57 (2004)
12. Luo, F., Yang, Y., Chen, C.F., Chang, R., Zhou, J., et al.: Modular organization of protein interaction networks. Bioinformatics 23, 207–214 (2007)
13. Bader, G.D., Hogue, C.W.: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4, 2 (2003)
14. King, A.D., Przulj, N., Jurisica, I.: Protein complex prediction via cost-based clustering. Bioinformatics 20, 3013–3020 (2004)
15. Enright, A.J., Dongen, S.V., Ouzounis, C.A.: An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30(7), 1575–1584 (2002)
16. Mukhopadhyay, A., Ray, S., De, M.: Detecting Protein Complexes in PPI Network: A Gene Ontology-based Multiobjective Evolutionary Approach. Molecular BioSystems 8(11), 3036–3048 (2012)
17. Zhang, Y., Lin, H., Yang, Z., Wang, J., Li, Y., Xu, B.: Protein Complex Prediction in Large Ontology Attributed Protein-Protein Interaction Networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 10(3), 729–741 (2013)
18. Uetz, P., et al.: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403(6770), 623–627 (2000)
19. Ito, T., et al.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Genetics 98(8), 4569–4574 (2001)
20. Ho, Y., et al.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868), 180–183 (2002)
21. Krogan, N.J., et al.: Global Landscape of Protein Complexes in the Yeast Saccharomyces cerevisiae. Nature 440(7084), 637–643 (2006)
22. Gavin, A.C., et al.: Proteome survey reveals modularity of the yeast cell machinery. Nature 440(7084), 631–636 (2006)
23. Dwight, S.S., et al.: Saccharomyces Genome Database (SGD) provides secondary gene annotation using Gene Ontology (GO). Nucleic Acids Research 30(1), 69–72 (2002)
24. Ivanoska, I., Trivodaliev, K., Kalajdziski, S.: Protein Function Prediction Using Semantic Driven K-Medoids Clustering Algorithm. International Journal of Machine Learning and Computing 4(1), 52–56 (2014)
25. Resnik, P.: Using information content to evaluate semantic similarity. In: IJCAI 2005, pp. 448–453 (1995)
26. Witsenburg, T., Blockeel, H.: K-means based approaches to clustering nodes in annotated graphs. In: Kryszkiewicz, M., Rybinski, H., Skowron, A., Raś, Z.W. (eds.) ISMIS 2011. LNCS (LNAI), vol. 6804, pp. 346–357. Springer, Heidelberg (2011)
27. Langfelder, P., Zhang, B., Horvath, S.: Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24(5), 719–720 (2008)

Modeling the Speedup for Scalable Web Services

Sasko Ristov[1], Marjan Gusev[1], and Goran Velkoski[2]

[1] Ss. Cyril and Methodious University, FCSE
Rugjer Boshkovikj 16, 1000 Skopje, Macedonia
{sashko.ristov,marjan.gushev}@finki.ukim.mk
[2] Innovation LTD, Vostanicka 118, 1000 Skopje, Macedonia
goran.velkoski@innovation.com.mk

Abstract. Cloud Computing enables scaling of web services. A typical customer would expect that the performance of web services on the cloud will be directly proportional to the availability of rented resources. However, we obtained that achieved performance is not always directly proportional to the scaling, by realising series of experiments varying the server load by changing the message size and the number of concurrent messages. The goal is to analyse the performance of web services utilising different hardware resources on the cloud for the same server load. We set a hypothesis about expected performance behaviour and then analyse and discuss the results about optimal resources when scaling the load. Interestingly, the results show different behaviour in determined regions, and also that there is a region where web services hosted on the cloud achieve superlinear speedup (speedup greater than the number of scaled hardware resources), meaning that the customers will get more performance than expected. Moreover, a region where input parameters are smaller without scaling the resources, provides an even better performance compared to scaled resources.

Keywords: Cloud Computing, Performance, Speedup.

1 Introduction

The benefits of using web services hosted on the cloud environment are numerous. The mechanisms for web service communication must be platform-independent, secured, and as lightweight as possible due to the distributed and heterogeneous nature of Web [4]. Migrating the services on the cloud reduces the cost since the companies can invest their money into business rather than in expensive and under-utilised or over-utilized complex data centres. Web service hosting platforms must standardise application architecture to ensure scalability [2]. The cloud infrastructure can support web services to be scalable.

Although most of the cloud services have periods of especially stable performance, performance variability of production cloud services is an important challenge [10]. The goal of this research is the performance analysis of different web services hosted on a scalable cloud environment. That is, we try to determine if the achieved speedup while scaling the resources will be linear, the

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_18

same as the price of the rented scaled resources. The performance analysis of cloud hosted web services is performed by series of experiments for computation-intensive and memory-demanding web services, while scaling the virtual machine (VM) instance hardware resources. Varying the server load by changing the message size and the number of concurrent messages, we measure the response time to analyse the performance and also to determine the region of input parameters that achieves maximum speedup for different scaling factor (the number of processors).

The rest of the paper is organised as follows. In Section 2, we present the related work that we have found in the literature. Section 3 briefly presents the testing methodology. Our theoretical analysis of speedup and its limits is elaborated in Section 4. The results of the experiments which confirms our theoretical analysis are presented in Section 5. We discuss the results in Section 6. Finally, we conclude our work and present the plans for future work in Section 7.

2 Related Work

There are several papers addressing the performance of web services. An optimization model for optimal resource allocation across a set of web service classes running on the same physical server in virtual environment is proposed by Almeida et al. [1]. Bonneta et al. [3] presented S service scripting language, compiler, and runtime system that can efficiently exploit today's multi-core parallel architectures to scale the number of concurrent requests. Ristov et al. [13] show that migrating the web services on the cloud reduces their performance using the same hardware resources.

Although the cloud can scale its resources, it does not guarantee that the performance will scale the same as the scaling factor. Virtualization is another layer that also produces performance discrepancy. Performance isolation is necessary in a cloud multi-tenant environment [15] since the same VM on the same hardware at different times among the other active VMs will not achieve the same performance [11]. Gusev and Ristov [7] have reported a phenomenon represented as almost 10 times better performance when web services are hosted on several VM instances and the concurrent requests are balanced among them, compared to the environment where the same amount of resources are allocated to a single VM instance. Also VM granularity significantly affects the workload's performance for small network workload [14]. Underutilization of resources by adding more nodes can considerably improve the performance implementing more parallelism [9]. In this paper, we determine a region where over-utilization of the cloud node is still better than under-utilization since we determine superlinear speedup.

Superlinear speedup for parallel execution is found when the same problem hosted on the cloud is scaled on more cores [5, 12]. We have set a hypothesis that there is also a superlinear region for particular input parameters even in the case of different web services.

3 Testing Methodology

The testing environment is based on a client-server architecture deployed on OpenStack open source cloud platform with KVM hypervisor. OpenStack is deployed on dual node, i.e., one physical server is the cloud controller, which schedules the VM instances, while the other physical server is cloud node, which runs the VM instances.

Server nodes consist of Intel(R) Xeon(R) CPU X5647 @ 2.93GHz with 4 cores and 8GB RAM installed. The platform consists of Linux Ubuntu Server 11.04 operating system and Apache Tomcat 6 as the application server.

3.1 Test Cases

Three document style Java web services are developed with criteria to test both computation-intensive and memory-demanding web services:

- *Concat* web service, which accepts two strings and returns their concatenation. It is a memory-demanding web service that depends on the input parameter size M with complexity $O(M)$.
- *Sort* web service, which also accepts two strings and returns their concatenation alphabetically sorted. It is both memory-demanding and computation-intensive web service with complexity $O(M \cdot log_2 M)$.
- *Math* web service is computation-intensive only web service. It accepts two integers parameters A and B, and returns the result of:

$$x = [\sin(A) \% \cos(B) \cdot \sqrt{\sin(A) \cdot \cos(B)}]^A \cdot \\ \cdot \ln[\sin(A) \% \cos(B) \cdot \sqrt{\sin(A) \cdot \cos(B)}]^B \quad (1)$$

3.2 Test Plan

We employ tests on VM instances with one, two and four CPUs respectively. N messages are sent with M bytes each, with variance 0.5, meaning that the number of concurrent messages will vary by $N/2$; it will increase to $3 \cdot N/2$, then decrease to $N/2$, and finally end with N within 60 seconds, i.e. the end of the test. Each test case runs for 60 seconds.

Parameter size M is measured in KB and with the following values 0, 1, 2, 4, 6, 8 and 10 for Concat web service and 0, 1, 2, 4 and 6 for Sort web service. Math web service is computation-intensive and does not need varying parameter sizes as input.

Both Concat and Sort web services are loaded with $N = 12$, 100, 500, 1000, 1500 and 2000 requests/second for each message size. Math web service tests is loaded additionally with $N = 2500, 5000$ and 10000 requests/second.

The performance of these services is calculated by measuring the average response time T and average CPU utilization for experiments using the same VM image with different resources. Next, we calculate the *Speedup* $S(P) = T_1/T_P$, where $P \in \{2, 4\}$ denotes the number of processors (cores) allocated in a VM.

T_1 denotes the average response time for particular load on web service hosted on VM with 1 processor (without scaling the resources), and T_P denotes the average response time for particular load on web service hosted on VM with P processors (with scaling the resources).

4 Speedup Limits

In this section, we analyze the speedup distribution and limits in a *Scaled system*, i.e. VM with P processors compared to *Conventional system*, i.e. VM with 1 processor.

The speedup domain is $S(P) \in (0, \infty)$ and we model it with five sub-domains:

- *Drawback* $0 < S(P) < 1$ - the new scaled system achieves less performance than the conventional system;
- *No Speedup* $S(P) = 1$ - the new scaled system achieves the same performance as the conventional system;
- *Sublinear* $1 < S(P) < P$ - the new scaled system performs better than the conventional system, but smaller than linear speedup (as expected according to Gustafson's scaled speedup);
- *Linear* $S(P) = P$ - the new scaled system achieves P times better performance than the conventional system;
- *Superlinear* $S > P$ - the new scaled system achieves greater performance than scaled size resources [6].

Despite communication and memory bound presented in [8], we introduce additional *Compute bound* dependency that we expect to appear for huge web service load. The speedup will stop rising and will saturate in this domain.

We aim to experimentally confirm the hypothesis about performance behavior and identify existence of the following 4 different regions for scaling parameter performance:

- *Under-utilized* - speed is lower in the scaled system rather than in conventional system due to communication bound;
- *Proportional* - speedup increases as input parameter size increases;
- *Superior* - speedup is greater than expected - due to memory capacity bound;
- *Saturated* - speedup decreases with increase of parameter size - due to compute bound.

Figure 1 depicts the expected speedup regions depending of web service load. X-axis presents the server load (parameters M or N), while Y-axis presents the achieved speedup while the resources are scaled with factor P.

Let us explain our hypothesis. The scaled system provides lower speed for minimal load due to communication bound. Increasing the load will reduce the computation time in the scaled system and the speedup will increase. Increasing the load even more, the response time in conventional system will grow more than the response time on the scaled system since the response time increases more for greater load than for small load [13], thus producing superlinear speedup.

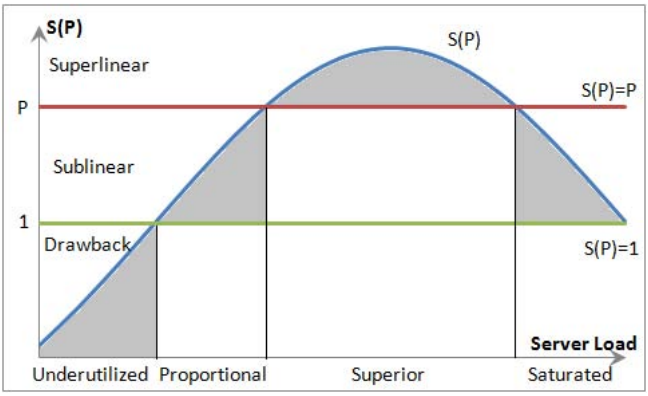

Fig. 1. Expected speedup of the scaled system compared to conventional

When both response times in conventional and scaled systems are huge, their ratio decreases which also decreases and saturates the speedup. Therefore, we expect two unusual regions, i.e., under-utilized and superior.

5 Experimental Results

This section describes the results of testing the performance impact of scaling the resources on the cloud for particular web service load, i.e. the speedup achieved on the scaled system with more resources. We also analyze the results to understand the performance impact of different message sizes and number of concurrent messages on the three web services described in Section 3.1.

5.1 Scaling the Concat Web Service

Figure 2 depicts the speedup for Concat web service when the resources are scaled with factor 2. We observe three speedup regions: constant, increase and decrease. The speedup retains the value when one input parameter increases and the other is constantly small; increases when one parameter is huge and the other begins to increase; and also decreases when both input parameters are huge.

We can educe two important conclusions. The first addresses a superlinear speedup region (speedup greater then scaled number of resources) with maximum speedup 5.67 (Linear speedup is 2), and the second addresses a region with less performance (speedup less than 1) with minimum value of 0.99.

Similar speedup regions are observed as those identified by scaling factor 2. The region with decreasing performance is also found for small input parameters and the speedup achieves its minimum value of 0.97. There is a superlinear speedup region and maximum speedup of 6.7 (Linear speedup is 4).

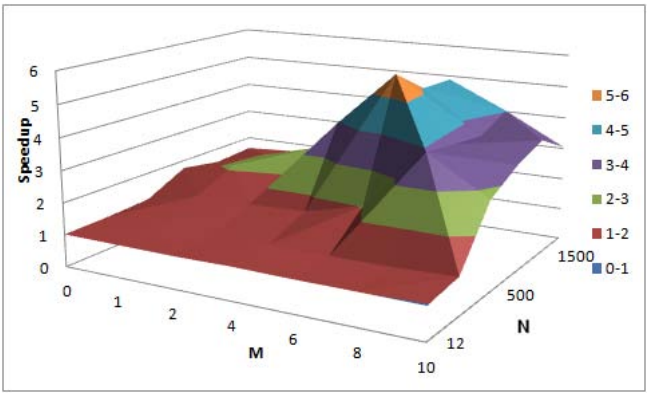

Fig. 2. Concat web service speedup when scaled with factor $P = 2$ as a function of request rate N (in requests/second) and string size M in KB

5.2 Scaling the Sort Web Service

Figure 3 depicts the speedup for Sort web service when the resources are scaled with factor 2. The same speedup distribution is determined, but the maximum speedup appears for smaller message size. The speedup is smaller than Concat web service since it requires additional CPU operations for sorting. Maximum measured superlinear speedup is 2.29 (Linear speedup is 2), but also there is a region with decreased performance with minimum value of $S(2) = 0.88 < 1$.

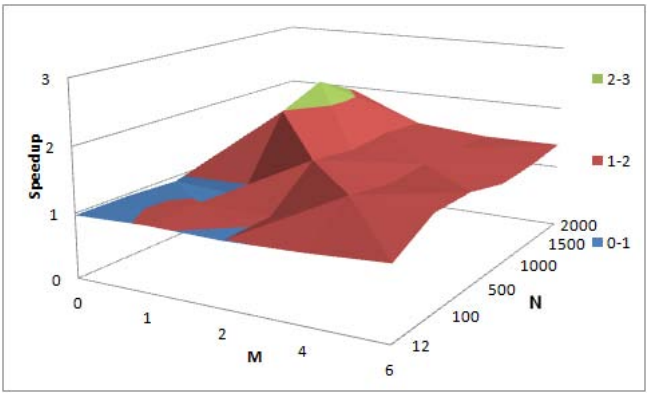

Fig. 3. Sort web service speedup when scaled with factor $P = 2$ as a function of request rate N (in requests/second) and string size M in KB

Similar results are achieved for speedup when the resources are scaled with factor 4. We have measured a huge decrease in performance up to 0.39. Also, we observe a huge maximum superlinear speedup of 7.23 (Linear speedup is 4).

5.3 Scaling the Math Web Service

Figure 4 depicts the speedup for Math web service when the resources are scaled with factor 2 and 4. There is only sublinear speedup for Math web service, i.e., the maximum observed speedup is 1.62 for $S(2)$ (Linear speedup is 2) and 1.56 for $S(4)$ (Linear speedup is 4). However, we also observe a region with decreased performance in the case of 12 concurrent messages with minimum value of 0.99 and 0.95 when hosted on VM with 2 and 4 CPUs correspondingly.

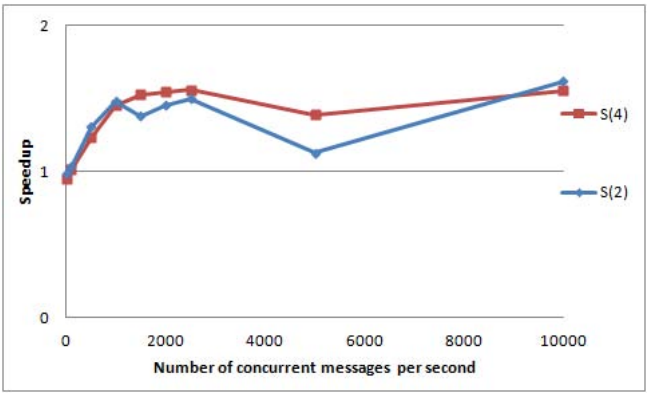

Fig. 4. Math web service speedup when scaled with factor $P = 2$ and $P = 4$ as a function of request rate N (in requests/second)

6 Analysis and Discussion

This section presents the speedup regions obtained by experiments by real scaling of the server load, varying one or both input parameters for Sort web service.

Figure 5 depicts the speedup for Sort web service hosted on a scaled system with 4 cores. Input parameter $M = 1K$ is constant and N varies. The experimental results confirm the theoretical analysis and we observe all three speedup regions (Drawback, Sublinear and Superlinear) and three scaling parameter regions (Underutilized, Proportional and Superior). The speedup begins to decrease and has a trend to saturate.

The speedup for Sort web service hosted on a scaled system with 2 cores is depicted in Figure 6. Input parameter $M = 2K$ is constant and N varies. The experimental results confirm the theoretical analysis and we clearly observe all three speedup regions and all four scaling parameter regions.

Figure 7 depicts the speedup for Sort web service hosted on a scaled system with 4 cores. Input parameter $N = 1000$ is constant and M varies. The experimental results also confirm the theoretical analysis and we clearly observe all three speedup regions and all four scaling parameter regions.

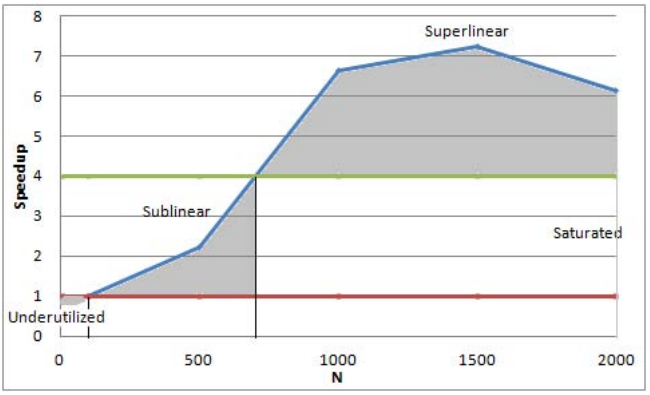

Fig. 5. Real speedup for Sort web service on a scaled system with 4 cores ($M = 1K$)

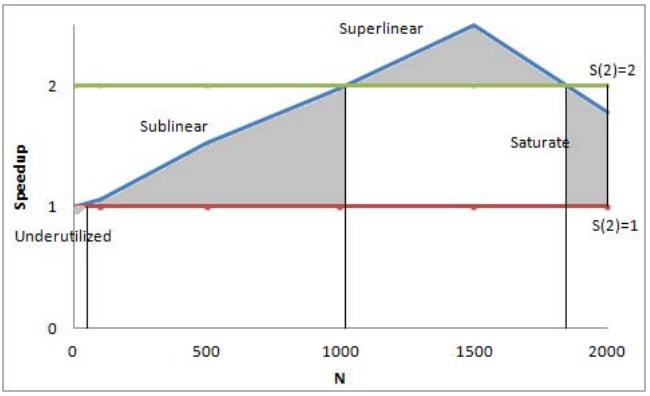

Fig. 6. Real speedup for Sort web service on a scaled system with 2 cores ($M = 2K$)

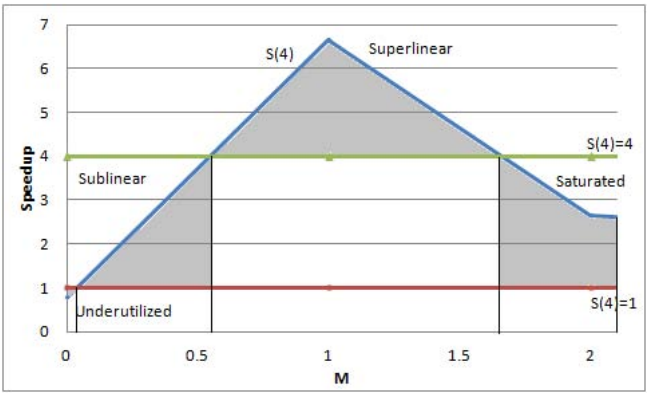

Fig. 7. Real speedup for Sort web service on a scaled system with 4 cores ($N = 1000$)

7 Conclusion and Future Work

This paper analyzes the performance of 3 different types of web services that utilize the cloud resources differently: *Concat* web service - memory-demanding only web service; *Sort* web service - memory-demanding and computation-intensive web service; and *Math* web service - computation-intensive only web service.

We have experimentally confirmed the hypothesis about typical performance behaviour of cloud web services. Although one would expect that cloud services can offer only decreased performance (due to increased I/O operations), we found existence of superior behaviour meaning that the web service hosted on the cloud performs much better when executed on more parallel resources. This is contrary to the existing law for bounded linear speedup given by Gustafson. We introduced an additional *Compute bound* dependency that appear for huge web service load.

Superlinear speedup is obtained only for Concat and Sort web services in the region where one CPU execution approaches 100% of utilization and saturates due to overload, but parallel execution on two and more cores can handle more load. Thus measuring the performance and comparing these cases we obtained the superlinear speedup effect. All three web services provide also drawback region when hosted on VMs with 2 and 4 cores for small number of concurrent messages. In this case the system is underutilized.

We found three different regions according to speedup increasing or decreasing for memory-demanding web services. For small load the response times are small for each number of processors providing sublinear speedup. Increasing the load, especially the number of concurrent messages, increases the speedup since the VM with 1 processor saturates and needs more time to compute all requests while a VM with more processors still works in normal mode. For huge load all VMs saturate providing huge response times which decreases also the speedup.

We will continue our research with other web services, such as chain web services or communicating with databases. We will also analyze on different cloud platforms, hypervisors, operating systems, and web servers. Although our research is conducted with small scaling factors $P \in \{2, 4\}$, we will analyze if our hypothesis is true for greater scaling factor and real world heterogeneous cloud. This modeling will be used to train the intelligent agent that will balance the load among cloud VMs in the superlinear region to maximize the performance.

References

1. Almeida, J., Almeida, V., Ardagna, D.: Cunha, ı., Francalanci, C., Trubian, M.: Joint admission control and resource allocation in virtualized servers. J. Par. Distr. Comp. 70(4), 344–362 (2010)
2. Birman, K.: Can web services scale up? Computer 38(10), 107–110 (2005)
3. Bonetta, D., Peternier, A., Pautasso, C., Binder, W.: S: a scripting language for high-performance RESTful web services. SIGPLAN Not. 47(8), 97–106 (2012)
4. Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., Weerawarana, S.: Unraveling the web services web: An introduction to SOAP, WSDL, and UDDI. IEEE Internet Comp. 6(2), 86–93 (2002)

5. Gusev, M., Ristov, S.: Superlinear speedup in Windows Azure cloud. In: 2012 IEEE 1st International Conference on Cloud Networking (CLOUDNET), Paris, France, pp. 173–175 (2012)
6. Gusev, M., Ristov, S.: A superlinear speedup region for matrix multiplication. In: Concurrency and Computation: Practice and Experience, pp. N/A–N/A (2013), http://dx.doi.org/10.1002/cpe.3102
7. Gusev, M., Ristov, S., Velkoski, G., Simjanoska, M.: Optimal resource allocation to host web services in cloud. In: Proc. of the 2013 IEEE 6th Int. Conference on Cloud Computing, CLOUD 2013, USA, pp. 948–949 (2013)
8. Gustafson, J., Montry, G., Benner, R.: Development of parallel methods for a 1024-processor hypercube. SIAM Journal on Scientific and Statistical Computing 9(4), 532–533 (1988)
9. Iakymchuk, R., Napper, J., Bientinesi, P.: Improving high-performance computations on clouds through resource underutilization. In: Proc. of the 2011 ACM Symposium on Applied Computing, SAC 2011, pp. 119–126 (2011)
10. Iosup, A., Yigitbasi, N., Epema, D.: On the performance variability of production cloud services. In: 11th IEEE/ACM Int. Symp. on CCGrid, pp. 104 –113 (May 2011)
11. Koh, Y., Knauerhase, R., Brett, P., Bowman, M., Wen, Z., Pu, C.: An analysis of performance interference effects in virtual environments. In: IEEE Int. Symp. on ISPASS 2007, pp. 200–209 (April 2007)
12. Ristov, S., Gusev, M.: Performance vs cost for Windows and Linux platforms in Windows Azure cloud. In: 2013 IEEE 2nd International Conference on Cloud Networking (CloudNet), San Francisco, USA (November 2013)
13. Ristov, S., Velkoski, G., Gusev, M., Kjiroski, K.: Compute and memory intensive web service performance in the cloud. In: Markovski, S., Gushev, M. (eds.) ICT Innovations 2012. AISC, vol. 207, pp. 215–224. Springer, Heidelberg (2013)
14. Wang, P., Huang, W., Varela, C.: Impact of virtual machine granularity on cloud computing workloads performance. In: 2010 11th IEEE/ACM Int. Conf. on Grid Computing (GRID), pp. 393–400 (October 2010)
15. Wang, W., Huang, X., Qin, X., Zhang, W., Wei, J., Zhong, H.: Application-level cpu consumption estimation: Towards performance isolation of multitenancy web applications. In: 2012 IEEE 5th Int. Conf. on Cloud Computing (CLOUD), pp. 439–446 (June 2012)

Urban Policy Modelling: A Generic Approach

Marjan Gusev[2], Goran Velkoski[1], Ana Guseva[1], and Sasko Ristov[2]

[1] Innovation LTD, Vostanicka 118, 1000 Skopje, Macedonia
{goran.velkoski,ana.guseva}@innovation.com.mk
[2] Ss. Cyril and Methodious University, FINKI
Rugjer Boshkovikj 16, 1000 Skopje, Macedonia
{marjan.gushev,sashko.ristov}@finki.ukim.mk

Abstract. This paper presents a methodology and case studies of modelling an optimisation intended to support the urban policy at cities, as a part of the research and development activities within the FP7 FUPOL project. The innovations are driven by both the demand of citizens and political decision makers in order to support the policy domains in urban regions based on ICT applications in urban policy modelling. The case study presents the mathematical model and modelling the optimisation, such as the best low cost project to be financed with the highest impact among several candidates. Urban planning optimisation functions can be deployed by complex mathematical processing or by simulation and visualisation models. The main contribution is in the generalised procedure to derive a proper optimisation model.

Keywords: FUPOL, Optimisation, Urban Policy, Computer simulation.

1 Introduction

Assume there is a list of possible proposals for a given urban environment that will improve citizen satisfaction and wellbeing. Usually, local authorities struggle with limited budgets and have to decide which projects will have the highest impact and are candidates to be financed and realised. The goal in this project is the decision making process on how to choose the best solution among several candidates. It is a typical optimisation problem and we present a generic procedure for the use case of urban policy modelling.

Optimisation problems may be solved in several ways, and basically, there are at least three approaches: solving complex mathematical systems of equations and inequalities; using artificial intelligence [2]; or by using the simulation and visualisation approach.

Solving relatively complex optimisation problems is a difficult programming task from one side, and time consuming task from the other side. A careful analysis has to be undertaken to analyse the processing demands and obtain results in real time. This approach usually finishes with introducing more constraints and limits to eliminate the redundant processing.

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_19

The simulation and visualisation approach is used recently in several research oriented projects, such as, the FUPOL project [3]. This project proposes a comprehensive new governance model to support the policy design and implementation life-cycle. The innovations are driven by the demand of citizens and political decision makers to support the policy domains in urban regions with appropriate ICT technologies [10]. It specifically targets domains such as sustainable development, land use, urban planning, urban segregation and migration.

However, the same principles used in problem analysis for the computer processing approach is also used for the simulation and visualisation approach. It is essential for both approaches, since it defines the real goal and objectives by establishing a proper mathematical model. The first approach is based on providing an exact mathematical solution using IT processing tools to solve a set of inequalities, and the second approach builds an information technology model, upon which simulation and visualisations provide results.

The rest of the paper is organised as follows. The generic approach for optimisation model is presented in Section 2 and case studies in Section 3 about Vodno recreational activities and bike usage as transport means at City of Skopje. The discussion about FUPOL approach and conclusions are specified in Section 4.

2 A Generic Approach

The generic model is developed to unify several modelling objectives under one umbrella. It should be an approach to develop a methodology addressing the relevant needs and different perspectives in the urban policy. The following issues have been identified for the urban policy: resource capacity, conflict resolution solution, and selection of the best project proposal with highest impact. To start with the analysis for modelling purposes, we define the necessary actors and environment:

- *Users* are groups of people with similar preferences about using the urban environment, identified by U_m, for $m = 1, \ldots, U_{max}$, where U_{max} is the number of user groups;
- *Activities* performed by the users for recreational and everyday living, identified by A_i, for $i = 1, \ldots, A_{max}$, where A_{max} is the maximum number of activities; and
- *Resources* from the urban environment, where the users are performing their activities identified with R_j, for $j = 1, \ldots, R_{max}$, where R_{max} is the maximum number of resources.

There are several relations among user groups, activities and resources:

- *Number of users per user group*, denoted as a function $NU(U_m)$, for the user group U_m.
- *Connectivity*, denoted by a boolean function $CN(U_m, A_i, R_j)$, where true value (or 1) is identified if a user group U_m performs an activity A_i using the resource R_j; and otherwise false (or 0).

2.1 User Analysis

Further analysis targets a specific time slot t_k, where $k = 0, 1, \ldots, 23$ and all the analysis will use average values calculated for the corresponding time slot. The maximum number of users $Nmax_{ij}$ that can ideally perform the activity A_i on a resource R_j can be calculated by (1), where $i = 1, \ldots, A_{max}$ and $j = 1, \ldots, R_{max}$.

$$\forall A_i, R_j \quad Nmax_{ij}(t_k) = \sum_{m=1}^{U_{max}} NU(U_m)CN(U_m, A_i, R_j), \tag{1}$$

Next we analise the number of users within a given time frame t_k. Not all users would perform an activity A_i on the resource R_j. It always depends on the following *willingness* factors:

- *temperature*, presenting the heat conditions for a given activity;
- *weather*, addressing the actual climate conditions, like sunny, cloudy, rainy, storm etc.;
- *timing*, determined by the time frame in a day when the activity can be performed;
- *working availability*, identified by the fact if the user group is available, such as, a recreation activity might be performed if it is weekend; and
- *attractiveness*, presenting the resource quality to perform a corresponding activity.

These factors are addressed by the following general input coefficients:

W - Weather conditions index is associated to the actual *weather* factor and determined by the weather type, such as sun, rain, wind, storm, etc. It's value ranges from 0 to 1. Value 0 presents that nobody will perform an activity due to severe weather conditions like storm. 1 is assigned to bright weather, meaning that everybody from the given user group would like to perform an activity on the given resource.

T - Time of the day index presents the *timing* factor and is associated to a time slot, ranging from 0 to 1. Time slots with frequent usage are assigned with higher values, where the maximum is 1. For example, if an activity is not performed at all, such as during the night, then 0 is assigned to this coefficient.

D - Day of the week index associated to the *working availability* is used to determine when the activity is performed more frequently. It ranges from 0 to 1, where 0 means no activity is performed in the particular day and 1 that all users in the group will perform the activity.

M - Monthly average temperature index addresses the *temperature* factor. It is actually presenting the dependence on the average temperature, ranging from 0 to 1, where 0 means that the average temperature is so low or so high, that nobody would like to perform the corresponding activity, and 1 that everyone form the analysed group is willing to perform the analysed activity.

Q - Quality index is actually presenting the *attractiveness* factor. The values are within the range between 0 and 1, where 0 means that the resource infrastructure is not suitable for the analysed activity and 1 that the resource infrastructure is ideal for performing the activity.

2.2 Resource Capacity

Using the defined willingness factors and introduced coefficients, the number of users N_{ij} from a group U_m that perform an activity A_i on a resource R_j in a given time slot t_k can be calculated by (2).

$$N_{ij}(t_k) = W \cdot T \cdot D \cdot M \cdot Q \cdot Nmax_{ij}(t_k) \tag{2}$$

Each urban policy addresses activities on a given resource. Denote by RC_j the capacity of the resource R_j measured for an one-hour time frame, where $j = 1, \ldots, R_{max}$. The resource capacity RC_j on a resource R_j will be reached if the corresponding number of users NR_j, for each activity A_i reaches its maximum, where $i = 1, \ldots, A_{max}$. The sufficient conditions are explicitly defined by the inequality in (3).

$$NR_j(t_k) = \sum_{i=1}^{A_{max}} N_{ij}(t_k) \leq RC_j \tag{3}$$

2.3 Conflict Resolution

Conflicts may occur if two opposing activities are to be performed on the same resource. The following analysis starts by denoting a pair of opposing activities A_i and A_l, where $i, l = 1, \ldots, A_{max}$. The function $CM(A_i, A_l)$ is a boolean function that shows if two activities are conflicting or not, where false (0) stands for no conflicting activities and true (1) for conflicting activities.

(4) defines if there is a conflict between users from the user group U_{mi} performing an activity A_i and users from the user group U_{ml} performing and activity A_l performed on the same resource R_j, where $mi, ml = 1, \ldots, U_{max}$, $i, l = 1, \ldots, A_{max}$ and $j = 1, \ldots, R_{max}$.

$$FN_{i,j,l} = CN(U_{mi}, A_i, R_j)CN(U_{ml}, A_l, R_j)CM(A_i, A_l) \tag{4}$$

We are interested to calculate the number of users $NC_j(t_k)$ that perform conflicting activities on a resource R_j, where $j = 1, \ldots, R_{max}$ in a given time slot t_k, where $k = 0, 1, \ldots, 23$. A naive interpretation may lead to a conclusion that the value can be presented as a sum of the corresponding users in conflicting user groups. Let use case 1 consist of 3 users in one user group performing activity A_i, which are in conflict with 5 users from another group performing activity A_l on the same resource. Also assume that in use case 2 there are 2 persons performing A_i which are conflicted with 6 persons performing A_l. There is a typical question raising on this naive conclusion, what makes a worse solution, the number of conflicting users or the number of conflicts. For example, in both use cases we can conclude that 8 users are conflicting, but there are 15 conflicts in use case 1, and 12 in use case 2. In this paper, we introduce a *conflict measure function* for a given resource R_j, calculated by (5), corresponding to their product, instead of sum.

$$NC_j(t_k) = \sum_{i=1}^{N_A} \sum_{l=1}^{N_A} FN_{i,j,l} N_{ij}(t_k) N_{lj}(t_k) \tag{5}$$

The total sum of these functions for all resources is a relative conflict measure NC_{all} that evaluates the number of conflicts for the analysed use case. The lower the value, the less conflicts are found. (6) defines how it can be calculated for all resources.

$$NC_{all}(t_k) = \sum_{j=1}^{N_R} NC_j(t_k) \tag{6}$$

2.4 Environmental Protection

We have included several environmental protection parameters in our model, which does not exclude the possibility to include more for a specific urban model. We found that pollution, noise pollution, criminal safety, physical safety, natural preservation and cleanliness are the most commonly used. The next definitions address their influence on the model.

Contamination (pollution) is the parameter that refers to the quantity of contaminants into the natural environment that cause adverse change. To model the influence, we define the contamination (pollution) coefficient C_{PL} for a specific activity A_i and resource R_j. In this study we do not analyse the real parameters, such as sulphur dioxide, nitrogen dioxide, particulate matters, carbon monoxide, ozone, or similar. For simplicity in our model we define just the relative ratio of average behaviour that the activities impact on environment pollution. Therefore the values of C_{PL} in our analysis ranges between 0 and 1, meaning that the minimal value of 0 produces no pollution, while the maximal value 1 corresponds to an activity that produces maximum pollution. For example: walking causes 0 contamination, while motorbikes cause 1 contamination.

The overall pollution function for each analysed user group U_m, where $m = 1, \ldots, U_{max}$ is calculated by (7) for a given time frame t_k, where $k = 0, 1, \ldots, 23$.

$$PL = \sum_{i=1}^{N_A} \sum_{j=1}^{N_R} C_{PL}(A_i, R_j) N_{ij}(t_k) \tag{7}$$

Noise pollution is a factor which addresses sounds that an activity is producing and their effect on the environment. Similarly to the previous case, the noise pollution coefficient C_{NP} depends on activity A_i and resource R_j. The values associated to the noise pollution are proportional to the produced sound levels and C_{NP} gets relative values in the range $[0, 1]$, where 0 corresponds to the minimum sound levels, while 1 to the maximal.

The overall noise pollution function is calculated by (8) if the user group U_m is analysed in a given time frame t_k, where $m = 1, \ldots, U_{max}$ and $k = 0, 1, \ldots, 23$.

$$NP = \sum_{i=1}^{N_A} \sum_{j=1}^{N_R} C_{NP}(A_i, R_j) N_{ij}(t_k) \tag{8}$$

Physical Safety factor C_{PS} addresses users within the user group U_m, where $m = 1, \ldots, U_{max}$. It is a relative factor where the values range between minimal

value 0, meaning a situation there is no physical danger while they are performing the activity A_i on the resource R_j and 1 reaches maximum danger caused by conflicts or other factors. The overall physical safety function is calculated by (9) for the users from the user group U_m in a given time frame t_k, $k = 0, 1, \ldots, 23$.

$$PS = \sum_{i=1}^{N_A} \sum_{j=1}^{N_R} C_{PS}(A_i, R_j) N_{ij} \tag{9}$$

Natural preservation factor expresses how users from a user group U_m preserve the nature while performing activities. The nature preservation relative factor C_{NR} gets values from 0, meaning that the activities do not preserve nature, up to 1, meaning that activities preserve the nature, in a most efficient way. When users from the user group U_m perform activities in a given time frame t_k, where $m = 1, \ldots, U_{max}$ and $k = 0, 1, \ldots, 23$, the natural preservation function is calculated by (10).

$$NR = \sum_{i=1}^{N_A} \sum_{j=1}^{N_R} C_{NR}(A_i, R_j) N_{ij}(t_k) \tag{10}$$

There are functions that can also be calculated in a similar way, such as crime safety, cleanliness, etc.

2.5 Selection of the Best Project Proposal

To find the best project, one needs a clear definition of project proposal and its costs. The process of selection of the best project proposal is calculated by evaluating several optimisation functions.

The project proposal cost function is a business related function, which relates to the cost C of a given project P_h, where $h = 1, ..., N_{projects}$ is an integer expressing one of the project proposals. Usually the budget B is limited and only a subset of project proposals can fit in the budget. Therefore, the cost function limitations is expressed by (11), as a function that fits a selection of N_h projects, where the set H expresses the selected projects.

$$\sum_{h \in H} C(P_h) \leq B, \tag{11}$$

(11) presents an optimisation to select a subset of a given set with defined costs that fit in a given budget. Finally, the selection of the best project is an optimisation problem, where one would like to have the highest benefit, usually reached as the highest number of users calculated by (2) and the projects with lowest pollution and noise pollution, calculated by (7) and (8) and the projects with highest natural preservation, calculated by (10) and (9). In addition to these functions, (12) expresses minimum conflicts to be reached, defined by (6).

$$Max(N_{ij}(t_k)), Min(NC_{all}(t_k)), Min(PL), Min(NP), Max(PS), Max(NR) \tag{12}$$

The constraints for these optimisation functions are summarised in (13).

$$C_{PL}, C_{NP}, C_{PS}, C_{NR} \in [0, 1] \tag{13}$$

Choosing a subset of projects that satisfies the constraints (11) and (13) with the highest benefit (12) is the final optimization goal. The next section presents case studies and implementation of this approach.

3 Case Studies

In this section we present two case studies. The first case study is about optimisation model that aims at organising the recreational activities at Vodno Mountain while preserving its natural environment and avoiding conflicts as much as possible. The second case study refers to activities that increase the overall bike users in City of Skopje.

3.1 A Case Study: Modelling the Vodno Recreation Activities

City of Skopje aims at minimising the conflicts when different user groups perform recreational activities on the Vodno mountain. The expected output is an optimised schedule of different recreational activities on the Vodno Mountain, satisfying the defined constraints.

According to the Federal Highway Administration and the National Recreational Trails Advisory Committee technical report [12], there are 12 principles for minimizing conflicts on multiple-use trails: 1. Recognize Conflict as Goal Interference; 2. Provide Adequate Trail Opportunities; 3. Minimize Number of Contacts in Problem Areas; 4. Involve Users as Early as Possible; 5. Understand User Needs; 6. Identify the Actual Sources of Conflict; 7. Work with Affected Users; 8. Promote Trail Etiquette; 9. Encourage Positive Interaction Among Different Users; 10. Favor "Light-Handed Management"; 11. Plan and Act Locally; and 12. Monitor Progress.

For the purposes of the FUPOL project, in [6] we have defined $A_{max} = 36$ recreational activities, a total of $R_{max} = 81$ resources including locations, transport facilities, parking lots, hiking trails etc. To identify the occurrences of conflicts, we have analysed $U_{max} = 89$ user groups with different preferences, and established their typical behaviour by specifying timing, activity and resource.

The fundamental conflict we want to resolve is that some activities are incompatible with others. Finally, we define operational rules under which conflicts between the defined activities occur. They are listed as follows:

- The resources should be used economically, because the nature should be protected and preserved;
- Mountain bikers use the tracks for riding at high speeds which poses a danger for climbers and especially for families with little children who'd like to run free and carelessly;
- Motorbikers drive at high speeds and disturb all other recreationists;

- Hiking and trekking trails should be as less crowded as possible;
- Skiers can ski only when there's snow (December-February);
- Mountain-bikers do not use Vodno when there's snow hikers, trekkers, climbers are in much smaller number when there's snow; and
- Number of visitors is increased if weather is nice, on weekends, especially in summer when it is very hot in the city, or when after severe cold weather, the sun offers nice stay in Vodno.

In addition to the established environmental protection functions, in the Vodno case study we use the cleanliness and crime safety index parameters. The final goal is to minimise pollution and noise pollution, and maximise physical and crime safety; cleanliness and natural preservation, presented by (7), (8), (9) and (10). The essential optimisation function is to maximise the number of users calculated by (2) keeping the minimum conflicts, calculated by (6). The Vodno model does not require the selection of the best project proposal, rather it specifies a schedule of resource usage [6].

3.2 A Case Study: Modelling the City of Skopje Bike Usage

The plans of the administration of City of Skopje include several measures and activities to provide a healthy environment for the citizens of Skopje [11] and increase the overall bikers of all those requiring transport. A lot of project proposals have been specified, such as those that increase the quality of bike infrastructure, or those that build new bike paths, introduce bike docking stations for bike renting and bike parking, or proposals that foster bike intermodality.

In contrast to the Vodno model, this model aims at selection of the best project proposal due to the limited budgets. The goal is to find a subset of the proposals that will increase the number of bikers the most, and hopefully reach the target of 5% bikers of all commuters (European city average). The starting point is low, as recent research shows that bike usage in Skopje is between 1.4% to 2.5% [4], [8], [9]. The optimization model [7], realised for the purposes of the FUPOL project aims to find out:

- Optimal selection of projects that will increase the bicycle infrastructure;
- Optimal number of docking stations, their distribution and size (besides locations we consider sizing of slots for private bikes parking and bike rental);
- Optimal selection of new bike paths by building connections between stations (which one connects with which other) considering an approximation of the distance between them; and
- Optimal selection of improving the existing bike paths.

This model does not specify different activities, since only one is analysed: the bike usage a as transport mean. Resources are identified by 268 locations, and 363 bike tracks with all characteristic points that determine their exact map path. The number of user groups identified is $U_{max} = 77$. The model uses information about 9012 possible transport needs between two different locations [8].

The model [7] uses only the attractiveness, temperature and weather coefficients. It does not need to use the conflict resolution functions, since we assume that bike related resources will not be used by other users.

In addition to the already described functions, in this model we introduce another very important function T_{avg}, which defines the average transport time between selected predefined locations, measured as an average of times for several predefined trips in rush times, or those time intervals where there is increased traffic jam, such as 8-9h in the morning or 16-17h in the afternoon.

The overall optimisation function is defined as: Find a subset of projects $S_{opt} \subseteq S$ by optimising $Min(T_{avg}, PL, NP)$ and $Max(N_{ij})$ subject to constraints defined by (11) and (13).

4 Conclusion

The citizens' participation in the decision-making processes is deemed fundamental to promote sustainable development. Particularly it is on the local level, where citizens live and work, where basic services are provided and where enterprises are established. Citizens have, therefore, common interests at stake, to set objectives and work together in identifying solutions particularly aiming at improved access to services, a more balanced distribution of available resources, greater social cohesion and enhanced accountability and transparency of public authorities, including to accountability mechanisms.

One of the most important challenges of Local Authorities is certainly the sustainable urbanisation, being an important aspect of good local governance. Because of the rapid growth in urban population, Local Authorities in urban areas have a decisive role to address the challenges related to urbanisation, such as the needs of citizens living without adequate services and facilities. Use of tools to boost citizen participation and support decision making is essential [1].

A typical optimisation function to find a project with minimal costs that will have the highest impact out of several proposals, is an essential support of citizen involvement in policy design processes and in helping the local authorities to decide.

The conventional processing approach can be built on the "try all possibilities" algorithm for the system of inequalities, finding the optimal schedule or subset of project proposals. Since it is a time consuming function, with exponential complexity growth, we suggest using a greedy algorithm, which first calculates the optimisation functions of all individual projects and ranks the projects according to the optimising functions and fitting the cost function in the budget.

The FUPOL project uses the simulation and visualisation approach. It is based on the same mathematical model and uses user agent based simulation, based on the user preferences. The simulation takes specified weather conditions and temperature and calculates the number of the citizens that are willing to perform an activity on a given resource. The time slots for specifying the agent behaviour is elected to be 1 minute, or 5 minutes, depending on the computation processing power of the server. The visualisation uses this information

and presents a visual map of resource occupancies. Details about simulation and visualisation results are presented in [5]. Citizens can participate in decision making processes by using the developed simulation and visualization tools, providing essential input for Local Authorities. Out of the scope in this paper are project management parameters, such as time and quality of project life-cycle, identification of project stakeholders, risk management, etc.

In this paper we have presented a generic approach of urban policy modelling, by identifying actors (user groups), activities and resources, their connectivity and conflicting nature. We have introduced willingness factors, defined corresponding coefficients and optimisations functions. The defined constraints present resource capacity and conflict resolution functions. In addition to optimisation of environmental protection functions, we defined the procedure of selection of a subset of project proposals that fit in a given budget and optimising the related functions. Two case studies were presented using this approach.

References

1. Bazerman, M., Moore, D.A.: Judgment in managerial decision making. John Wiley & Sons, Inc. (2012)
2. De Oliveira, M., De Almeida Neto, A.: Optimization of traffic lights timing based on multiple neural networks. In: 2013 IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 825–832 (2013)
3. FUPOL: project web site (2014), http://www.fupol.eu/
4. GUP: General Urbanistic Plan of Skopje. Tech. rep., City of Skopje (2012)
5. Gusev, M., Velkoski, G., Avukatov, A., Ringov, M., Markic, L., Apostolova, M.: ICT tools for Policy Design in City of Skopje. In: Proc. 11th CiiT Conf. on Informatics and Information Technologies.s UKIM FINKI (2014)
6. Gusev, M., Veselinovska, B., Guseva, A., Gjurovikj, B.: Future Policy Modeling: A case study – Optimization of recreational Activities at the Vodno Mountain. In: Advanced ICT Integration for Governance and Policy Modeling. IGI Global (2014)
7. Guseva, A., Gusev, M., Veselinovska, B.: Fostering Bicycle Inter Modality in Skopje. In: Advanced ICT Integration for Governance and Policy Modeling. IGI Global (2014)
8. IDORM: Transport Master Plan for Greater Skopje. Tech. rep., City of Skopje (2010)
9. JP Ulici i Patista Skopje: Web site (2014), http://www.uip.gov.mk/
10. Sonntagbauer, P., Boscolo, P., Prister, G.: Fupol: an integrated approach to participated policies in urban areas (2012), https://www.fupol.de/sites/default/files/doc/eChallenges
11. Stefanoski, I.: Drafting the Bicycle Master Plan for Skopje. Tech. rep., City of Skopje (2004)
12. The Federal Highway Administration and The National Recreational Trails Advisory Committee: Conflicts on multiple-use trails. Tech. rep (2003)

Robustness of Speech Recognition System of Isolated Speech in Macedonian

Daniel Spasovski[1], Goran Peshanski[1], Gjorgji Madjarov[2], and Dejan Gjorgjevikj[2]

[1] Netcetera, Skopje, Macedonia
spasovski.daniel@gmail.com, pesanski_goran@yahoo.com
[2] Faculty of Computer Science and Engineering, Skopje, Macedonia
{gjorgji.madjarov,dejan.gjorgjevikj}@finki.ukim.mk

Abstract. Over five decades the scientists attempt to design machine that clearly transcripts the spoken words. Even though satisfactory accuracy is achieved, machines cannot recognize every voice, in any environment, from any speaker. In this paper we tackle the problem of robustness of Automatic Speech Recognition for isolated Macedonian speech in noisy environments. The goal is to exceed the problem of background noise type changing. Five different types of noise were artificially added to the audio recordings and the models were trained and evaluated for each one. The worst case scenario for the speech recognition systems turned out to be the babble noise, which in the higher levels of noise reaches 81.10% error rate. It is shown that as the noise increases the error rate is also increased and the model trained with clean speech, gives considerably better results in lower noise levels.

Keywords: speech recognition, robustness, isolated speech, signal-to-noise ratio, background noise.

1 Introduction

During the last few years, there is increased reliance and use of the automatic speech recognition systems. They tend to be standard part of our daily life and people gradually get used to it. The crucial reason is the voice oriented interface, which makes the man-machine interaction (MMI) very natural and straightforward for users. The automatic speech recognition systems are widely used into medicine, in the military, by the people with disabilities, in remote control systems, dictation services, telephone operators, automotive industry etc.

Automatic speech recognition is a process of converting the audio signal in sequence of text symbols. Generally, a speech recognition system is consisted of: audio signal processing, signal decoding and adaptation. The speaker produces sequence of words which are transmitted through the communication channel, whereby the waveform of the audio signal is generated. The waveform is then forwarded to the speech recognition component where a parameterized acoustic signal is obtained. By using stochastic methods, the speech decoding component transforms the parameterized acoustic signal into sequence of symbols.

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_20

Most of the speech recognition systems are constructed to be speaker independent, i.e. the system can be used from various users without the need of an adaptation to a specific user. The other portion of the commercially used speech recognition systems, require the speaker to assist in the process of training the model in order to achieve better accuracy for the speaker. Thus, the system is adapted on the voice characteristics of the speaker and the background environment. Many of the systems reach accuracy of 98-99% in ideal conditions. The ideal conditions are accomplished if the speaker voice characteristics are similar to the training data. Then the speaker can be adapted to the system and the system can operate in silent environment (environment without noise).

Doing a research in this field where the noisy environments can be compared and evaluated, and the influence of the different noises in the process of speech recognition can be expressed, is significant. It is significant to determine the critical noises that we meet every day. Here, the main representatives are the natural noises (babble noise, rain and wind noise), the industrial noises (traffic noise) and from scientific perspective the synthetic noises (white noise). The experiments were made with audio database of isolated speech in Macedonian. The database includes the 20 most frequently occurring names in Macedonia, with 2000 recordings for training the acoustic model and 845 recordings for testing the model. In many related work it is proved that the recognition of the isolated speech is independent from the subject (topic) of the audio samples.

It is interesting to designate the level of the signal to noise ratio (SNR) where a degradation of the speech recognition system can occur. Since SNR is can be easily measured and the range and the boundary in which the systems can achieve satisfactory results can be determined. The experiments are examined in cross-noise environments where the model is trained on one level of noise, and tested on different levels. It indicates the process of training the model for greater robustness if we know the noisy environment and the level of noise.

Sometimes, it is useful to compare the error rate in speech recognition systems relative to the level of the noise, in different environments. It is shown that the different environments cause changes to the error rate, when the level of noise is changed.

Here, only the impact of the noise to the speech recognition systems is presented. For complete picture of the system behavior, additional research is required in terms of the channel modification, and more generally the change of the channel.

In Section 2 a review of the related work is shown. Section 3 describes the datasets and the experimental setup. The experimental results are presented and discussed in Section 4. Finally, the conclusions are given in Section 5.

2 Related Work

The first evaluation for Speech in Noisy Environments (SPINE1) was conducted by the Naval Research Labs (NRL) in August, 2000. The purpose of the evaluation was

to test existing core speech recognition technologies for speech in the presence of varying types and levels of noise.

In terms of robustness of the speech recognition systems a lot has been done before, especially in the early 90'. P. Schwarz in his doctoral thesis [1] has presented some of the properties of the speech recognition systems in different condition. Results of many studies have demonstrated that automatic speech recognition systems can perform very poorly when they are tested using a different type of microphone or acoustical environment from the one with which they were trained [2].

Most of the proposed techniques for acoustical robustness are based on signal processing procedures that enable the speech recognition system to maintain a high level of recognition accuracy over a wide variety of acoustical environments [4] [7, 8, 9, 10]. Further, more accurate models were proposed with empirically-based methods and model-based methods, using a variety of optimal estimation procedures [11].

Some speech systems have been proposed for Macedonian language. One noticeable example is the speaker dependent digit recognition system that recognizes isolated words, designed by Krajlevski et al. [5, 6]. They proposed hybrid HMM/ANN architecture and achieved accuracy of around 85%.

3 Datasets and Experimental Setup

It was interesting to verify the behavior of the acoustic model with best performance in noisy environment. Thus, the existing test data samples were processed, and modified test data samples were produced by adding different type and level of noise (signal-to-noise ratio). The experiments were performed with Sphinx [12, 13].

3.1 Datasets

The audio database consists of audio recordings collected with an online web application. The experiments are performed on recordings from isolated speech with 20 different proper names in Macedonia, where more than 40 speakers were included. The most frequently used names in Macedonia were used. It is composed of 2845 audio samples of the proper in Macedonia from 50 male and 50 female samples for each name. The samples are collected with an online web application made especially for this purpose. 2000 of the samples make the training set and the other 845 make the test set, which withdraws around 50 different samples from different speakers for each name. The training dataset is used for training the acoustic model, while the performance of the acoustic model is evaluated on the corresponding test set.

The fact that the samples are collected with a web application, recorded with many different microphones in different conditions, indicates that the dataset is a relevant represent of the environment where such name recognition system would be used. The training and test audio recordings have the following quality: 8 kHz, 16 bit, 1 mono channel.

3.2 Experimental Setup

The next part of the process was to create language model for the recognizer. The language model consists of the words for the most frequent proper names in Macedonia, phonetic dictionary and fillers, which represent "empty" speech or silence. The phonetic dictionary contains of the phonemes the words are built from.

For the creation of the HMM based name recognition system in Macedonian we used the Sphinx framework. It consists of part that provides feature extraction, modeling and training an acoustic model, modeling a language model and a part for decoding that enables to test the performance of the system.

The feature extraction part separates the speech signal into overlapping frames and produces feature vector sequences each containing 39 MFCC features (coefficients). The 39-dimensional MFCC vector is created of the first 13 MFCC coefficients, 13 features that represent the speed of the signal (the first derivate of the first 13 MFCC coefficients) and 13 features that represent the acceleration of the signal (the second derivate of the first 13 MFCC coefficients). This 39-dimensional feature vector, which is the most widely used in speech recognition, is used as the basic feature vector for our speech recognition system in Macedonian.

We used the Baum-Welch algorithm (integrated in Sphinx) for the training of the acoustic models. This algorithm finds i.e. adjusts the unknown parameters of a HMM.

The performance of all acoustic models is evaluated by measuring the Word Error Rate - WER (equation 1).

$$WER = \frac{\#\ correctly\ recognized\ words}{\#\ number\ of\ samples\ in\ the\ test\ set}\ [\%] \tag{1}$$

We consider this evaluation relevant because of the versatile nature of the test set which, in some way, tells us how the acoustic models would behave in natural environment.

4 Results

4.1 Results in Noisy Environments

The performance of the model was evaluated on 4 different signal-to-noise (SNR) ratio levels and 5 different types of noise. The natural noises are represented with the rain, wind and babble noise. The main representative of industrial noises is traffic noise, and the synthetic noises are represented by the white noise. The results are shown on Fig. 1.

The obtained results show almost linearly relation between the error rate and the noise level. The error rate increases proportionally with the noise level, which indicates that the acoustic model changes with the same intensity as noise level is added to the test set, without any dependence from the type of the noise. It means that any model is capable of learning the patterns of the noise better than the other. The degradation of WER, for the different models is almost constant.

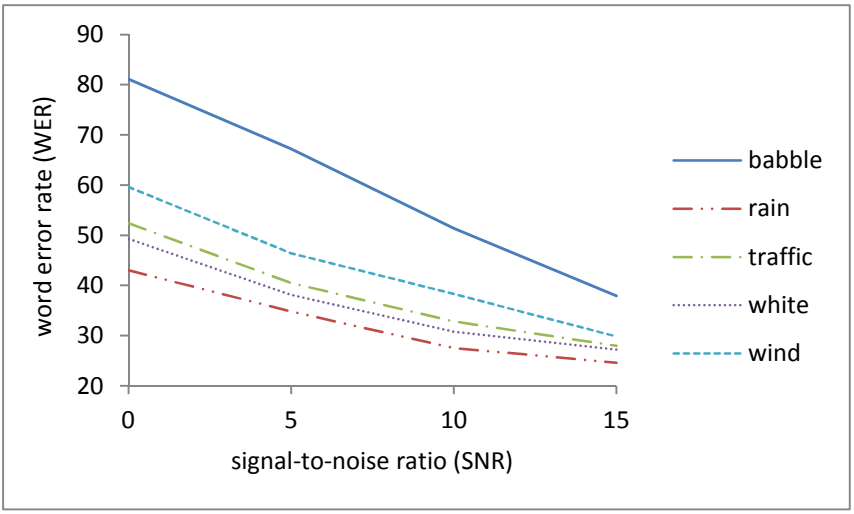

Fig. 1. Robustness of HMM based model for speech recognition with model trained on clean speech

The worst case scenario for the speech recognition systems turned out to be the babble noise, which in the higher levels of noise reaches 81.10% error rate (Table 1). It is obvious that the accuracy, almost for nearly every model linearly decreases as the noise increases. The rain noise doesn't affect the accuracy in the lower levels of noise. The other noises have similar tendencies. Quantization of information is useful technique for robustness improvement.

Table 1. Robustness of HMM model in speech recognition compared to the noise with model trained on clean speech, in noisy test environments with babble, rain, traffic, wind and white noise

SNR	babble	rain	traffic	white	wind
0	81.10	43.00	52.40	49.30	59.60
5	67.20	34.80	40.50	38.10	46.40
10	51.40	27.50	32.80	30.80	38.30
15	37.90	24.60	28.00	27.20	29.80

4.2 Results in Cross-Noise Environments

In the previous experiments the explored models were built in one training conditions and tested in many different test conditions. The training and testing conditions were invariant from the type of noise. The experiments are based on babble and traffic noise, which are one of the most common noises. The models are trained on five levels of noise, and are evaluated on each one of them. Table 2 and Table 3 present

the absolute WER. The obtained results show that to ensure particular WER, it is always better to train the system with greater noise level. The system than is capable for recognition of less distorted patterns with satisfactory WER.

Table 2. Robustness of HMM based system with noisy trained model with babble noise for different test and train noise level. The training conditions are in rows, and the testing conditions are in columns. The same training and test conditions are marked.

HMM	SNR0	SNR5	SNR10	SNR15	Clean
SNR0	**87.3**	71.6	64.4	58.8	66.6
SNR5	66.6	**57.9**	51.1	48.3	58.6
SNR10	61.5	49.6	**43.8**	42.8	53.3
SNR15	66.0	53.8	45.3	**42.4**	53.4
Clean	81.1	67.2	51.4	37.9	**26.0**

However, this doesn't work always, especially when the target conditions are clean speech. The patterns which the classifier is observing become totally different. In the clean speech and the invoiced parts of the speech, the logarithm of the energy may be close to $-\infty$.

Table 3. Robustness of HMM based system with noisy trained model with traffic noise for different test and train noise level. The training conditions are in rows, and the testing conditions are in columns. The same training and test conditions are marked.

HMM	SNR0	SNR5	SNR10	SNR15	Clean
SNR0	**47.9**	41.8	45.6	49.0	64.6
SNR5	41.5	**37.9**	34.3	35.3	47.6
SNR10	59.3	48.8	**42.8**	39.6	41.9
SNR15	78.0	64.0	55.1	**47.7**	48.2
Clean	52.4	40.5	32.8	28.0	**26.0**

In Table 2 the relative values of WER (babble noise) can be calculated where the noise level matches. WER on the current noise level is subtracted from WER of the training noise level. Obtained value greater than zero means the model trained on particular noise level (rows) can be applied successfully to the actual testing noise level (columns). The values above the main diagonal are cases where the training noise level is greater than the testing noise level. The error rate for the models trained with SNR 10 and SNR 15 is significantly improved when tests are performed over samples with SNR 0, 5, 10, 15. Worst results again tend to show models trained with SNR 0 and SNR 5.

The model which is trained with higher level of noise can be used in lower noise level conditions, with particular tradeoff in WER.

The WER of 26% on clean speech in normal environment is not so satisfactory. But, in term of isolated speech it is acceptable regarding the conditions in which the training set samples were collected.

The performances of the model trained with clean speech are very interesting as well. The same experiments applied with traffic noise, show similar behavior, but the results obtained are considerably better (Table 3). The behavior is almost the same as the former trained model with babble noise, which indicates that the models behave in similar fashion regardless the type, or the source, of the noise.

5 Conclusion

The discussion presented here illustrates the theory that it is better to train models on lower SNR to ensure WER for better SNR. However, to ensure fully robustness it is necessary to examine the behavior of the model over the changes in the transmission channel or more generally the change of the channel.

Here is also shown that as the noise increases the error rate is also increased and the model trained with clean speech, gives considerably better results in lower noise levels. That's the reason why this type of trained model mainly is used by the speech recognition systems.

In our future work we will try to evaluate the different transmission channel. We will collect more training and evaluation data. It is interesting to experiment with other types of background noise, and with continuous speech. Our final goal is to create different recognizers that will be robust on particular noise level, independent from the type of noise.

References

1. Schwarz, P.: Phoneme recognition based on long temporal context. PhD Thesis, Faculty of Information Technology. Department of Computer Graphics and Multimedia, Brno University of Technology 47–60 (2008)
2. Acero, A.: Acoustical and Environmental Robustness in Automatic Speech Recognition. Phd. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania (1990)
3. Sethy, S.N., Parthasarthy, S.: A split lexicon approach for improved recognition of spoken names. Integrated Media Systems Center, Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, United States. AT&T Labs-Research, Florham Park. Speech Communications 48(9) (2006)
4. Liu, F., SternR., H., Huang, M., AceroA, X.: Efficient Cepstral Normalization for Robust Speech Recognition. Department of Electrical and Computer Engineering, Carnegie Mellon University (1992)
5. Kraljevski, I., Mihajlov, D., Gjorgjevik, D.: Hybrid HMM/ANN speech recognition system in Macedonian. Faculty of Electrical Engineering, St. Cirilus and Methodius, Skopje. Veterinary Institute, Skopje (2000); Краљевски, И., Михајлов, Д., Ѓорѓевиќ Д.: Хибриден HMM/ANN систем за препознавање на говор на македонски јазик. Електротехнички факултет, Универзитет Св. Кирил и Методиј, Скопје. Ветеринарен институт, Скопје (2000)

6. Gerazov, B., Ivanovski, Z., Labroska, V.: Modeling of the intonation structure of the Macedonian language on intonation phrases level. Faculty of Electrical engineering and Information technology, St. Cyrilus and Methodius University, Skopje. Intstitute of Macedonian Language Krste Misirkov, Skopje (2012)
7. Геразов, Б., Ивановски, З., Лаброска, В.: Моделирање на интонацискатаструктура на македонскиотјазик на ниво на интонациски фрази. Институт за електроника, Факултет за електротехника и информацискитехнологии, Универзитет Св. Кирил и Методиј, Скопје. Институт за македонски јазик Крсте Мисирков, Скопје(2012)
8. Kumar, N., Andreou, A.G.: Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. Speech Communication 26, 283–297 (1998)
9. Compernolle, V.D.: Noise Adaptation in a Hidden Markov Model Speech Recognition System. Computer Speech and Language (1989)
10. Hermansky, H., Sharma, S.: Temporal patterns (traps) in asr of noisy speech. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP), Phoenix, Arizona, USA (1999)
11. Jyh-Shing, R.J.: Audio Signal Processing and Recognition. CS Dept., Tsing Hua University, Taiwan (1996), http://neural.cs.nthu.edu.tw/jang/books/audiosignalprocessing/index.asp (accessed 2013)
12. Moreno, P.J.: Speech Recognition in Noisy Environments. Department of Electrical and Computer Engineering. Carnegie Mellon University (1996)
13. Sphinx – 4. Speech Recognizer written in Java™, http://cmusphinx.sourceforge.net/sphinx4/ (accessed 2013)
14. CMUSphinx Wiki. Document for the CMU Sphinx speech recognition engines, http://cmusphinx.sourceforge.net/wiki/ (accessed 2013)

The Influence the Training Set Size Has on the Performance of a Digit Speech Recognition System in Macedonian

Daniel Spasovski[1], Goran Peshanski[1], and Gjorgji Madjarov[2]

[1] Netcetera, Skopje, Macedonia
spasovski.daniel@gmail.com, pesanski_goran@yahoo.com,
[2] Faculty of Computer Science and Engineering, Skopje, Macedonia
gjorgji.madjarov@finki.ukim.mk

Abstract. Automatic speech recognition (ASR) systems became an important part of our lives and are used by millions of people. However, scientists still try to improve their accuracy using many different techniques. In this paper, we focus on the influence the training set size has on the performance of Hidden Markov Model (HMM) based digit recognition system in Macedonian. The experiments are conducted using dataset consisting of 3093 samples divided in several different-sized training sets and one test set. Additionally, the behavior of several classification techniques was evaluated for the same issue. The best result was 19.9% error rate for 1500 samples in the training set using HMM based ASR system. This indicates that for this particular problem using the specified dataset the ideal number of samples for the training set is around 1500.

Keywords: automatic speech recognition, training set size, Hidden Markov Model, Word Error Rate.

1 Introduction

The people's desire to communicate with machines in the same way they communicate with other people led to the appearance of the automatic speech recognition (ASR). During the last decade or so, the ASR evolved rapidly to a level where it became a significant part of our everyday life [1]. ASR systems are now integrated in mobile devices, cars, planes, kitchen appliances and are used by millions of people around the world.

Maybe the most important part of the ASR system is the audio datasets used to create the system. The more relevant and appropriate datasets you have for a specific speech recognition problem the better the performance of the ASR system will be for that problem. The quality of the datasets depends on many different parameters and conditions like the environment in which the samples are recorded, the quality of the recording device, the level of noise present in the recorded samples, the number of

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_21

speakers, the total number of samples and many others. Thus, it was very interesting for us to make a research about how the performance of an ASR system depend on the number of samples in the datasets, with focus on the training set, for a specific speech recognition problem.

The main goal in this research is to examine how the performance of digit recognition system in Macedonian depend on the number of samples in the training set and to find an interval for the number of samples for which the performance of the digit recognition system would be the best. It is supposed to serve as a reference for someone who wants to quickly design ASR for isolated words in Macedonian.

The audio database created for the experiment consists of 3093 audio recordings of isolated speech in Macedonian. It includes the digits from 0 to 9 with 2000 recordings for the training sets and 1093 recordings for the test set. A HMM based digit recognition system in Macedonian is created. The system is designed to be speaker independent i.e. to be able to recognize digits in Macedonian regardless of the speaker. Several acoustic models are created, trained with training sets containing different number of samples. The HMM based ASR system is evaluated as experiments are conducted with relevant test set and the Word Error Rate (WER) is measured for each of the acoustic models.

Additionally, several classification techniques like Support Vector Machine (SVM), k Nearest Neighbors (k-NN) and Multilayer Perceptron (MLP) are experimentally evaluated for the same issue. The same datasets used to train the acoustic models and test their performances are used for the classification techniques as well.

Section 2 discuss about research projects that deal with similar kind of problems. Section 3 describes the datasets used in the experiments. The experimental setup and the results are presented in Section 4. Finally the conclusions and the plans for further work are given in Section 5.

2 Related Work

The most common and accurate technique used for ASR is the Hidden Markov Model (HMM). The HMM based ASR systems proved to be more efficient than other techniques used for ASR such as Dynamic Time Warping (DTW) and Artificial Neural Networks (ANN) [2, 3].

In recent years the majority of the research projects are focused on improving the efficiency of the techniques used for ASR by adjusting and optimizing or combining the algorithms that these techniques use. Also, many scientists are focused on providing better combination of features that will improve the performance of the ASR system.

Regarding digit recognition, there are digit recognition systems in many different languages and they are mainly used in telecommunication services [4].

As for the training set optimization, Nagorski and Boves [5] proposed a method for optimizing the training of ASR systems based on Principal Component Analysis (PCA).

However, not much has been done in the area of ASR in Macedonian. One noticeable example is the speaker dependent digit recognition system that recognizes isolated words, designed by Krajlevski et al. [6]. They proposed hybrid HMM/ANN architecture and achieved accuracy of around 85%.

3 Datasets and Experimental Setup

3.1 Datasets

For the needs of the experiments a dataset was created, that is composed of 3093 audio samples of the digits in Macedonian from around 60 different male and female speakers. The samples are collected with an online web application made especially for this purpose. 2000 of the samples make the training set and the other 1093 make the test set. This training set is used to create several smaller training sets with different number of samples per digit. The number of samples for each training set and the number of female and male samples per digit are shown in Table 1.

Table 1. Training sets. Total number of samples in the training set and number of female and male samples per digit.

Total # of samples	F (samples per digit)	M (samples per digit)
600	30	30
800	40	40
1000	50	50
1500	75	75
2000	100	100

On one hand, the fact that the samples are collected with a web application, recorded with many different microphones in different conditions implicates that there is a lot of noise in the samples and it was later reflected in the performances of the ASR. On the other hand, it indicates that the dataset is a relevant represent of the environment where such digit recognition system would be used.

Apart from the dataset mentioned above, another dataset is created. Unlike the first dataset, this dataset is created in controlled conditions using only one microphone. The dataset contains 1046 audio samples from 11 male and 9 female speakers. 600 samples are used for training and the other 446 for testing.

All audio samples that compose the datasets are of the following quality:

- 8 KHz sample rate;
- 8 bit audio depth;
- Mono-channel audio signal in .wav format.

Each training dataset is used for training the acoustic model, while the testing datasets are used for evaluating the predictive performances of the acoustic model.

3.2 Experimental Setup

The next part of the process was to create the statistical language model for the recognizer. The language model consists of the words for the digits in Macedonian, phonetic dictionary and fillers, which represent "empty" speech or silence.

Macedonian is a phonetic language, so there is a phoneme for each letter. Therefore, the phonetic dictionary contains 16 phonemes for each different letter in the words.

For the creation of the HMM based digit recognition system in Macedonian we used the Sphinx-4 framework [7]. It consists of part that provides feature extraction, modeling and training an acoustic model, modeling a language model and a part for decoding that enables you to test the predictive performance of the system.

The feature extraction part separates the speech signal into overlapping frames and produces feature vector sequences each containing 39 MFCC features (coefficients). The 39-dimensional MFCC vector is created of the first 13 MFCC coefficients, 13 features that represent the speed of the signal (the first derivate of the first 13 MFCC coefficients) and 13 features that represent the acceleration of the signal (the second derivate of the first 13 MFCC coefficients). This 39-dimensional feature vector, which is the most widely used in speech recognition, is used as the basic feature vector for our digit recognition system in Macedonian.

We used the Baum-Welch algorithm [8] (integrated in Sphinx-4) for training the acoustic models. This algorithm finds i.e. adjusts the unknown parameters of a HMM.

The performance of all acoustic models is evaluated by measuring the Word Error Rate - WER (equation 1).

$$WER = \frac{\#\ correctly\ recognized\ words}{\#\ number\ of\ samples\ in\ the\ test\ set}\ [\%] \tag{1}$$

We consider this evaluation relevant because of the versatile nature of the test set which, in some way, tells us how the acoustic models would behave in natural environment.

For many different issues the classification techniques prove themselves as the right approach to finding a solution with high precision. Because of the fact that digit recognition in Macedonian is a problem with finite number of possible outputs it was interesting to examine the behavior of the classification techniques like SVM, k-NN and MLP when applied to this issue.

The evaluation of the classification techniques was performed using their implementations in Weka [9]. For training the SVMs, we used the SMO implementation. In particular, we used SVMs with a radial basis kernel. The kernel parameter gamma and the penalty C, for each combination of dataset and method, were determined by 10-fold cross validation using only the training set. The values 2^{-15}, 2^{-13}, … , 2^{1}, 2^{3} were considered for gamma and 2^{-5}, 2^{-3}, … , 2^{13}, 2^{15} for the penalty C. The number of neighbors in the k-NN method for each dataset was determined from the values 1 to 9 with step 2 [10]. The Neural Networks are represented by MLP with 25 neurons in the hidden layer and value for the validation threshold of 10. After determining the best parameters values for each method on every dataset by 10-fold cross validation, the classifiers were trained using all available training examples and were evaluated by recognizing all test

examples from the corresponding dataset. In this part of our research the main accent in set on the classification techniques that are used for speech recognition i.e. digit classification, and not on the feature extraction process. For the feature extraction we used the tool MARSYAS, a framework for audio processing and speech analysis with special emphasis on music information retrieval. The feature vector consists of the first 13 MFCC coefficients and additional 39 features that are calculated as a combination of the mean value and the standard deviation of the first 13 features.

The evaluation process for the classifiers is the same as the one for the HMM based digit recognition system. The same test set and the same measurement (WER) are used to examine and compare the performances of the classifiers.

4 Results

4.1 Results from the Experiments with the HMM Based Model

In Table 2, the results of the evaluation of the acoustic models are shown. We can notice that the model trained with the training set that consists of 1500 samples (75 male and 75 female samples per digit) shows the best results. In other words its WER is the smallest, only 19.9%.

Table 2. WER of the acoustic models depending of the size of the training set

Training sets (# of samples)	WER (%)
600	28,2
800	27,5
1000	26,7
1500	**19,9**
2000	24,4

This can better be seen from the graph in Figure 1 where we can clearly see that for our dataset of audio samples, the digit recognition system in Macedonian, shows best performance when its acoustic model is trained with a training set that has size of around 1500 samples (75 male and 75 female per digit).

The interesting thing that should be noticed here is that there is a dip for the value of the WER when the number of samples in the training set is around 1500. It means that the WER value reaches its local optimum (optimum for this dataset) at that point.

The additional 500 samples only increased the noise level in the dataset and decreased the performances of the ASR. However, maybe this value would vary if the size of the whole dataset increases and the quality of the recordings is better.

The results provided from the experiments, conducted with the dataset collected under controlled conditions and with only one microphone, show that the conditions in which the acoustic model is trained and tested are also very important for the performance of an ASR system. The WER here is only 13.3%.

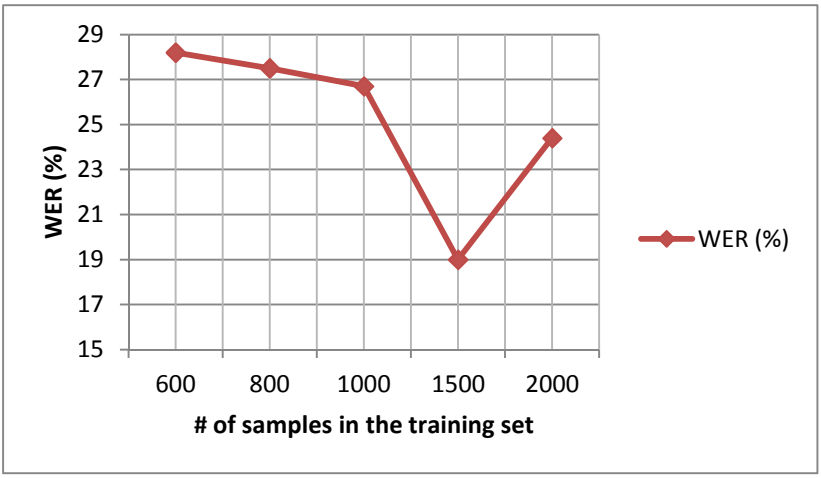

Fig. 1. WER in HMM based acoustic models depending on the size of the training set

4.2 Results from the Experiments with the Classification Techniques

The classifiers are trained with the same different-sized 5 training sets and are evaluated in the same way, by measuring the WER when tested against the same test set. It was interesting to examine the performance of techniques that are rarely used for speech recognition. In the table in Table 3, the results obtained from the evaluation process of the classification techniques are shown. It can be noticed that the best results i.e. best accuracy comes from the SVMs, 24.89% WER for 2000 samples in the training set, while the results obtained using MLP are the worst.

Table 3. WER of the classifiers depending on the size of the training set

Training set (# of samples)	**SVM**	*k-NN*	MLP
600	57,37	67,16	61,4
800	51,33	65,15	55,72
1000	51,88	65,05	56,82
1500	39,16	46,84	48,95
2000	**24,89**	29,1	45,48

However, the behavior of all three classifiers is similar and their performances depend on the size of the training set in the same way. Their WER drops down with the increasing of the number of samples in the training set. This behavior can be noticed easily from the graph in Figure 2.

We can see that the WER is the smallest for the training set with 2000 samples (100 male and 100 female per digit) and maybe it will decrease if the size of the

training set increases. But, because of the worse performance and the time needed for the training process (significantly more than the HMM), the classifiers are not as good as the HMMs for solving this kind of problem. Also, they are not suitable for real time usage.

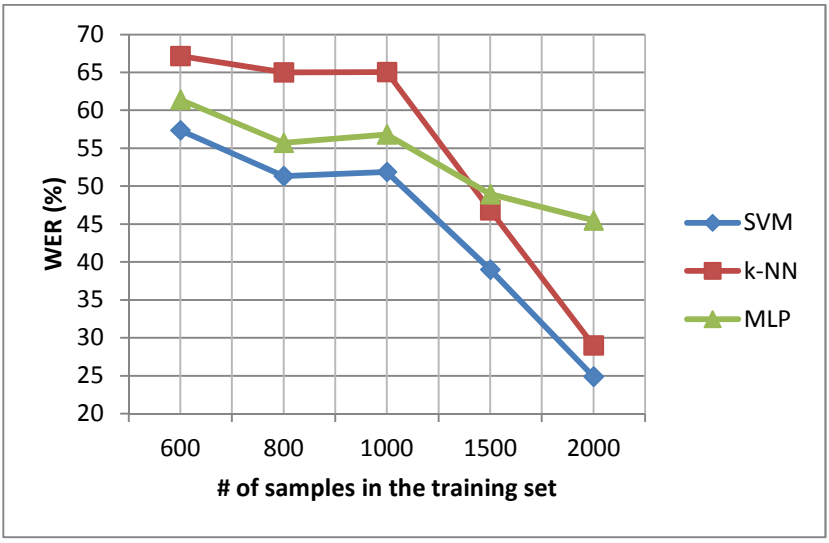

Fig. 2. WER of the classifiers depending on the size of the training set

5 Conclusion and Further Work

In this paper, we discussed about the importance of the size of the training set and how it influences the performance of a digit recognition system in Macedonian. Experiments were conducted with several different-sized training datasets and corresponding test dataset.

First we conducted the experiments with HMM base ASR system and then the process was repeated for the classification techniques (SVMs, k-NN, MLP). As expected, the HMM based ASR system provided better predictive performance in comparison to the other classification techniques, with a WER of 19.1% for 1500 samples in the training dataset. This means that the local optimum for the HMM based ASR system evaluated using the particular datasets could be reached the point of 1500 samples for the training set. This means that for number of samples around 1500 the HMM based digit recognition system provides the best accuracy.

On the other hand, the classification techniques behave in the same way depending on the size of the training dataset. With the increasing of the number of samples in the training set, the WER drops down. The SVM showed the best performance, but none of these techniques can be used in real time applications for digit recognition because of the time they need for the training process.

These results can serve as a future reference for problems with similar complexity as the digit recognition in Macedonian. The future work will involve further experiments with larger datasets on more complex problems. We will also try to conduct this kind of experiments for continuous speech recognition system.

References

1. Juang, B.H., Rabiner, L.R.: Automatic Speech Recognition - A Brief History of the Technology Development. Rutgers University and the University of California, Santa Barbara (2004)
2. Plannerer, B.: An Introduction to Speech Recognition. ver. 1.1, Munich, Germany (2005)
3. Lippmann, R.P.: Neural Networks Classifiers for Speech Recognition. The Lincoln Laboratory Journal 1(1) (1988)
4. Rabiner, L.R.: Applications of Speech Recognition in the Area of Telecommunications. AT&T Labs (1997)
5. Nagórski, A., Boves, L., Steeneken, H.: Optimal Selection of Speech Data For Automatic Speech Recognition Systems. Department of Language and Speech. University of Nijmegen, The Netherlands (2002)
6. Krajlevski, I., Mihajlov, D., Djordjevikj, D.: Hybrid Hmm/Ann System for Speech Recognition of Macedonian Language. In: Fifth National Conference With International Participation ETAI, Ohrid (2000)
7. Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., Woelfel, J.: Sphinx-4: A Flexible Open Source Framework for Speech Recognition. Sun Microsystems Inc. (2004)
8. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77(2) (1989)
9. University of Waikato, New Zealand: Machine learnings of software written in Java (Version 3.6) "Weka" (1997)
10. Madjarov, G.M.: Advanced methods for building hierarchical multi-label classifiers. Phd thesis, Faculty of Computer Science and Engineering, Skopje, Macedonia, pp. 48–50 (2012)

Combined AES + AEGIS Architectures for High Performance and Lightweight Security Applications

Furkan Şahin[1], H. Fatih Uğurdağ[1], and Tolga Yalçın[2]

[1] Department of Electrical and Electronics Engineering, Ozyegin University, Istanbul, Turkey
furkan.sahin@ozu.edu.tr, fatih.ugurdag@ozyegin.edu.tr
[2] University of Information Science and Technology "St. Paul the Apostle", Ohrid, Macedonia
tolga.yalcin@uist.edu.mk

Abstract. AES has been the prominent block cipher since its introduction as the standard. It has been *the* cipher used in almost all new applications that require solid, unbreakable security with reasonable resource usage. Several versions of AES have been implemented in both hardware and software platforms with all kinds of design targets varying from high-performance to lightweight. With the widespread Internet, authenticated encryption (AE) has gained an unprecedented popularity, making AES the logical choice for AE implementations. While there already exists standardized modes that allow AES to be used for AE, more recently, special AE schemes that utilize AES in its native form (or with minimal modifications) have emerged. While these modes claim better performance and resource usage, very few implementations exist to support these claims, yet. In our work, we combine AES with one of the most recent AE ciphers, namely AEGIS, in an effort to analyse the combined performance of the two ciphers.

Keywords: Encryption, authenticated encryption, AES, AEGIS, high performance, lightweight, security, FPGA, ASIC.

1 Introduction

Security is more important than ever. Every modern computation and communication device relies on security. It may be a cloud server storing millions of users' critical data. It might be the Internet backbone of a multi-billion Euro enterprise. It can be our cell phones. Or it can be the pace maker that keeps us alive. All of these devices, and many more, work on the principle and assumption that they are secure against any kind of *attacks*.

This may in fact be true. Modern cryptography has introduced so many rock-solid algorithms, most of which seem unbreakable for many more decades to come. The most well-known and well-analyzed one is of course the Advanced Encryption Standard (AES) [1] which was selected amongst several competitor

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_22

algorithms at the end of a long competition process in 2001. Since then, countless efforts have been in the hopes of breaking AES, or at least finding some weakness. None has been successful so far.

While AES is mainly designed to provide encryption (and decryption), it can also be used in conjunction with one of the block cipher modes of operation to provide authentication as well. For example, the most popular combined mode of operation CCM, uses AES in both counter and CBC mode, for encryption and authentication, respectively, thereby practically turning it into a authenticated encryption (AE) cipher [2]. This is quite important, as AE ciphers form the backbones of Internet Protocol Security Suite [3].

Another alternative to implementing an AE cipher is to use a block cipher together with a hash function. However, neither of the alternatives offer resource efficient solutions. The former solution requires running AES twice for each data block, once for encryption and once for authentication, halving the throughput; while the latter solution requires a second module – the hash function, doubling the resource usage.

More recent attempts have introduced specialized AE ciphers. They come in too many forms and configurations. Some of them use hash functions in AE mode, while some of them propose completely new structures tailored for AE. Of particular interest to us is the third class of AE ciphers, which modifies the native AES somehow in order to come up with an AE scheme that relies on the proven (yet unbroken) security of AES. ALE and AEGIS are two such examples [4][5].

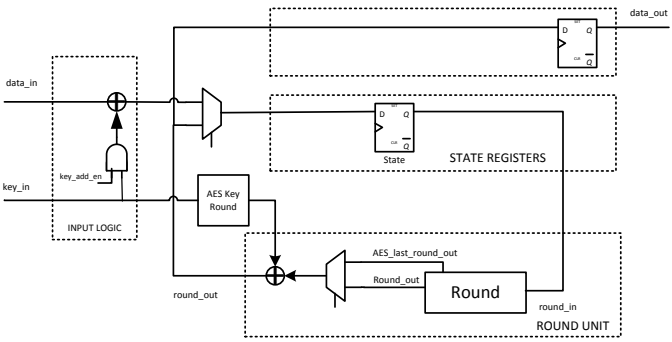

Fig. 1. Commonly Implemented Iterative AES Architecture

One other claim from the creators of these AES-based AE ciphers is that they can easily be realized from an existing AES implementation with minimal effort and resulting in favorable resource savings. However, due to the infancy of such ciphers, these claims are yet to be proven.

In our work, we choose AEGIS, the most recent AES-based AE proposal (at the time we started this work, so to say), and combine it with naive AES cores in order to investigate the validity of such claims. We further aim to propose

combined architectures for both high-performance and lightweight applications that will allow the users to switch between block cipher and AE modes by means of a simple switch.

In the following section, we present a brief overview of both AES and AEGIS. We then explain in detail our combined AES+AEGIS architecture for high-performance applications. It is followed by the combined architecture for lightweight applications. We conclude with performance figures and future directions.

2 AES and AEGIS Overview

AES is a symmetric cipher that supports 128, 192, 256 bits of key sizes and fixed 128 bits data blocks [1]. AES consists of round iterations, and the number of round iterations is 10, 12, 14 for the key sizes of 128, 192, 256, respectively. Each intermediate cipher result is called *state*. A roundkey for each round iteration is also generated from the encryption key. Each round consists of SubBytes, ShiftRows, and MixColumns operations performed on state, finally adding the state and roundkey, which is called AddRoundKey operation. However, the last round skips MixColumns.

AEGIS is a dedicated authenticated encryption algorithm, which is constructed from the AES encryption round function [5]. AEGIS-128, AEGIS-256, and AEGIS-128L use 5, 6, 8 AES round functions, respectively. According to [5], these AEGIS algorithms offer high levels of security. Intermediate cipher results are called state also in AEGIS, however, in contrast to AES, a state consists of 5, 6, 8 16-byte data blocks in AEGIS-128, AEGIS-256, and AEGIS-128L, respectively. A function called State_Update performs 5, 6, 8 AES rounds on state in AEGIS-128, AEGIS-256, and AEGIS-128L, respectively. Each AEGIS algorithm consists of initialization, processing the authenticated data, encryption, and finalization steps. Depending on the AEGIS algorithm and the length of data to be processed, each step also consists of different numbers of State_Update iterations. For a detailed explanation of AEGIS, we refer the reader to [5].

3 AES-128 and AEGIS-128 Encryption Core Architecture

We have designed and implemented a core that can do both AES and AEGIS encryption depending on a configuration select input. A state in AEGIS-128 consists of five 16-byte data blocks. AEGIS-128 State_Update128 function consists of five AES round functions without addition with the roundkey.

Well-known iterative architecture of AES encryption core is shown in Figure 1. The architecture consists of Input Logic, Round Unit, Key Round Unit, State and Output Registers. Input Logic performs AddRoundKey operation in the load phase. Round Unit performs AES round operation which consists of SubBytes, ShiftRows, MixColumns and AddRoundKey operations. Key Round Unit performs key expansion operation and produces round key for each round. State Registers store 16-byte state of AES. After 10 round operations, cipher text

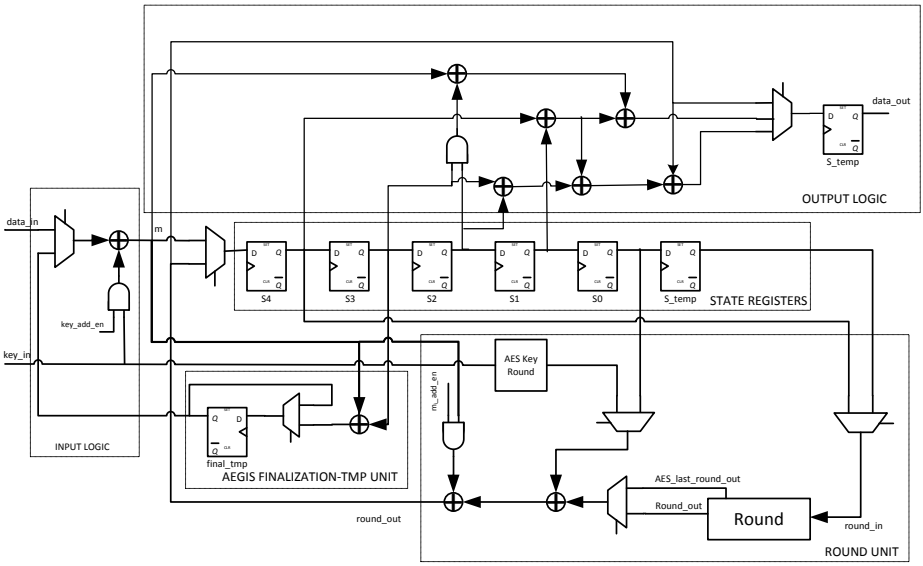

Fig. 2. AES-128 and AEGIS-128 Encryption Core Architecture

outputted via data_out register. This AES architecture is our starting point for designing combined AES + AEGIS architecture for high performance applications. We examined the AEGIS and tried to design a modified architecture that can perform both AES and AEGIS. To make possible this we made some additions and modifications on standalone AES encryption core.

The architecture of our AES-128/AEGIS-128 encryption core is depicted in Figure 2. This core is for 128-bit keys. All datapaths and registers in our design are 128-bit. The core consists of five components: input logic, state registers, round unit, AEGIS finalization-tmp unit, output logic. The core operates in five phases. Cycle counts for each phase is given in Table 1.

Table 1. Cycle counts for AES-AEGIS Operation Phases

Phases	AEGIS	AES
0: Load	5 cycles	1 cycle
1: Initialization	10×5 cycles	–
2: AD processing	$u \times 5$ cycles	–
3: Encryption	$v \times 5$ cycles	10 cycles
4: Finalization	7×5 cycles	–

3.1 Input Logic

The input logic unit performs input data selection for state registers, key addition during load phases of AES and AEGIS, and initialization of AEGIS.

– **Data Input:** The data input is used for plain text input in AES and initialization vector, const0, const1, associated data, plain text input in AEGIS. Table 2 shows data input depending on phase, state update, cycle, AEGIS, AES.

Table 2. Data Input for AES-AEGIS Operation Phases

phase	state_update	cycle	data input (AEGIS)	data input (AES)
0	- (load)	0	init. vector (IV)	plain text
		1	const1	–
		2	const0	–
		3	const0	–
		4	const1	–
1	2, 4, 6, 8, 10	0	–	–
	1, 3, 5, 7, 9	0	init. vector (IV)	–
		1-4	–	–
2	all	0	assoc. data (AD)	–
	all	1-4	–	–
3	all	0	plain text	–
	except last	1-4	–	–
	last	4	(adlen || msglen)	–
4	all	all	–	–

– **Key Input:** 128-bit key input for both AES and AEGIS. In AES, this initial key is also input for AES Round Key unit.
– **Msg Output:** Table 3 shows msg output of the input logic unit depending on phase, state update, cycle, AEGIS, AES.

3.2 State Registers

This unit consists of six 16-byte shift registers. First five registers, which are S4, S3, S2, S1 and S0, store a state of AEGIS-128. S_temp is an additional 16-byte register for storing the previous S0. It is a necessity coming from the AEGIS-128 State_Update128 function. As stated in [5], AEGIS-128 State_Update128 function rounds firstly $S_{i,4}$, then it rounds $S_{i,0}$. So we designed our AES-AEGIS encryption core as follows: In the first cycle of each State_Update128 operation, S4, which stores $S_{i,4}$, is fed into the Round Unit. Since $S_{i,0}$ must be rounded in the second cycle of State_Update128, it is shifted from S0 to S_temp at first

Table 3. Msg Input of Input Logic for AES-AEGIS Operation Phases

phase	state_update	cycle	msg out (AEGIS)	msg out (AES)
0	- (load)	0	key $\oplus$ data$_{in}$	key $\oplus$ data$_{in}$
		1	data$_{in}$	–
		2	data$_{in}$	–
		3	key $\oplus$ data$_{in}$	–
		4	key $\oplus$ data$_{in}$	–
1	2, 4, 6, 8, 10	0	key_in	–
	1, 3, 5, 7, 9	0	key $\oplus$ data$_{in}$	–
		1-4	–	–
2	all	0	data$_{in}$	–
	all	1-4	–	–
3	all	0	data$_{in}$	–
	except last	1-4	–	–
	last	4	data$_{in}$	–
4	all	0	finalization	–
	all	1-4	–	–

cycle, and stored in S_temp. Except the first cycle, S_temp always contains the proper one fifth part of a state, which must be fed into Round Unit. Contents of registers, depending on the cycle, are given in Table 4. Register contents are shifted each cycle, and parts of a state is propagated through the S_temp register.

Table 4. State Register Contents for AES-AEGIS Operation Cycles

cycle	S4	S3	S2	S1	S0	S_temp
0	$S_{i,4}$	$S_{i,3}$	$S_{i,2}$	$S_{i,1}$	$S_{i,0}$	$S_{i-1,4}$
1	$S_{i+1,0}$	$S_{i,4}$	$S_{i,3}$	$S_{i,2}$	$S_{i,1}$	$S_{i,0}$
2	$S_{i+1,1}$	$S_{i+1,0}$	$S_{i,4}$	$S_{i,3}$	$S_{i,2}$	$S_{i,1}$
3	$S_{i+1,2}$	$S_{i+1,1}$	$S_{i+1,0}$	$S_{i,4}$	$S_{i,3}$	$S_{i,2}$
4	$S_{i+1,3}$	$S_{i+1,2}$	$S_{i+1,1}$	$S_{i+1,0}$	$S_{i,4}$	$S_{i,3}$

3.3 Round Unit

This unit performs the AES round function. An AES round consists of SubBytes, ShiftRows, MixColumns, and AddRoundKey transformations. The Round box in this unit performs SubBytes, ShiftRows, and MixColumns transformations. AddRoundKey is separately done as shown in Figure 2. In the first cycle of

a State-Update128 operation, round function is applied on the contents of S4 register, whereas it is applied on the contents of S_temp register contents in all cycles.

3.4 AEGIS Finalization-Tmp Unit

This unit computes and stores the 16-byte tmp value in the finalization step of the AEGIS. *tmp* is defined as $S_{u+v,3} \oplus$ (adlen || msglen) in [5]. The $S_{u+v,3}$ is stored in S3 register and the (adlen || msglen) is fed to the unit via data input, whereas the || symbol represents concatenation.

3.5 Output Logic

Output Logic unit performs computation of output values cipher text and tag. The proper output is selected by the multiplexer with respect to the selected algorithm (AES or AEGIS) and output type (ciphertext or tag). In the case of AEGIS, ciphertext is output at the first cycle of each State_Update in encryption phase (phase 3). For AES, there is no State_Update function. However, since ten round iterations are performed for AES in two State_Update of AEGIS, the last round of AES corresponds to last round of second state update of phase 3 and ciphertext is output after that round. Data output values are given in Table 5.

Table 5. Data Output for AES-AEGIS Operation Phases

phase	state update	cycle	data output (AEGIS)	data output (AES)
0	all	all	–	–
1	all	all	–	–
2	all	all	–	–
3	all	0	$\text{data}_{in} \oplus$ S1 $\oplus$ S4 $\oplus$ (S2 & S3)	–
	last	4	–	round_{out}
4	7	4	$\text{round}_{out} \oplus$ S4 $\oplus$ S3 $\oplus$ S2 $\oplus$ S1	–

4 Lightweight AES and AEGIS Encryption Core Architecture

Several 8-bit FPGA and ASIC implementations of AES are reported in literature [6][7][8][9]. Based on our knowledge, the lowest power and lowest area implementation of AES encryption hardware core have been reported in [10], where 8-bit datapaths are employed and one AES round performed in 16 clock cycles. Their design supports 128-bit keys.

We designed and implemented a lightweight hardware core that performs both AES-128 and AEGIS-128 encryption based on a 1-bit selection input. Basically, we modified some parts of the design in [10] and added some new parts to make AEGIS-128 encryption possible. Our new lightweight design is a compact combination of the design in [10] and our AES/AEGIS core design approach, which was presented in the previous section. Our lightweight core also employs 8-bit datapaths, and one AES round is performed in 16 clock cycles. High-level architecture of our design is depicted in Figure 3.

4.1 Top Level I/O ports

Top level I/O ports of our design are data input, key input and data out. Initialization vector, (adlen || msglen), associated data in AEGIS, plain text in both AES and AEGIS fed to the core as serial bytes via data input port. Encryption key is fed into the core via key input port. Authentication tag in AEGIS, ciphertext in both AES and AEGIS output via data out port. One AES round is performed in 16 clock cycles. Input schedule for data input port is given in Table 6.

Table 6. Input Schedule for Data Input

phase	state update cycle		data input (AEGIS)	data input (AES)
0	- (load)	0	init. vector (IV)	plain text
		1-4	–	–
1	10	0	(adlen || msglen)	–
	2, 4, 6, 8	all	–	–
	1, 3, 5, 7, 9	0	init. vector (IV)	–
	all	1-4	–	–
2	all	0	assoc. data (AD)	–
	all	1-4	–	–
3	all	0	plain text	–
	except last	1-4	–	–
4	all	all	–	–

4.2 S-Box

This unit performs SubBytes transformation. There are two identical S-box in our design same as in [10]. Sbox-1 used in state rounding and Sbox-2 is used in AES key rounding.

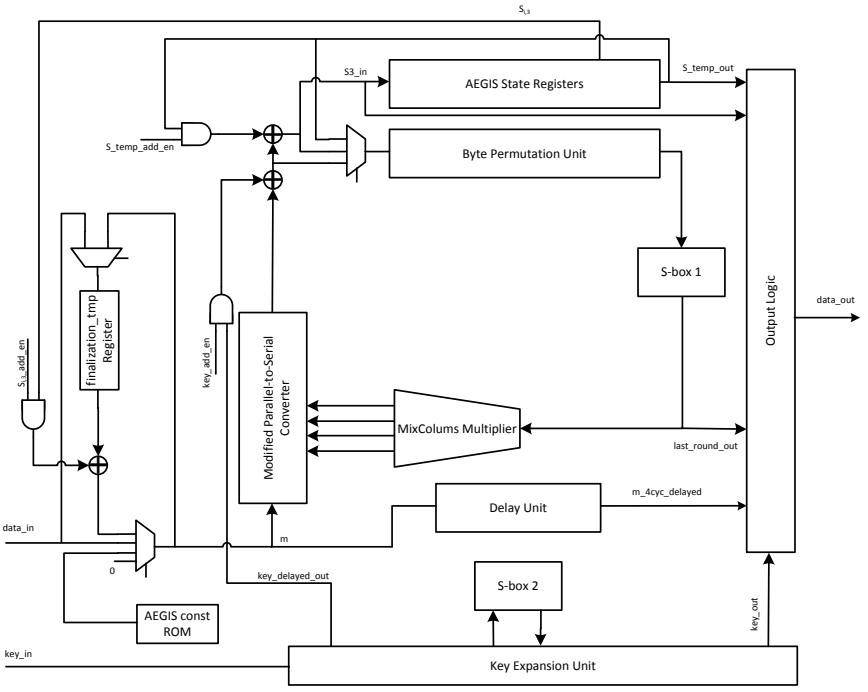

Fig. 3. Lightweight AES-128 and AEGIS-128 Encryption Core Architecture

4.3 AEGIS Constant ROM

AEGIS const ROM stores 32 bytes const value, which is the Fibonacci sequence module 256. First 16 bytes of it is called *const0* and the last 16 bytes is called *const1*. During the initialization phase, it is fed into the encryption core via multiplexer in input.

4.4 Key Expansion Unit

Key expansion unit in our design is almost the same as the design presented in [10]. Since there is no key rounding in AEGIS, our key expansion unit works like a ring shift register when AEGIS is employed. After the encryption key is loaded once into the unit, if the core is working in AES mode, the unit performs key rounding, otherwise it just shifts the register contents. Since it is the last register output fed into first register input in AEGIS mode, encryption key is shifted in a ring. rk_delayed_out and rk_last_out outputs are the same as the corresponding outputs in [10].

This unit performs SubBytes transformation. There are two identical S-boxes in our design, the same as in [10]. Sbox-1 used in state rounding, and Sbox-2 is used in AES key rounding.

4.5 Modified Parallel-to-Serial Converter

This unit is basically very similar to the parallel-to-serial converter presented in [10]. However, since our core supports also AEGIS-128 authenticated encryption, we made some modifications. In AEGIS, associated data, plain text, and finalization tmp is added to $S_{i,0}$ in proper state_update operation. Since the rounded part of an AEGIS state is transferred from MixColumns multiplier to parallel-to-serial converter, adding m to $S_{i,0}$ is performed by this unit. Modified circuit diagram of the unit is depicted in Figure 4.

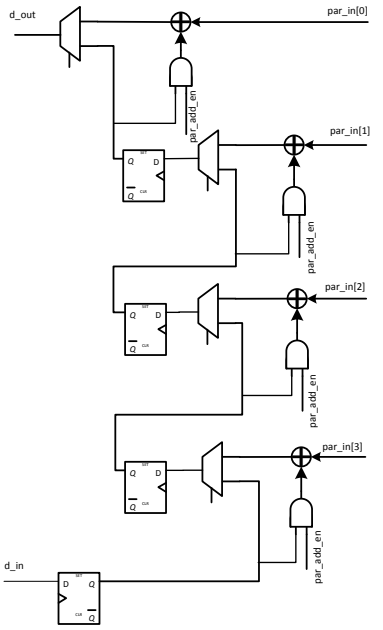

Fig. 4. Modified Parallel-to-Serial Converter

4.6 Byte Permutation Unit

This unit is the same as the byte permutation unit in [10]. 12-bytes of a state is stored and ShiftRows operation is performed by this unit.

4.7 MixColumns Multiplier

This unit performs MixColums multiplication. This is also the same as in [10]. Since bytes of a column comes serially, one column multiplication takes 4 clock cycles.

4.8 AEGIS State Registers

This unit stores AEGIS-128 state parts. S3, S2, S1 and S0 consists of 16 8-bit registers. At the beginning of each state update, S3, S2, S1, and S0 store $S_{i,3}$, $S_{i,2}$, $S_{i,1}$ and $S_{i,0}$, respectively. Each clock cycle, register contents are shifted and bytes of AEGIS state is propagated through S_temp_out. $S_{i,4}$ is stored partially in Byte Permutation Unit and is partially Modified Parallel-to-Serial Converter. These registers are used for the same logical reason, which was mentioned in our parallel AES-AEGIS hardware encryption core.

4.9 Finalization Tmp Register

This unit stores 16-byte AEGIS finalization tmp, which is used during state updates in finalization phase. *tmp* computed at the beginning of finalization phase, then it is stored and added to first round output of each state_update. Before the finalization phase, this unit stores (adlen || msglen). As shown in Tableřefinpsch, (adlen || msglen) is fed to the core at the first round of phase 1. By using (adlen || msglen), control module of core computes u and v, which are number of State_Update128 iterations – AD Processing phase (phase 2) and encryption phase (phase 3). In the first round of AEGIS finalization phase, to compute the finalization tmp $S_{i,3}$ is exored with (adlen || msglen) and the result stored again in finalization tmp register unit.

4.10 Delay Unit

This unit has four 8-bits back to back registers. It just delays its input 4 clock cycles. Since the latency of MixColums Multiplier is 4 clock cycles, this delay unit aligns the input coming from data input and MixColumns Multiplier output for calculations in Output Logic.

4.11 Output Logic

Output logic unit is designed to perform the AEGIS tag calculation in the finalization phase (phase 4) and cipher text calculation in the encryption phase (phase 3) for both encryption algorithms. Data output schedule in given in Table 7.

Table 7. Output Schedule for Data Output

phase	state update	round	data output (AEGIS)	data output (AES)
0	all	all	–	–
1	all	all	–	–
2	all	all	–	–
3	all	0	m_4cyc_d $\oplus$ S1 $\oplus$ $S3_{in}$ $\oplus$ (S2 & S3)	–
	last	4	–	last_rnd_o $\oplus$ rk_last_o
4	7	4	$S3_{in}$ $\oplus$ S3 $\oplus$ S2 $\oplus$ S1 $\oplus$ S0	–

5 Conclusion and Future Work

We implemented both of our architectures using UMC 90 nm low-leakage standard cell library and Cadence RTL Compiler.

The high-performance version occupies a total area of 19.6K GE and can run up to a maximum frequency of 91 MHz. This corresponds to a maximum throughput of 1059 Mbps. The lightweight version occupies a total area of 9K GE. Since it was targeted for lightweight applications, we did not test its highest frequency. Instead, we synthesized it for a fixed target frequency of 100 KHz. At this frequency, it offers a throughput of 72.7 Kbps.

In the high-throughput case, the area of the combined architecture is only 50% higher than that of a standalone AES module, while for the lightweight case, the area is almost tripled. For both cases, there is no loss in terms of throughput. These figures support the claims of the designers of AES-based AE cipher schemes in general.

In further steps of our study, we will implement our architectures using other cell libraries as well as on various FPGA platforms, in order to verify our initial observations. We will also add power consumption figures. Furthermore, we are planning to investigate ways of integrating other AES-based AE schemes in our architecture with minimal additional resource usage.

References

1. Daemennn, J., Rijmen, V.: The Design of Rijndael: AES - The Advanced Encryption Standard. Springer, Heidelberg (2002)
2. Formal Specification of the CCM Mode of Operation (2005)
3. Information Technology - Security techniques - Authenticated Encryption (2009)
4. Bogdanov, A., Mendel, F., Regazzoni, F., Rijmen, V., Tischhauser, E.: Lightweight aes-based authenticated encryption. In: Fast Software Encryption (FSE), Singapore (March 2013)
5. Wu, H., Preneel, B.: Aegis: A fast authenticated encryption algorithm. Cryptology ePrint Archive, Report 2013/695 (2013), `http://eprint.iacr.org/`
6. Good, T., Benaissa, M.: Aes on fpga from the fastest to the smallest. In: Rao, J.R., Sunar, B. (eds.) CHES 2005. LNCS, vol. 3659, pp. 427–440. Springer, Heidelberg (2005)
7. Farhan, S.F., Khan, S.A., Jamal, H.: An 8-bit systolic aes architecture for moderate data rate applications. Microprocess. Microsyst. 33(3), 221–231 (2009)
8. Feldhofer, M., Dominikus, S., Wolkerstorfer, J.: Strong authentication for rfid systems using the aes algorithm. In: Joye, M., Quisquater, J.-J. (eds.) CHES 2004. LNCS, vol. 3156, pp. 357–370. Springer, Heidelberg (2004)
9. Feldhofer, M., Wolkerstorfer, J., Rijmen, V.: AES implementation on a grain of sand. IEE Proceedings / Information Security 152, 13–20 (2005)
10. Hamalainen, P., Alho, T., Hannikainen, M., Hamalainen, T.D.: Design and implementation of low-area and low-power aes encryption hardware core. In: Proceedings of the 9th EUROMICRO Conference on Digital System Design, DSD 2006, pp. 577–583. IEEE Computer Society, Washington, DC (2006)

Novel Methodology for CRC Biomarkers Detection with Leave-One-Out Bayesian Classification

Monika Simjanoska and Ana Madevska Bogdanova

Ss. Cyril and Methodious University, Faculty of Computer Science and Engineering, Rugjer Boshkovikj 16, 1000 Skopje, Macedonia
m.simjanoska@gmail.com, ana.madevska.bogdanova@finki.ukim.mk

Abstract. In our previous research we developed a methodology for extracting significant genes that indicate colorectal cancer (CRC). By using those biomarker genes we proposed an intelligent modelling of their gene expression distributions and used them in the Bayes' theorem in order to achieve highly precise classification of patients in one of the classes carcinogenic, or healthy. The main objective of our new research is to subside the biomarkers set without degrading the sensitivity and specificity of the classifier. We want to eliminate the biomarkers that do not play an important role in the classification process. To achieve this goal, we propose a novel approach for biomarkers detection based on iterative Bayesian classification. The new Leave-one-out method aims to extract the biomarkers essential for the classification process, i.e. if they are left-out, the classification shows remarkably degraded results. Taking into account only the reduced set of biomarkers, we produced an improved version of our Bayesian classifier when classifying new patients. Another advantage of our approach is using the new biomarkers set in the Gene Ontology (GO) analysis in order to get more precise information on the colorectal cancer's biomarkers' biological and molecular functions.

Keywords: Colorectal Cancer, Biomarkers Detection, Bayesian Classification, Gene Ontology.

1 Introduction

Recently, the scientists provide intensive gene expression profiling experiments in order to compare the malignant to the healthy cells in a particular tissue. The advantage of the microarray technologies enables simultaneous observation of thousands of genes and allows the researchers to derive conclusions whether the disorder is a result of the abnormal expression of a subset of genes.

In our previous work we have used gene expression data from Affymetrix Human Genome U133 Plus 2.0 Array to perform analysis of CRC and healthy tissues [1]. During the research we developed a methodology for biomarkers detection based on the two types of tissues, carcinogenic and healthy. The obtained set of biomarkers was then used to build a machine learning (ML) based classifier capable of distinguishing between carcinogenic and healthy patients.

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_23

Since the classification analysis resulted in very high accuracy when classifying both CRC and healthy patients, we proceeded to inspect whether the biomarkers we discovered play important biological role in the CRC development [2]. For that purpose, we provided GO analysis and inspected the molecular functions and the biological processes of a particular set of genes that showed to be over-represented among all biomarkers. Considering the colorectal cancer significance of the biomarker genes, we confirmed few biomarkers to be tightly related to the disease: $CHGA$, $GUCA2B$, $MMP7$, $CDH3$ and PYY.

In this paper we address another issue - reducing the biomarkers set in order to improve the classification reliability and to extract new information from the Gene Ontology analysis. We developed a new methodology for subsiding the biomarkers set without degrading the sensitivity and specificity of the built classifier. We want to eliminate the biomarkers that do not play an important role in the classification process. To achieve this goal, we propose a new approach, based on iterative Bayesian classification. In order to eliminate the non-informative genes, we use a Leave-one-out method. Taking into account only the reduced set of biomarkers (the subsided gene set), we produced an improved version of our Bayesian classifier when classifying new patients.

The GO helps us to find the biological meaning of the gene data and their role in the functions connected to the CRC. The GO analysis [2] has showed the possible biological and molecular functions connected to the CRC. The novel proposed methodology, gives us an advantage in the GO analysis, because we can obtain more precise knowledge about the expressed genes in the CRC disease.

The rest of paper is organized as follows. In Section 2 we present the literature related to CRC and GO analysis. The analysis flow is presented in Section 3, whereas the GO analysis are described in Section 4. In Section 5 we discuss the experiments from both the classification and the GO analysis. In Section 6 we present the conclusion from our work and our plans for further research.

2 Related Work

In this section we give a review of the recent literature that relates to CRC and GO analysis.

Authors in [3] sum up the biomarkers results from 23 different researches on CRC and GO analyses. Even though most of them show diversity in the significant genes revealed, the authors in their research take into account the unique biomarkers, which are nearly 1000, and perform GO analysis by using few different tools. They mainly hold on to the ontology results of the enriched set of genes, rather than verifying the biomarkers with classification methods so that we can compare our results.

Similarly, in [4] the researchers use Affymetrix microarray data from 20 patients to reveal significant gene expression, which resulted in 1469 biomarkers. From the GO results they ranked top 10 most important pathways. Comparing our results to theirs, we realized that there is no overlap between ours and their biomarkers sets. Even though they lack a classification analysis, we may

include their biomarkers in our future work and test the ability of the Bayesian approach to make an appropriate modelling using different biomarkers revealing procedure.

Since the non overlapping between the biomarkers sets discovered in different scientific papers is very common, a new meta-analysis model of CRC gene expression profiling studies is proposed in [5]. As the authors ranked the biomarker genes according to various parameters, the gene CDH3 which we found to play role in the CRC [2] is also found by their meta-analysis model.

Another interesting approach maintained with classification analysis is presented in [6], where the authors constructed disease-specific gene networks and used them to identify significantly expressed genes. A particular attention is given to five biomarkers, from which one of them, IL8 was also detected by the GO enrichment analysis of our new subsided set of biomarkers. In order to test the power of the colon cancer-specific gene network biomarkers revealing ability, they use five different classifiers: Diagonal linear discriminate analysis, 3 Nearest neighbours, Nearest centroid, Support Vector Machines and Bayesian compound covariate.

3 The Methodology

In this section we present the new approach for biomarkers detection which is an extension of a previously defined methodology [1].

3.1 The Previous Methodology

The biomarkers which we use in this paper were detected by using the following methods:

- *Quantile normalization.* Since our aim is to unveil the difference in gene expression levels between the carcinogenic and healthy tissues, we proposed the Quantile normalization (QN) as a suitable normalization method [7].
- *Low entropy filter.* We used low entropy filter to remove the genes with almost ordered expression levels [8], since they lead to wrong conclusions about the genes behaviour.
- *Paired-sample t-test.* Knowing the facts that both carcinogenic and healthy tissues are taken from the same patients, and that the whole-genome gene expression follows normal distribution [9], we used a paired-sample t-test.
- *FDR method.* False Discovery Rate (FDR) is a reduction method that usually follows the t-test. FDR solves the problem of false positives, i.e., the genes which are considered statistically significant when in reality there is not any difference in their expression levels.
- *Volcano plot.* Both the t-test and the FDR method identify different expressions in accordance with statistical significance values, and do not consider biological significance. In order to display both statistically and biologically significant genes we used volcano plot visual tool.

3.2 The Novel Approach

After we discovered the significant genes, we proposed a generative approach by building the prior distributions of the two classes (CRC and healthy) which we used in the Bayes' Theorem to classify new patients.

Given the classes C_i for $i = 1, 2$ and a vector $\boldsymbol{x}$ of biomarkers gene expression values, we calculated the prior distributions, $p(\boldsymbol{x}|C_i)$, as a product of the continuous probability distributions of each biomarker distinctively:

$$p(\boldsymbol{x}|C_i) = \prod f_1 f_2 ... f_n \tag{1}$$

Once we determined the class-conditional densities, we applied them in the Bayes' theorem to obtain the a posteriori probability $P(C_i|\boldsymbol{x})$:

$$p(C_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_i) * P(C_i)}{\sum_{i=1}^{2} p(\boldsymbol{x}|C_i) * P(C_i)} \tag{2}$$

For the prior probabilities $P(C_i)$, we defined two test cases:

- Test Case 1: Since we have equal number of tissues into both of the classes, the prior probabilities are also equal $P(C_1) = P(C_2) = 0.5$;
- Test Case 2: The prior probabilities are estimated according to the statistics in [10]. Therefore, $P(C_1) = 0.0002$ and $P(C_2) = 0.9998$, where C_1 denotes the carcinogenic class, and C_2 denotes the healthy class.

A new set of biomarkers gene expression, $\boldsymbol{x}$, is classified according to the rule of maximizing the a posteriori probability (MAP):

$$C_i = \max p(C_i|\boldsymbol{x}) \tag{3}$$

By using this methodology we achieved very high classification accuracy whose results are presented in Section 5.

The high sensitivity and specificity results from the Bayesian classifier intrigued us to go into more detail and make the prior distributions even more precise. In order to achieve our aim, we need to reduce the number of genes whose prior distribution varies when compared to the prior distributions of the majority of the genes.

That is the origin of the idea to use the generative Bayesian model as an additional method for biomarkers detection. As described in Figure 1, the method in based on iterative leave-one-out classification until we reach a set of genes whose classification power is higher than the initial set of biomarkers. Therefore the method was applied as follows.

Let's have n number of biomarkers. Since we wanted to reduce the biomarkers set as well as to sustain the good features of Distribution models for Bayes classification process, we performed $n * \frac{3}{4}$ iterations. We chose this number of iterations to be fixed since the analysis reported in our previous research [2] showed that this number of biomarkers is approximately optimal for Bayesian classification.

In each iteration $i = 0, ..., n * \frac{3}{4}$ we perform $n - i$ retrainings and classifications by cutting-off one biomarker in each one. Once we obtain the results from all the classifications in the particular iteration, we chose the biomarker that have degraded the classification results at most, and put it into the new set of biomarkers. That biomarker is excluded from the initial set of biomarkers. The initial set of biomarkers is now reduced by one biomarker and we can proceed to the next iteration. Eventually, after $n * \frac{3}{4}$ iterations we complete the new set of biomarkers which will consist of $n * \frac{3}{4}$ biomarkers - the subsided gene set.

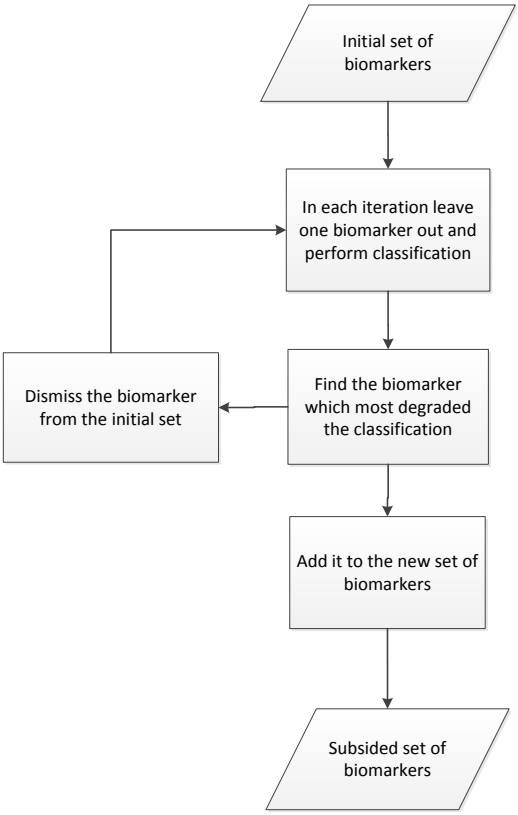

Fig. 1. The novel approach

4 GO Analysis

In the past the analyses of single markers have been in the focus of the genome-wide association studies. However, it often lacks the power to uncover the relatively small effect sizes conferred by most genetic variants. Therefore, using prior biological knowledge on gene function, pathway-based approaches have been developed with the aim to examine whether a group of related genes in the same functional pathway are jointly associated with a trait of interest [11].

The goal of the GO Consortium is to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing [12]. The GO project provides ontologies to describe attributes of gene products in three non-overlapping domains of Molecular Biology [13]:

1. **Molecular Function** describes activities, such as catalytic or binding activities, at the molecular level. GO molecular function terms represent activities rather than the entities that perform the actions, and do not specify where, when or in what context the action takes place.
2. **Biological Process** describes biological goals accomplished by one or more ordered assemblies of molecular functions.
3. **Cellular Component** describes locations, at the levels of subcellular structures and macromolecular complexes.

There are many tools based on GO resource; however, in this research we use the freely accessible Gene Ontology Enrichment Analysis Software Toolkit (GOEAST). It is a web based tool which applies appropriate statistical methods to identify significantly enriched GO terms among a given list of genes. Beside the other functions, GOEAST supports analysis of probe set IDs from Affymetrix microarrays. It provides graphical outputs of enriched GO terms to demonstrate their relationships in the three ontology categories. In order to compare GO enrichment status of multiple experiments, GOEAST supports cross comparisons to identify the correlations and differences among them [14].

We use cross comparisons of the old and new GO analyses to derive conclusions of the three Molecular Biology domains acquired from the subsided set of biomarkers.

5 Experiments and Results

In this section we present the experiments and the results obtained from the previously defined methodologies.

5.1 Microarray Data Analysis

In order to extract significant colorectal cancer genes we used gene expression profiling of 32 colorectal tumors, adenomas, and matched adjacent 32 non-tumor colorectal tissues probed with Affymetrix Human Genome U133 Plus 2.0 Array. It contains 54,675 probes, but the unique genes observed are 21,050.

The gene expression values were preprocessed according to the methodology described in Section 3.1. The methods produced a set of 138 biomarkers. The prior distributions of the biomarkers were modelled and a generative Bayesian classifier was produced whose results are reported in Table 1. In order to test the classifier on completely new patients, we used additional gene expression values from 239 CRC and 12 healthy patients.

Table 1. Old Sensitivity and Specificity

Chip	Performance	Sensitivity	Specificity	Test Cases
Affymetrix	Tissues	1	0.84	Test case 1
		0.94	1	Test case 2
	Patients	0.98	0.92	Test case 1
		0.90	1	Test case 2

Table 2. New classification results

Results	Test Case 1	Test Case 2
32 CRC Tissues	100%	96.87%
32 Healthy Tissues	81.25%	100%
239 CRC Patients	97.90%	94.14%
12 Healthy Patients	100%	100%

Table 3. New Sensitivity and specificity

Chip	Performance	Sensitivity	Specificity	Test Cases
Affymetrix	Tissues	1	0.81	Test case 1
		0.97	1	Test case 2
	Patients	0.98	1	Test case 1
		0.94	1	Test case 2

The new methodology from Section 3.2 reduced the set of genes by retaining the most important ones - the subsided biomarkers set. In Table 2 we present the results from the classification procedure where we performed classification analysis of 64 tissues from 32 patients, and additional 251 patients that weren't involved in the training process.

We evaluated the classificator performance through relative trade-off between the true positives and the false positives. True positive rate (TPR), the sensitivity, refers to the classifier's ability to correctly classify CRC tissues, whereas the ability of the classifier to correctly classify healthy tissues is measured in terms of specificity. The results from the new approach that improved the sensitivity and the specificity of the classifier are presented in Table 3.

Figure 2 depicts the comparison of the new sensitivity (green) and specificity (violet) results with the old sensitivity (blue) and specificity (red) results.

5.2 GO Results

In order to compare the GO results from the analysis of subsided set of biomarkers, we performed comparisons with the GO analysis of the old set of 138 biomarkers.

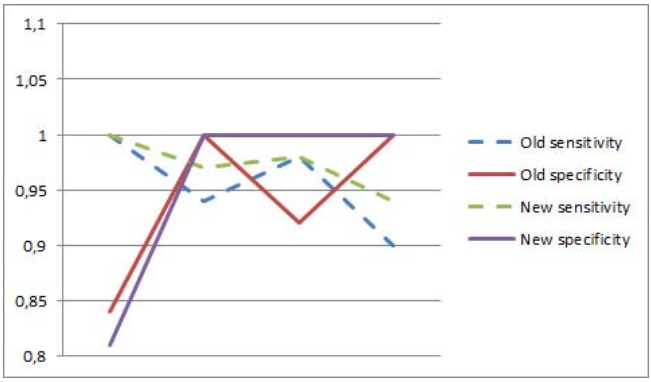

Fig. 2. Performance comparison

Biological Processes (BP). Figures 3 and 4 present the comparison of the biological processes (BP) of the two sets of biomarkers. Even though the subsided set of 100 genes is a subset of the old set of 138 genes, some of the processes that were not enriched in the previous analysis, now show significant enrichment. The results from the old analysis are marked with red, whereas the results from the new analysis are marked with green. All the common enriched terms are labelled with yellow. The newly enriched BP and their GO descriptions are as follows:

- **Negative regulation of cell proliferation** - Any process that stops, prevents or reduces the rate or extent of cell proliferation.
  Genes: SST, MSX2, CCL23, FABP6, IL8, SCG2.
- **Transmembrane receptor protein serine/threonine kinase signaling pathway** - A series of molecular signals initiated by the binding of an extracellular ligand to a receptor on the surface of the target cell where the receptor possesses serine/threonine kinase activity, and ending with regulation of a downstream cellular process, e.g. transcription.
  Genes: GREM2, MSX2, CHRDL1.
- **Indole-containing compound biosynthetic process** - The chemical reactions and pathways resulting in the formation of compounds that contain an indole (2,3-benzopyrrole) skeleton.
  Genes: TPH1

Molecular Functions (MF). Considering the comparison of the molecular functions, we see the following enriched functions are no longer present in the new analysis:

- **G-protein coupled receptor binding** - Interacting selectively and non-covalently with a G-protein coupled receptor.
- **Hormone activity** - The action characteristic of a hormone, any substance formed in very small amounts in one specialized organ or group of cells and

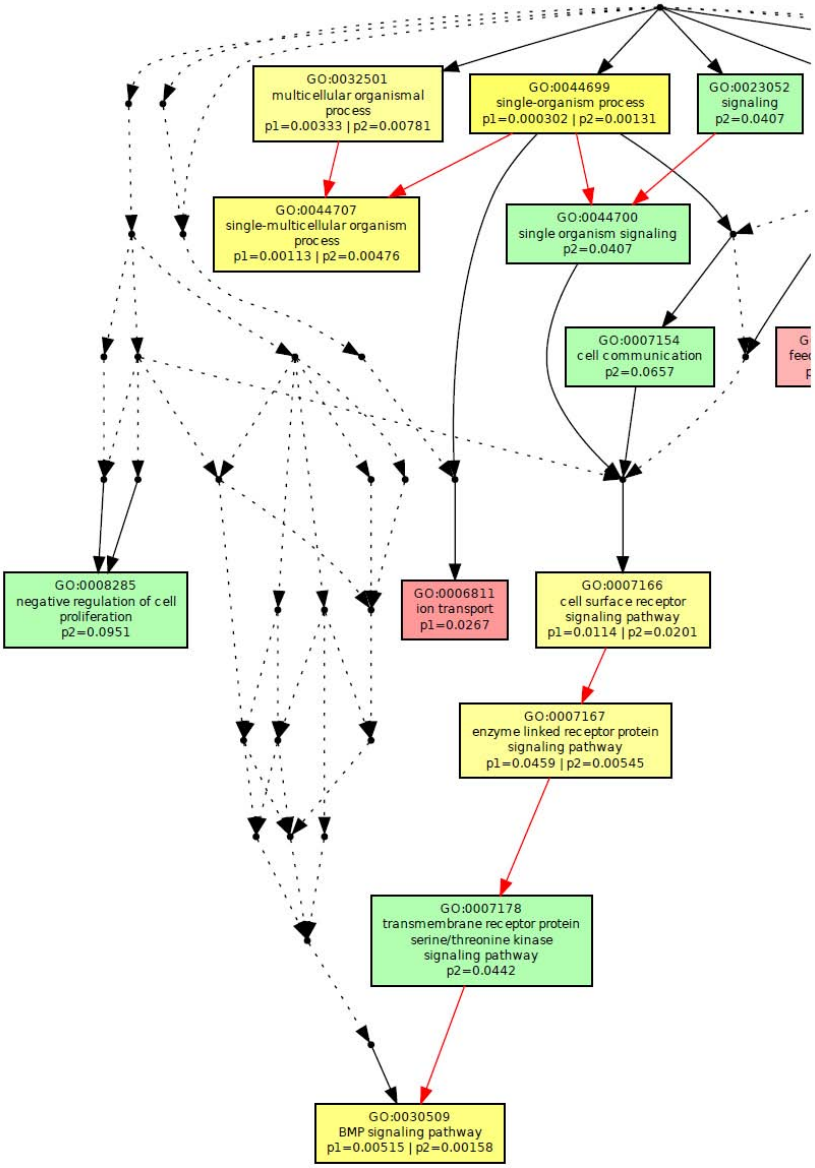

Fig. 3. Biological Processes Comparison (part 1)

carried (sometimes in the bloodstream) to another organ or group of cells in the same organism, upon which it has a specific regulatory action.

– **Alcohol dehydrogenase activity, zinc-dependent** - Catalysis of the reaction: an alcohol + NAD+ = an aldehyde or ketone + NADH + H+, requiring the presence of zinc.

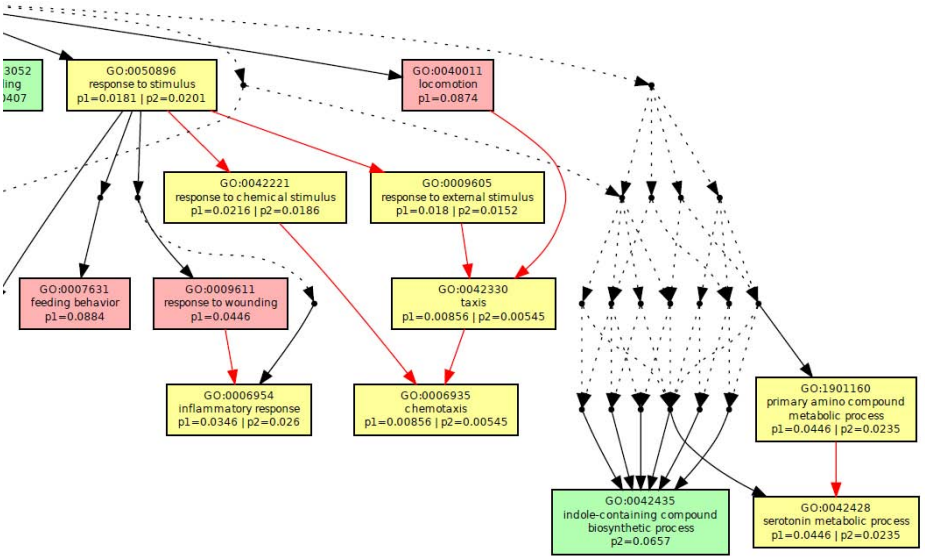

Fig. 4. Biological Processes Comparison (part 2)

– **Sodium channel activity** - Catalysis of facilitated diffusion of a sodium ion (by an energy-independent process) involving passage through a transmembrane aqueous pore or channel without evidence for a carrier-mediated mechanism.

Cellular Components (CC). Eventually, we compared the cellular components results and we found the following enriched terms are excluded in the new results:

– **Apical part of cell** - The region of a polarized cell that forms a tip or is distal to a base. For example, in a polarized epithelial cell, the apical region has an exposed surface and lies opposite to the basal lamina that separates the epithelium from other tissue.
– **Apical plasma membrane** - The region of the plasma membrane located at the apical end of the cell.
– **Sodium channel complex** - An ion channel complex through which sodium ions pass.

Further analysis of the relation of the new GO results to CRC will be presented in our future work, where we will discuss the results from a biological point of view.

6 Conclusion

The aim of this paper was to enforce the classification system created for CRC diagnosis - the Bayes classification process that uses the chosen biomarker set

[1]. We used the built generative Bayesian model as an additional method for meaningful reduction of the biomarkers set. We addressed this issue by choosing the biomarkers that contribute to the classification process the most. To achieve this goal, we proposed a new approach, based on iterative Bayesian classification. In order to eliminate the non-informative genes, we used a Leave-one-out method - we picked the ones that degrade the classification process when excluded from building the classification system. Taking into account only the reduced set of biomarkers (subsided set of biomarkers), we produced an improved version of our Bayesian classifier when classifying new patients and tissues.

We also engaged the GO analysis to understand the biological processes, the molecular functions and the cellular components when using the subsided biomarkers set. We compare the GO analysis of the initial set of biomarkers [2] with the analysis from the novel proposed methodology, with the subsided biomarkers set. The novel approach gives us an advantage in the GO analysis, because we can obtain more precise knowledge about the expressed genes and processes they are connected to in the CRC diagnosis. We obtained newly enriched BP and their GO descriptions and found out about enriched functions that are no longer present in the new analysis for the CC and MF.

Future work includes the investigation if the subsided biomarkers set can improve the methodology for CRC stages diagnostics [15], and a close collaboration with the Molecular Biology experts, that will validate our results for molecular diagnostics, evaluation and prognostic purposes in patients with colorectal cancer.

References

1. Simjanoska, M., Madevska Bogdanova, A., Popeska, Z.: Bayesian posterior probability classification of colorectal cancer probed with Affymetrix microarray technology. In: 2013 36th International Convention on Information & Communication Technology Electronics & Microelectronics (MIPRO), pp. 959–964. IEEE (2013)
2. Simjanoska, M., Madevska Bogdanova, A., Panov, S.: Gene ontology analysis of colorectal cancer biomarkers probed with Affymetrix and Illumina microarrays. In: Proceedings of the 5th International Joint Conference on Computational Intelligence, IJCCI 2013, pp. 396–406. IJCCI (2013)
3. Lascorz, J., Chen, B., Hemminki, K., Försti, A.: Consensus pathways implicated in prognosis of colorectal cancer identified through systematic enrichment analysis of gene expression profiling studies. PloS One 6(4), e18867 (2011)
4. Xu, Y., Xu, Q., Yang, L., Liu, F., Ye, X., Wu, F., Ni, S., Tan, C., Cai, G., Meng, X., et al.: Gene expression analysis of peripheral blood cells reveals toll-like receptor pathway deregulation in colorectal cancer. PloS One 8(5), e62870 (2013)
5. Chan, S.K., Griffith, O.L., Tai, I.T., Jones, S.J.: Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers. Cancer Epidemiology Biomarkers & Prevention 17(3), 543–552 (2008)
6. Jiang, W., Li, X., Rao, S., Wang, L., Du, L., Li, C., Wu, C., Wang, H., Wang, Y., Yang, B.: Constructing disease-specific gene networks using pair-wise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements. BMC Systems Biology 2(1), 72 (2008)
7. Wu, Z., Aryee, M.: Subset quantile normalization using negative control features. Journal of Computational Biology 17(10), 1385–1395 (2010)

8. Needham, C., Manfield, I., Bulpitt, A., Gilmartin, P., Westhead, D.: From gene expression to gene regulatory networks in arabidopsis thaliana. BMC Systems Biology 3(1), 85 (2009)
9. Hui, Y., Kang, T., Xie, L., Yuan-Yuan, L.: Digout: Viewing differential expression genes as outliers. Journal of Bioinformatics and Computational Biology 8(suppl. 01), 161–175 (2010)
10. GLOBOCAN (2008), http://globocan.iarc.fr/factsheets/cancers/colorectal.asp
11. Wang, K., Li, M., Hakonarson, H.: Analysing biological pathways in genome-wide association studies. Nature Reviews Genetics 11(12), 843–854 (2010)
12. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature Genetics 25(1), 25 (2000)
13. Harris, M., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., et al.: The gene ontology (go) database and informatics resource. Nucleic Acids Research 32(Database issue), D258 (2004)
14. Zheng, Q., Wang, X.J.: Goeast: a web-based software toolkit for gene ontology enrichment analysis. Nucleic Acids Research 36(suppl. 2), W358–W363 (2008)
15. Simjanoska, M., Bogdanova, A.M., Popeska, Z.: Bayesian multiclass classification of gene expression colorectal cancer stages. In: Trajkovik, V., Anastas, M. (eds.) ICT Innovations 2013. AISC, vol. 231, pp. 177–186. Springer, Heidelberg (2013)

Evaluating an Ordered List of Recommended Physical Activities within Health Care System

Igor Kulev[1], Elena Vlahu-Gjorgievska[2], Saso Koceski[3], and Vladimir Trajkovik[1]

[1] Faculty of Computer Science and Engineering,
University "Ss Cyril and Methodious", Skopje, Macedonia
{igor.kulev,trvlado}@finki.ukim.mk
[2] Faculty of Administration and Information Systems Management,
University "St.Kliment Ohridski", Bitola, Macedonia
elena.vlahu@uklo.edu.mk
[3] Faculty of Computer Science,
University "Goce Delcev", Stip, Macedonia
saso.koceski@ugd.edu.mk

Abstract. Information and communication technologies make it possible to bridge the gap and time barriers in the flow of health information and knowledge, allowing every involved part in the health process to have access to the information. This approach provides the knowledge of the individual to contribute effectively to the improvement in human health. But also, helps the collective knowledge effectively to solve health problems on individual level. In this paper we are evaluating the algorithm that generates recommendation for users. We are using simulations on generic data to see how different types of activities are affecting the accuracy of the algorithm. On the basis of the performed activities and blood glucose measurements, our recommendation algorithm should determine list of activities that have bigger influence on the change of the blood glucose levels. Generic data for our simulations are based on modeling of food intake and physical activity influence over the blood glucose level.

Keywords: recommendation algorithm, blood glucose level, evaluation.

1 Introduction

The advances in communication and computer technologies have revolutionized the way health information is gathered, stored, processed, and communicated to decision makers for better coordination of healthcare at both the individual and population levels [1]. Pervasive health care takes steps to design, develop, and evaluate computer technologies that help citizens participate more closely in their own healthcare [2], on one hand, and on the other to provide flexibility in the life of patient who lead an active everyday life with work, family and friends [3].

Life style with moderate eating habits and increased physical activity plays a key role in disease management. Some clinical conditions (like diabetes, metabolic

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_24

syndrome, chronic heart failure) can be prevented by proper diet and regular physical activity. There are number of studies that have shown that increased physical activity and diet modification (termed as 'lifestyle interventions'), independent of other risk factors, has a protective effect against the development of chronic diseases as diabetes and metabolic syndrome [4, 5]. Guidance and interactive training regarding appropriate choices of diet and exercise plans combined with encouragement and monitoring of progress, can empower patients to make beneficial lifestyle modifications [6].

The recommendation algorithm, evaluated in this paper, is part of the Collaborative Health Care System Model – COHESY [7]. COHESY is "a tool" for personal heathcare. It is deployed over three basic usage layers. The first layer consists of the bionetwork (that reads parameters' values from various body sensors) and a mobile application (that collects users' bio data and parameters of performed physical activities). The second layer is presented by the social network and the third layer enables interoperability with the primary/secondary health care information systems. The usage of social network and its' recommendation algorithm are the main advantages of COHESY. The social network enables different collaboration within the end user community. It allows communication between users, exchange of their experiences and gathering of large amount of data about their health parameters, food intake and performed activities. The recommendation algorithm takes into account the effects of food and physical activity on health parameters (e.g. blood glucose level), and based on prior knowledge (data gathered from social network and clinical centers) recommend physical activity that will improve the users' health.

The next section gives a brief overview of the recommendation algorithm. In the third section experimental methodology and result will be discussed. The fourth section is the conclusion of the paper.

2 Recommendation Algorithm

The main purpose of this algorithm is to find the dependency of the users' health condition, food intake and physical activities they perform. The algorithm incorporates collaboration and classification techniques in order to generate recommendations and suggestions for preventive intervention. To achieve this we consider the data read by the bionetwork (parameters values), the data about the user's physical activities, the user's medical record (obtained from a clinical centre) and the data contained in the user profile on the social network (so far based on the knowledge of the social network). We use classification algorithms on these datasets to group the users by their similarity. Use of classified data when generating the recommendation provides more relevant recommendations because they are enacted on knowledge for users with similar medical conditions and reference parameters.

Generally in our algorithm we use a similarity metrics in order to find the most similar users to the active user according to their medical history. We assume that if two users had the same combination of parameter values in the past, there is bigger probability that similar latent factors affect their health condition. For each user from the set of similar users we keep the details about the physical activities he performed

and the measurements of his health parameters. Further, we use only data from the active user and from the users most similar to him, and we calculate the usefulness of each type of physical activity. We analyze the history of activities and measurements of each user and we want to find the type of influence of each type of activity on each of the health parameters. For this purpose two measurements (value of the parameter) are selected for each activity – the most recent measurement before the execution of the activity and a measurement performed a particular time period after the execution of the activity. We do not choose the first measurement after the activity because a time is needed for the activity to show its effect. The difference between the next and the previous measurement approximates the influence of the activity on the parameter change. After this, we use the information about the usefulness of each activity in order to generate recommendations. For each user from the set of similar users (plus the active user) we obtain the most useful activity that could potentially improve his health condition. The activity which is declared as the most useful to most of the users is recommended to the active user.

Simulations made for the evaluation presented in next section are for one user that has food intake 3 times per day. Although the algorithm has more steps, which are in details explained in [8], for the evaluation covered in this paper the step for calculating the benefits of performing the activity is important. This step is presented below.

Find the benefits of performing an activity. The benefit of performing activity *a* by the user *u* for parameter *p* is calculated by Eq.(1).

$$V_{u,a,p} = \frac{importance_p \cdot \sum_{a_u} \left(\frac{|next_p(a_u) - prev_p(a_u)|}{timeSpan(duration(a_u))} \cdot validity(a_u) \cdot dir(a_u,p) \cdot intensity\,(a_u) \right)}{num(a_u)}, \quad (1)$$

$$validity(a_u) = validityPrev(a_u) \cdot validityNext(a_u),$$

$$\forall a, a \in A', \forall p, p \in P''$$

- *A'* – set of different activities;
- *P'* – set of health parameters;
- $next_p(a_u)$ - function that returns the value of the parameter after performing the activity *a*;
- $prev_p(a_u)$ - function that returns the value of the parameter before performing the activity *a*;
- $duration(a_u)$ - the duration of the activity *a*;
- $intensity\,(a_u)$ - the value of the intensity of performed activity *a*, calculated by an appropriate formula;
- $num\,(a_u)$ - the number of reviewed readings of activity *a*;
- $validity(a_u)$ - function that returns the validity of the measured activity *a*;
- $importance_p$ – the importance of this parameter for the user (the bigger importance of the parameter for the health of the user - the higher its coefficient is);
- $timeSpan(x)$ – logarithmic function.

$dir(a_u, p)$= -1 when we are calculating the benefits of an activity that decreases the value of the parameter. When the activity increases the value of the parameter the value of this function is 1.

3 Methodology

We assume that there are two factors that affect the parameter value: food intakes and activities. Food intakes tend to increase and activities tend to decrease the parameter value. We assume that after consuming the food, there is a short period of time where it causes rapid increase of the parameter value, and after that there is a short period of time where it causes decrease of the parameter value. After this "unstable" period of increase and decrease, as a result there is a small increase of the parameter value. The final change of the parameter value caused by the food intake happens 6-9 hours after the food intake and depends on the type of the food intake. Three food intakes happen each day: breakfast, lunch and dinner. These food intakes affect the parameter value with different intensity. Lunch has the largest effect and causes the biggest increase after the unstable period. Dinner causes smaller increase than lunch and breakfast causes the smallest increase. Breakfast, lunch and dinner happen around 09:00, 12:30 and 18:00 accordingly. The exact moment of occurrence of breakfast, lunch and dinner is determined in each day of the simulation by using a Gaussian distribution. The parameter function affected only by food intakes that happen in one day is given on Fig. 1. We define the maximal increase to be the largest increase caused by food intake and the stable increase to be the increase of the parameter value after the "unstable" period. The stable increases for breakfast, lunch and dinner are 0.2, 0.4 and 0.3 accordingly.

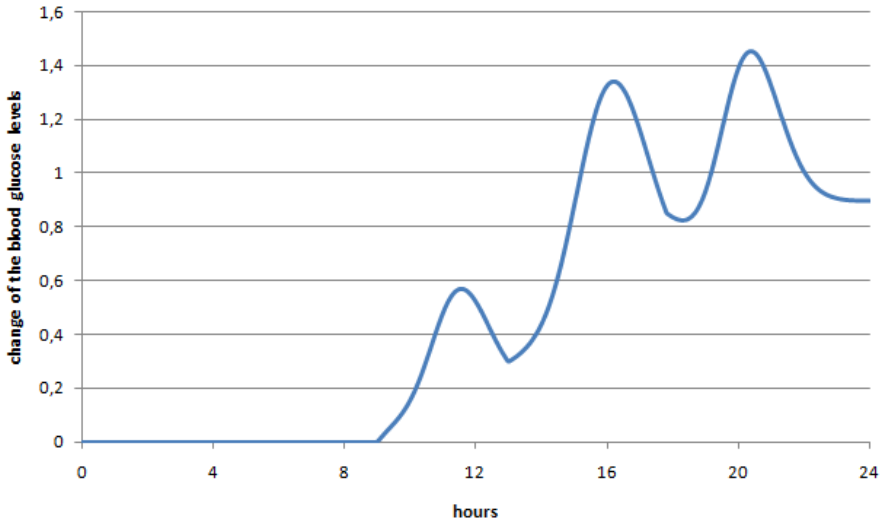

Fig. 1. Change of the blood glucose level during one day under the influence of breakfast, lunch and dinner and under no other kind of influence

Activities have opposite effect on the parameter value. After the activity there is a short period of time where it causes rapid decrease of the parameter value, and after that there is a short period of time where it causes increase of the parameter value. After this "unstable" period of decrease and increase, as a result there is a small decrease of the parameter value. The final change of the parameter value caused by the activities happens 3 days after the activity. Minimal decrease and stable decrease could be defined for activities in a similar way as for food intakes (Fig 2).

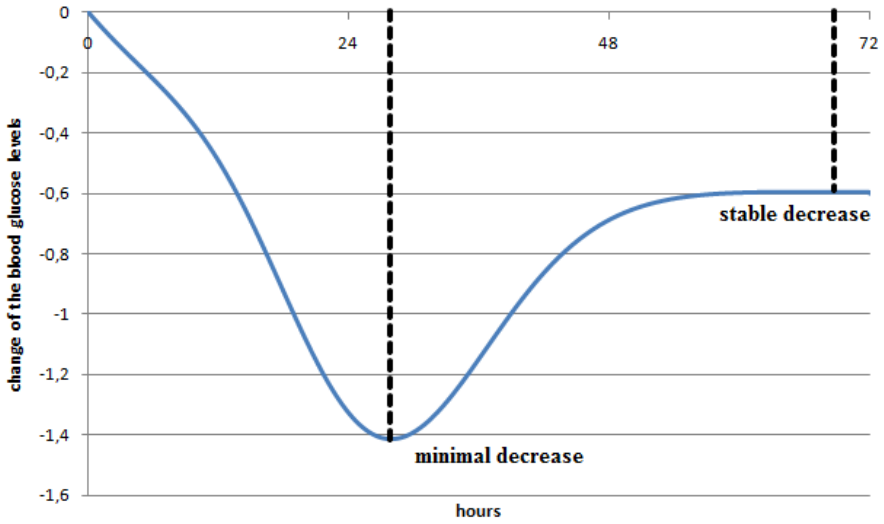

Fig. 2. Change of the blood glucose levels generated by one activity

We simulate 40 days in which the parameter value is changed by three different types of food intakes defined above and N different types of activities. The minimal decrease and the stable decrease caused by each type of activity are defined by the following formulas:

$$md_i = (-0.9) - \frac{(N-1)}{2} diffMD + (i-1) \cdot diffMD \quad (2)$$

$$sd_i = (-0.6) - \frac{(N-1)}{2} diffSD + (i-1) \cdot diffSD \quad (3)$$

where $1 \leq i \leq N$. Our simulator has three parameters: N, $diffMD$ and $diffSD$. $diffMD$ represents the difference between minimal decreases of two consecutive types of activities and $diffSD$ represents the difference between stable decreases of two consecutive types of activities. We generate the same number of activities of each type. When we choose the number of activities of each type we assume that the expected change of the parameter value in the end of the simulation is zero (or as close

to zero as possible). We calculate the number of activities of each type according to the formula:

$$numberOfActivitiesOfEachType = \frac{-(simulationLengthInDays \cdot (0.2+0.4+0.3))}{N \cdot (-0.6)} \quad (4)$$

The activities and 40 measurements are generated at random moments during one simulation. The change of the parameter value during one simulation is shown on Fig. 3a. The data that is provided to the recommendation algorithm (activities and measurements) is shown on Fig. 3b.

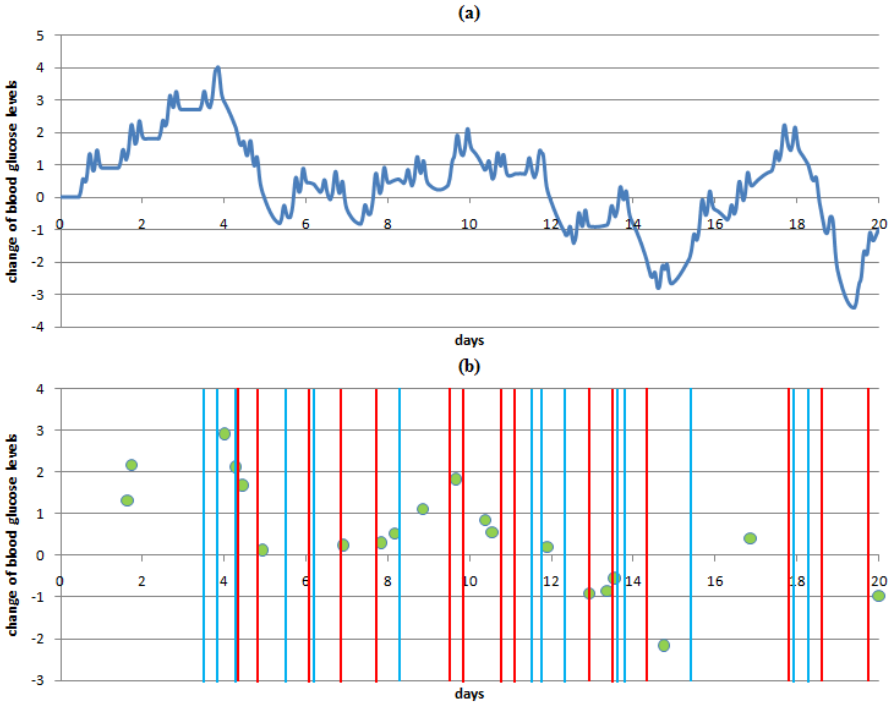

Fig. 3. a) Change of the blood glucose levels during one simulation; b) Data provided to the recommendation algorithm. The moments when the first type of activity has occurred are denoted with blue vertical lines. The moments when the second type of activity has occurred are denoted with red vertical lines. The measurements are denoted with green circles.

On the basis of the activities and measurements, our recommendation algorithm should determine the usefulness of all types of activities and should provide an ordered list of the types of activities according to their usefulness. If some type of activity has higher rank than other type of activity, this means that the algorithm has concluded that the first type of activity has bigger stable decrease than the second one. In this paper we want to evaluate how well the algorithm ranks the types of activities.

4 Results and Analysis

We have performed three different experiments. In the first experiment we have changed the value of $diffMD$ from 0.025 to 0.1 (in time intervals of length 0.025), in the second experiment we have changed the value of $diffSD$ from 0.025 to 0.1 (in time intervals of length 0.025) and in the third experiment we have changed both the values of $diffMD$ and $diffSD$ from 0.025 to 0.1 in the same time (in time intervals of length 0.025). In each experiment we have evaluated the quality of the ordered list using three evaluation metrics: Normalized Discounted Cumulative Gain (NDCG), Precision and Recall, and the Number of inversions.

4.1 Normalized Discounted Cumulative Gain

Normalized discounted cumulative gain (NDCG) measures the performance of a recommendation system based on the graded relevance of the recommended entities. It varies from 0.0 to 1.0, with 1.0 representing the ideal ranking of the entities. This metric is commonly used in information retrieval and to evaluate the performance of web search engines [9]. All three experiments are performed for different N from 2 to 12. The results of the evaluations using NDCG for all three experiments are shown on Fig. 4. It can be seen that the ranked list is relevant because the normalized discounted cumulative gain for each combination of values of N and $diffMD/diffSD$ is higher than the normalized discounted cumulative gain of random ordering of a list. Although we can conclude that the generated ordered list is relevant and better than random list, we cannot say how good the ordering is. Additionally, we cannot compare two NDCGs of results from experiments with different N. From Fig. 4 we can conclude that when we decrease $diffMD$ we get better results. Same happens when we decrease $diffSD$. In both cases we increase the absolute difference between $diffMD$ and $diffSD$. However, when we decrease $diffMD$ and $diffSD$ in the same time, we get worse results than if we decrease only one of the two ters: $diffMD$ or $diffSD$ (except in the case when $diffMD = -0.1$ and $diffSD = -0.1$). This is conclusion stands for all different values of N.

4.2 Precision and Recall

If we separate the types of activities in two groups: relevant and not relevant, then we can use the Precision and Recall measure to evaluate how much the relevant types of activities are ranked higher by the recommendation algorithm. Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved [10]. We have chosen that $N = 10$ in all simulations. We have marked the first 5 activities that have the highest $diffMD$ and $diffSD$ as relevant and we consider the others as not relevant. We have performed all three experiments and the results are shown on Fig. 5. These results show that the ordered list is relevant and they affirm the results obtained by the NDCG measure. Analyzing the Precision and Recall curves, we can also affirm the conclusion that the algorithm gives lower accuracy when we change $diffMD$ and $diffSD$ in the same time.

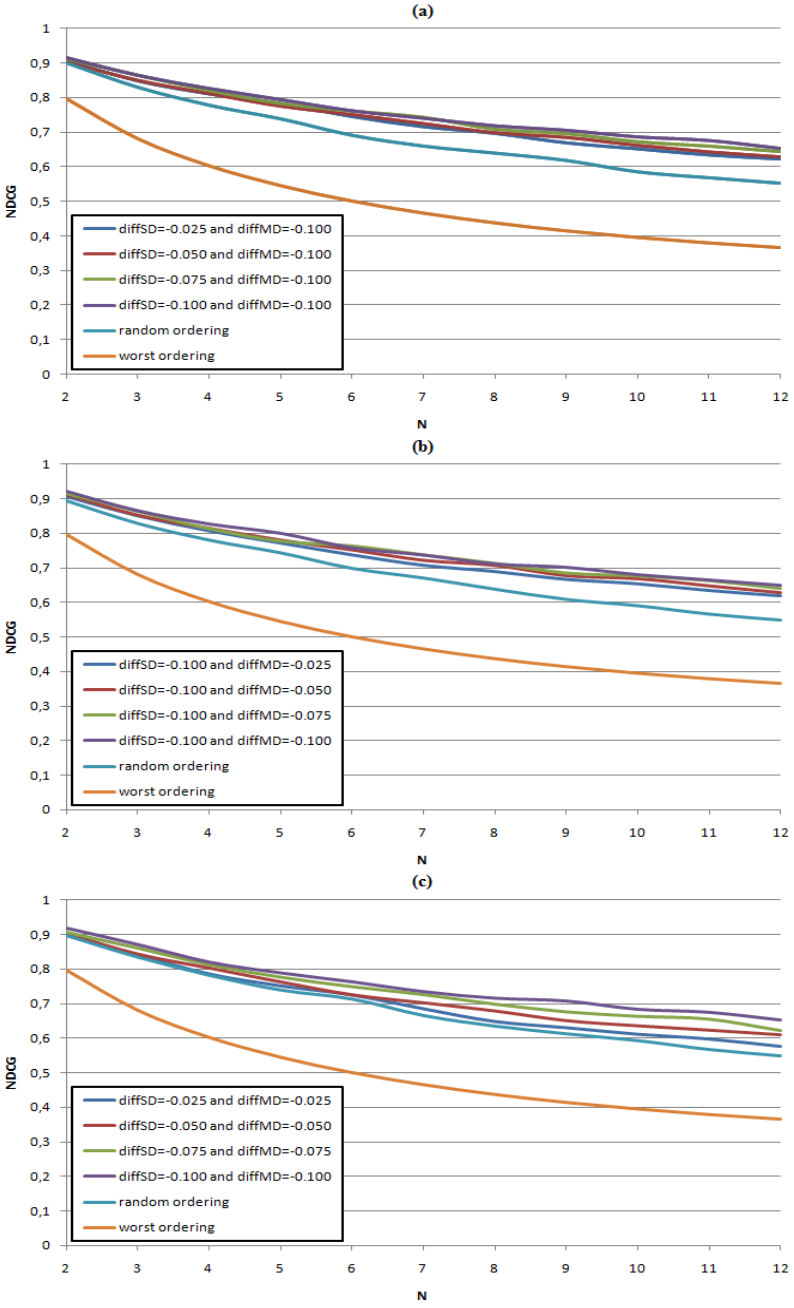

Fig. 4. Normalized Discounted Cumulative Gain for experiments with different number of activities

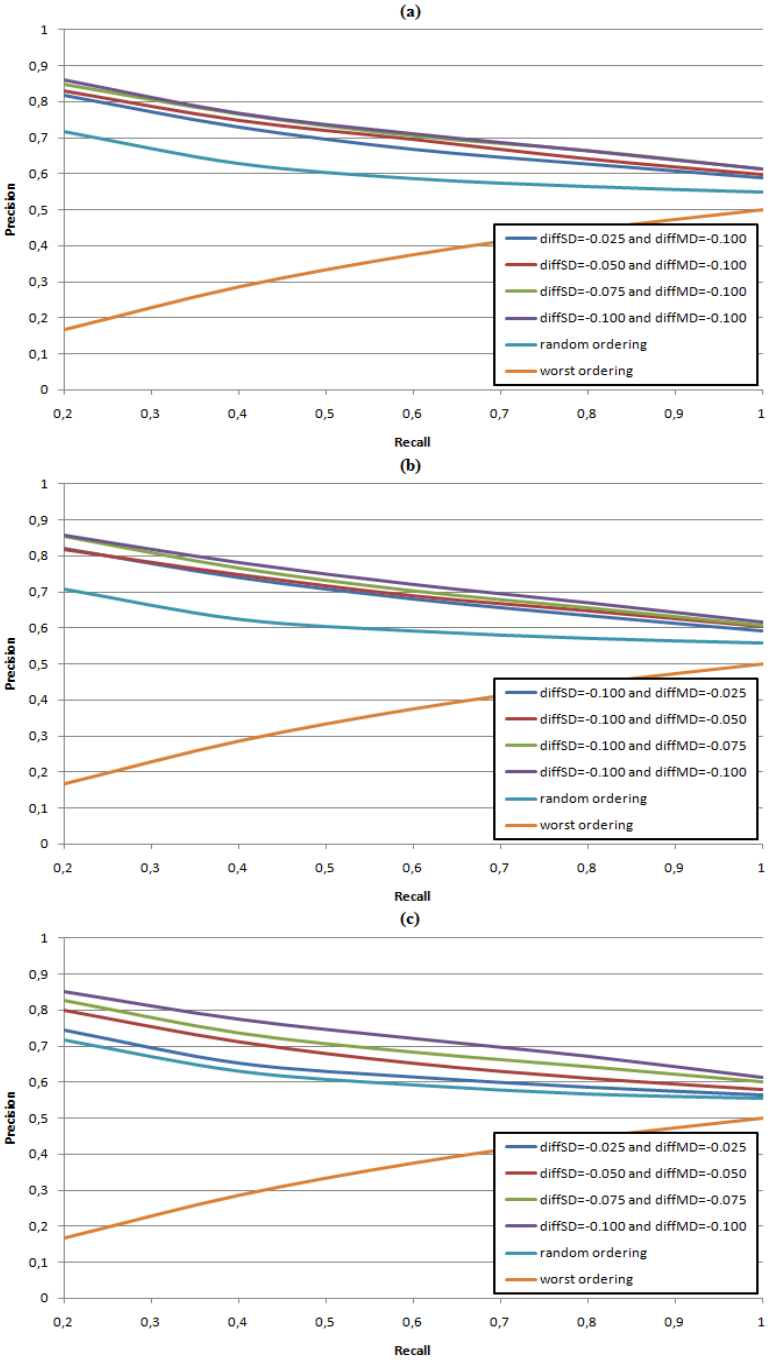

Fig. 5. Precision and Recall curves for experiments with 10 types of activities

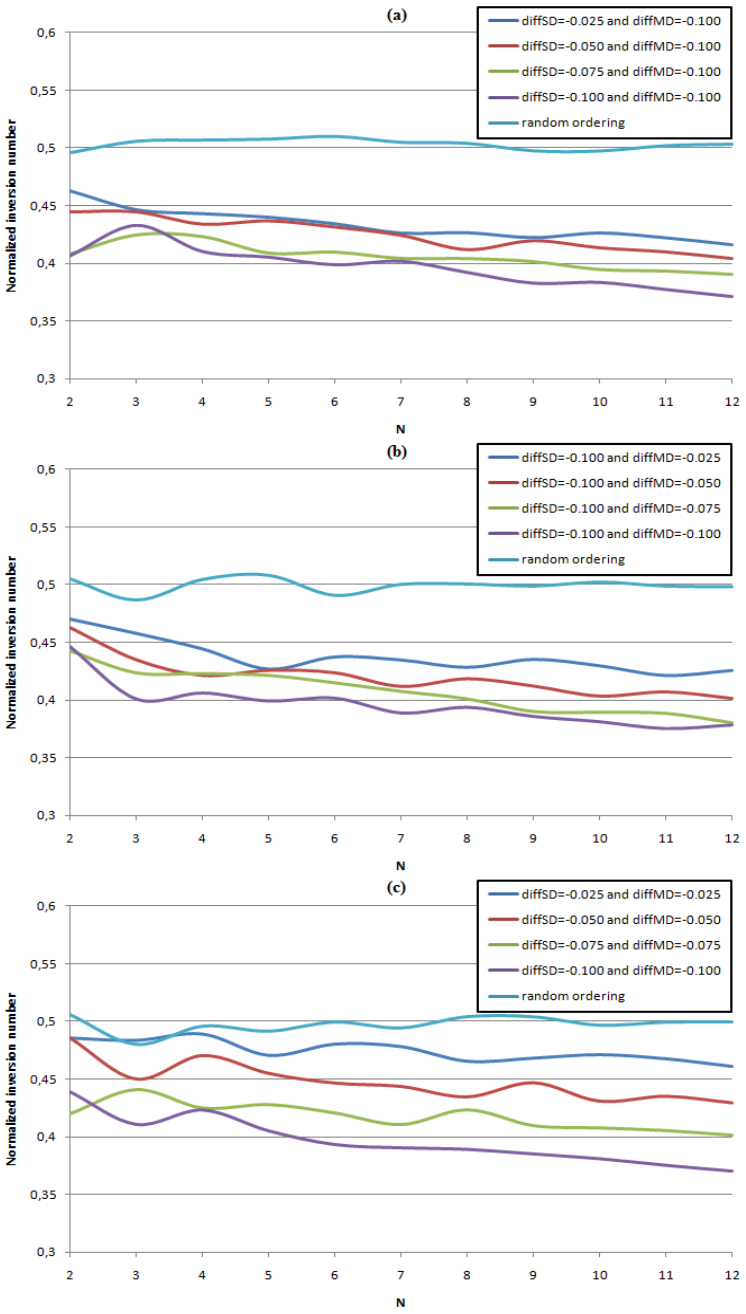

Fig. 6. Normalized inversion numbers for experiments with different number of activities

4.3 Number of Inversions

In computer science and discrete mathematics, an inversion is a pair of places of a sequence where the elements on these places are out of their natural order. The inversion number of a sequence is one common measure of its sortedness [11,12]. All three experiments are performed for different N from 2 to 12. The inversion number is normalized – the best ordering should have normalized inversion number 0, the worst ordering should have normalized inversion number 1 and the random ordering should have normalized inversion number 0.5. The results of our experiments are shown on Fig. 6. We can confirm the results obtained by the NDCG measure and the Precision and Recall measure. Additionally, using the normalized inversion number we can compare the performance of the algorithm for different N. We can see a linear decrease of the normalized inversion number in all curves obtained from our experiments. This means that as we increase the number of different types of activities we get better ordered list. The explanation of this result could be that as we increase the number of different types of activities, the difference between the type of activities with the biggest and smallest $diffMD/diffSD$ increases and it could be easier for the algorithm to conclude which of these activities produce a bigger/smaller stable decrease. Before the evaluation we weren't sure how the algorithm would behave when we increase N. This was because when the total number of activities increases, the global parameter function becomes more complicated so this could mean the algorithm might behave worse.

5 Conclusion

Our evaluation shows that the recommendation algorithm gives relevant ranking of the types of activities according to their usefulness. We have confirmed this conclusion by using three evaluation metrics: Normalized Discounted Cumulative Gain (NDCG), Precision and Recall, and the Number of inversions. We also conclude that the quality of the generated ordered list increases if the difference between the minimal decrease ($diffMD$) and the stable decrease ($diffSD$) is bigger or when the magnitude of both $diffMD$ and $diffSD$ is bigger. Increasing the number of different types of activities results in more complicated parameter function, but according to the simulation results it does not mean that the algorithm gives imprecise recommendations.

Acknowledgements. This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University. The authors would also like to acknowledge the contribution of the COST Action IC1303 - AAPELE, Architectures, Algorithms and Platforms for Enhanced Living Environments.

References

1. Fichman, R.G., Kohli, K., Krishnan, R.: Editorial Overview-The Role of Information Systems in Healthcare: Current Research and Future Trends. Information Systems Research 22(3), 419–428 (2011)
2. Ahamed, S.I., Haque, M.M., Khan, A.J.: Wellness assistant: a virtual wellness assistant using pervasive computing. In: Symposium on Applied Computing, pp. 782–787. ACM, USA (2007)
3. Ballegaard, S.A., Hansen, T.R., Kyng, M.: Healthcare in everyday life: designing healthcare services for daily life. In: Conference on Human Factors in Computing Systems, pp. 782–787. ACM, USA (2008)
4. Nachman, L., et al.: Jog Falls: A Pervasive Healthcare Platform for Diabetes Management. In: Floréen, P., Krüger, A., Spasojevic, M. (eds.) Pervasive 2010. LNCS, vol. 6030, pp. 94–111. Springer, Heidelberg (2010)
5. Orozco, L.J., Buchleitner, A.M., Gimenez-Perez, G.: Roqué i Figuls, M., Richter, B., Mauricio, D.: Exercise or exercise and diet for preventing type 2 diabetes mellitus. Cochrane Database of Systematic Reviews 3(Art. No.: CD003054) (2008)
6. Davidson, J.: Strategies for improving glycemic control: effective use of glucose monitoring. The American Journal of Medicine 118(suppl. 9A), 27S–32S (2005)
7. Vlahu-Gjorgievska, E., Trajkovik, V.: Towards Collaborative Health Care System Model – COHESY. In: 2011 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, pp. 1–6. IEEE Computer Society, Washington, DC (2011)
8. Kulev, I., Vlahu-Gjorgievska, E., Trajkovik, V., Koceski, S.: Development of a novel recommendation algorithm for collaborative health - care system model. Computer Science and Information Systems 10(3), 1455–1471 (2013)
9. Kalervo, J., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) 20(4), 422–446 (2002)
10. Nandish, C., Goyani, M.: Enhanced Multistage Content Based Image Retrieval (2013)
11. Philippe, F., Vitter, J.S.: Average-case analysis of algorithms and data structures (1987)
12. Wilhelm, B., Mutzel, P., Jünger, M.: Simple and efficient bilayer cross counting. Journal of Graph Algorithms and Applications 8(2), 179–194 (2004)

New Representation of Information Extracted from MRI Volumes Applied to Alzheimer's Disease

Katarina Trojacanec, Ivan Kitanovski, Ivica Dimitrovski, and Suzana Loshkovska

The Alzheimer's Disease Neuroimaging Initiative*
Ss. Cyril and Methodius University,
Faculty of Computer Science and Engineering, Skopje
Rugjer Boshkovik 16, P.O. Box 393,
Skopje, Macedonia
{katarina.trojacanec,ivan.kitanovski,
ivica.dimitrovski,suzana.loshkovska}@finki.ukim.mk

Abstract. The aim of the paper is to propose a new representation of the information extracted from the MRI volumes in the context of Alzheimer's Disease (AD). Two main stages are required: segmentation and estimation of the quantitative measurements. The representation is comprised of the quantitative measurements highlighted in the literature as valuable markers for distinguishing AD from healthy controls: cortical thickness of the separate parts of the brain cortex, the volume of the ventricular structures: left and right lateral ventricle, third and fourth ventricle, and the volume of the left and right hippocampus, left and right amygdala. Moreover, a representation that addresses the change of the patient's condition over the time is also proposed. An illustration and discussion of the proposed method is given by using the MRIs provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI).

Keywords: Medical volumetric data, VOI, Alzheimer's Disease, ADNI, segmentation, quantitative measurements.

1 Introduction

Functional and structural neuroimaging is a valuable marker for Alzheimer's Disease (AD) [1]. This leads to a vast amount of medical imaging collections available nowadays that need to be efficiently and precisely searched. The first and very important step to achieve that is appropriate representation of the clinically relevant information extracted from the medical volumes. This is exactly the main subject of research in this paper.

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:
http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_25

The standard procedure in such case, coming from the traditional Content-Based Image Retrieval (CBIR) systems [2], begins with generation of the image features derived from the visual cues contained in an image. This procedure is based on two type of features, photometric (exploiting color and texture cues), and geometric (using the shape-based cues) [2] extracted from the Volume of Interest (VOI). Characteristic examples from AD point of view are intensity and texture features extracted from VOIs [3,4,5], and shape features [6].

However, there are other markers that make possible the distinction between healthy controls and patients with AD. Huge amount of studies have been performed to detect and highlight possible indicators for AD, and subsequently to analyze their statistical dependence with respect to the disease [7,8,9,10,11]. These include cortical thickness [1], [7], [12,13], ventricular structures [1], [7], [10,11], hippocampus [1], [8,9,10], [13] amygdala [9]. Although the traditional way for generating image representation is possible in the case of the AD-based CBIR, the domain knowledge and the research conducted on the aforementioned structures/indicators could lead to another, alternative way for image representation. Some of them are used in the literature for classification separately (or combined with features from other imaging modalities or biomarkers), or as a combination of only few of them [14,15,16].

A good representation regarding the medical imaging domain and particularly AD subdomain should address several very important aspects: (1) key information extracted from the medical volume itself; (2) the change of the patient's condition; and (3) efficiency. A new representation of MRI volumes on the bases of the quantitative measurements estimated from the segmented valuable brain structures, addressing all three aforementioned aspects is proposed in this paper. This representation consists of the cortical thickness of separate parts of the brain cortex, as well as the volume of the left and right hippocampus, left and right amygdala, and the ventricular structures: left and right lateral ventricle, third ventricle, and fourth ventricle. An extended representation that incorporates the information about the changes in the patient's condition in certain time point is also proposed.

The paper is organized as follows. Section 2 represents the state of the art regarding the segmentation and quantitative measurements. Section 3 describes the proposed representation of the information extracted from the MRI volumes, the implementation details and the discussion about its application to the real volumes. Section 4 provides the concluding remarks and future directions.

2 State of the Art

The proposed representation of the information extracted from the medical volumetric data obtained from MRI is based on the quantitative measurements of particular structures in the human brain. The selection of the structures is based on the studies performed to analyze their statistical dependence with respect to the AD provided by the literature. Two main phases are required: segmentation and estimation of the quantitative measurements on the segmented structures. The state of the art is provided in the next subsections.

2.1 Segmentation

The first step to extract the key information is the segmentation of the anatomical structures in the human brain. The detection of the volumes of interest is very important to focus better and extract automatically localized visual information [17]. Moreover, the segmentation enables localized monitoring of the pathology or changing state of certain specific body part/region that usually is closely related to the progression of the disease/abnormality.

The segmentation that is suitable for brain anatomical segmentation is required to obtain the VOIs assumed to be valuable indicators for AD. There are several studies that provide methods/software tools for conducting segmentation of the anatomical structures important for AD detection and progression monitoring. For instance, the authors of [9] use a modified multi-atlas segmentation framework for hippocampus extraction. The authors of [18] have proposed a fully automated method for the extraction of the hippocampus using probabilistic and anatomical priors. The segmentation of 83 regions covering the whole brain is conducted by [19]. The results of their work are publicly available.

Additionally, several software tools exist and are widely used by researchers in this domain, enabling segmentation of different structures in the human brain such as FreeSurfer software package [20] used for cortical and subcortical segmentation [7], [12], Brain Ventricular Quantification (BVQ) software [21] for ventricular segmentation [11], Statistical Parametric Mapping (SPM) software package for White Matter (WM), Grey Matter (GM), and Cerebrospinal Fluid (CSF) segmentation [22,23], Automatic Lateral Ventricle delIneatioN (ALVIN) for lateral ventricle segmentation [10], as well as the FIRST tool as a part of FMRIB Software Library (FSL) [10].

2.2 Quantitative Measurements

After the segmentation has been performed, the estimation of the indicators for AD could be performed. For this purpose, some of the software packages mentioned in the context of segmentation provide useful measures such as ventricle volume [20,21], cortical thickness [20] etc.

3 The Proposed Feature Representation

3.1 Feature Representation

In this paper, we propose how to use the representation of the information extracted from the MRI volumetric data based on the quantitative measurements that are assumed to be valuable indicators for AD:

1. For the purpose of representation of the single scan, we propose to use the left and right lateral ventricle, the third and the fourth ventricle volume, the volume of the left and right hippocampus, left and right amygdala, as well as the cortical

thickness of the separate cortical structures (34 for each hemisphere). This leads to a total of 76 features as a representation appropriate to describe single scan of a patient. It should be noted that this dimensionality is relatively small in comparison to the traditional descriptors, e.x. 13312 features for 3D Gray Level Co-occurrence Matrices, 1920 for 3D Wavelet Transforms, 9216 for Gabor Transforms, and 11328 for 3D Local Binary Patterns for one volume [24].

2. To address the change of the patient's condition we propose extended representation. This is possible only if multiple scans for the same patient exist along certain period of time (e.x. the first visit scan, and scans obtained after three, six, nine, … months). In this case, we propose to use the same information as in the single scan scenario, but obtained from all scans and grouped by feature type. This leads to a structured representation where for each feature type, a tuple of measures is provided from all scans. The dimensionality of the tuples depends on the number of scans provided for the patients in the certain study. For example, if three scans are provided for each patient (e.x. the first visit scan, and the scans obtained after three and six months) for the feature type left lateral ventricle, a triple consists of the volume measures of the left lateral ventricle obtained from the three scans will take place in the feature representation, and so on for all proposed feature types. This is important to obtain full information about the changes that occur during the analyzed period of time. In this scenario, the dimension of the feature vector will be "the number of scans" times bigger. One very important question arises here, that is addressing the difference in the dimensionality if different number of scans is available for different patients. One trivial solution is to use only those subjects who have all scans for the examined time points. For example, if scans are available for the baseline, after 6, 12, 18 and 24 months for the most of the patients and for some of them only for the baseline, after 6, 12, and 24 months, one solution is to exclude the patients with no sufficient information available. The drawback of this solution might be excluding a possibly large group of patients. Another solution is generation of an artificial sample for the missing scan. This can be addressed at different levels. First direction here is making an artificial model of the brain (or the separate relevant structures) which will represent the missing scan and will be derived on the bases of the brain anatomical variation studies over the time. The second direction is deriving the information of the missing scan at the level of the quantitative measures. This can be conducted using the information obtained from the patients who have full series of scans to predict the missing information. The third solution is combing the information from the available scans for each patient to obtain each feature type as a template. This will result in the dimensionality that is the same as for the single scan case.

3.2 Implementation Details

The implementation details for the proposed method are described in the following subsections.

The Dataset. The imaging data used in the preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations as a $60 million, 5-year public–private partnership. Investigation on whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, such as cerebrospinal fluid (CSF) markers, APOE status and full-genome genotyping via blood sample, as well as clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and Alzheimer's Disease (AD) has been the primary goal of ADNI. Determination of sensitive and specific markers of very early AD progression is intended to assist the development of new treatments, simplify and increase the ability to monitor their effectiveness, and reduce the time and cost of clinical trials.

The principal investigator of the initiative is Michael W. Weiner, MD, Veteran's Affairs Medical Center and University of California, San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations. The adults, aged 55 to 90 years, have been recruited from over 50 sites across the U.S. and Canada, resulting in more than 1500 participants for ADNI and its followers ADNI-GO and ADNI-2. The participants include cognitively normal individuals, adults with early or late MCI, and people with early AD with different follow up duration of each group, specified in the protocols for ADNI-1, ADNI-2, and ADNI-GO. For up-to-date information, see http://www.adni-info.org.

Segmentation Applied to MRI and Output Usage for the Purpose of the Proposed Representation. For the purpose of the proposed method, the FreeSurfer software package was used to obtain the required measurements. Among other things, it enables cortical and subcortical segmentation, image registration, cortical thickness estimation and longitudinal processing.

Discussion. The segmentation procedure and estimation of the quantitative measurements on real patients were applied on the patients with diagnosed AD (an example shown on fig. 1 (a) and (c)) and normal controls (NL) (an example shown on fig. 1 (b) and (d)). Those images contain the subcortical structures of interest (left lateral ventricle (1), right lateral ventricle (2), the third ventricle (3), left (5) and right (6) hippocampus, and left (7) and right (8) amygdala are denoted on fig. 1 (a) and (b), while a good view on the fourth ventricle (4) is depicted on fig. 1 (c) and (d).

It should be noted that the structures labeled with 1, 2, 3, and 4, are clearly enlarged and the atrophy of the structures labeled as 5 and 6 is evident in the patient with AD (fig. 1 (a) and (c)) in comparison to NL (fig. 1 (b) and (d)). This confirms the validity of the use of these parameters as features in the feature vector (and is in accordance to the literature as well). However, the situation with amygdala (labeled as 7 (left) and 8 (right)) is opposite. Exactly for this reason, we propose to use a combination of the indicators that are stressed as relevant for AD, and not only a few of them.

From the point of view of the cortical analysis (the rest of the features in the proposed feature vector), the thickness map is depicted on fig. 2 for patient with AD (a) and for NL (b). The red color denotes thinner, while the yellow thicker. It is clear that the patient with AD has more thinner parts in comparison to NL.

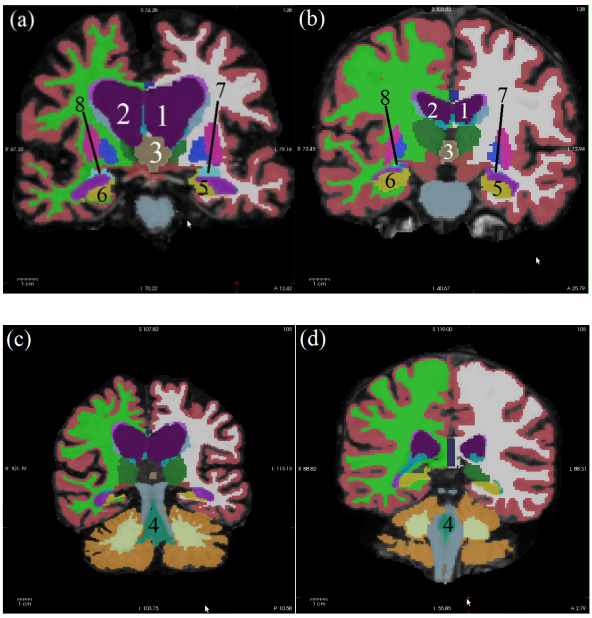

Fig. 1. Illustration of the left and right lateral ventricle, the third ventricle, the left and right hippocampus, and the left and right amygdala in patient with AD (a), and NL (b) and the fourth ventricle in patient with AD (c), and NL (d)

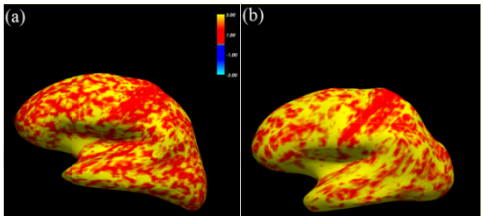

Fig. 2. Illustration of the cortical thickness map in patient with AD (a), and NL (b)

An example of the relevant VOIs extracted from MRI from patient who has been diagnosed with AD at the first visit/screening (a) and after 24 months (b) (coronal and sagittal view), is given on fig. 3. The progression of the disease is evident. The indicators that are clearly notable in this case include left (1) and right (2) lateral ventricle enlargement, and the third (3) ventricle enlargement. To be more precise, the graphical representation of the volume changes of the relevant structures is depicted on fig. 4. It confirms the ventricle enlargement over a period of 24 months (including the 6^{th} and the 12^{th} month). Considering the forth ventricle, the hippocampus and amygdala,

the volume/atrophy change is not always strictly increasing/decreasing. A possible solution to overcome this is a longitudinal processing including common information from the within-subject template, recommended to increase significantly the reliability and statistical power [25].

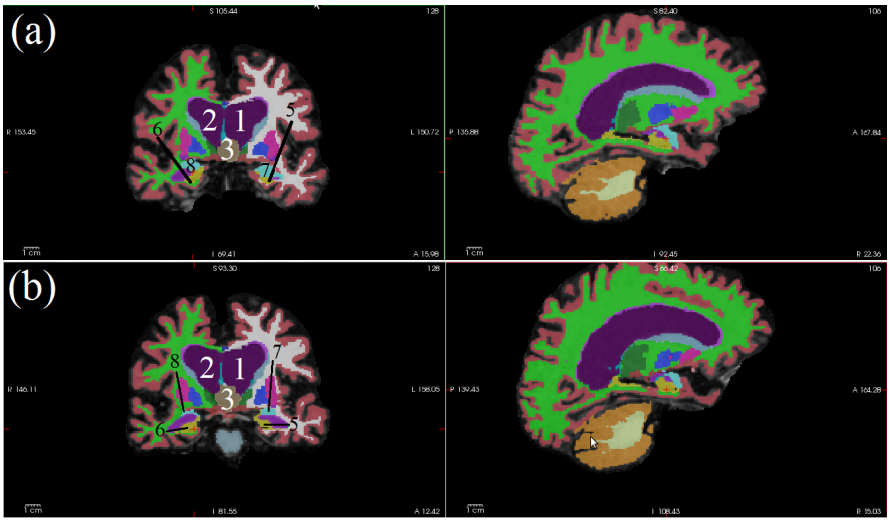

Fig. 3. Volume changes in patient with AD at the first screening (a) and after 24 months (b)

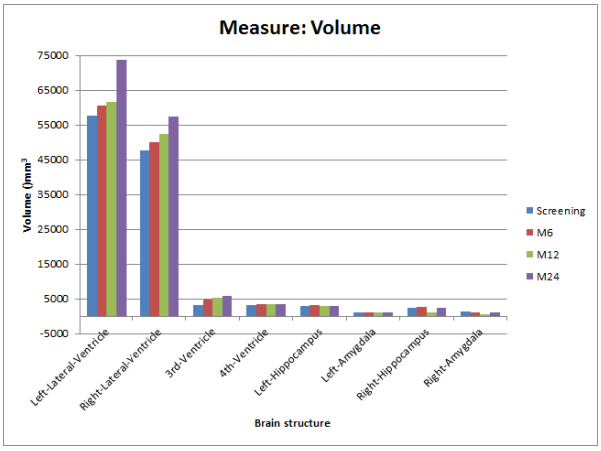

Fig. 4. The volume changes in patient with AD over 24 months: first visit, after six months, after 12 months, and after 24 months

Slightly changes in the cortical thickness again in patient with AD are visible on fig. 5. On this figure, the two time points are the first visit and the 24th month.

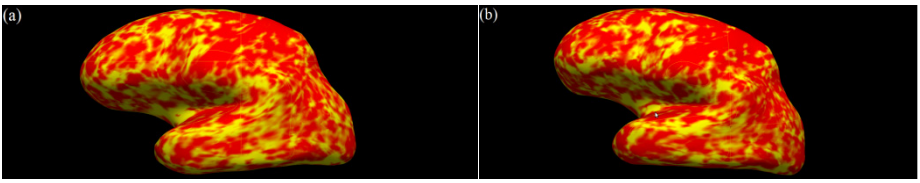

Fig. 5. Cortical thickness in patient with AD at the first screening (a) and after 24 months (b)

The proposed strategy to make a combination of the relevant indicators for AD as a representation is based on the clinically relevant information in the context of AD. The benefits that arise are regarding (1) its comprehensiveness (considering not only one or few relevant indicators), (2) efficiency, and (3) the suitability to be extended to address the patient's condition changes over the time.

4 Conclusion and Future Work

A new representation of the information extracted from the MRI volumes in the context of Alzheimer's Disease was proposed in the paper. The proposed method is suitable in both, single-scan studies, or multiple-scan studies (provided at certain time points). The representation includes the measures of the ventricle volume: left and lateral ventricle, third and fourth ventricle, the volume of the left and right hippocampus, and left and right amygdala, as well as the cortical thickness of the separate cortical structures. The main advantages of the proposed representation include comprehensiveness, suitability to reflect the patient's condition changes over the time, and efficiency.

The future work will be aimed to incorporate this kind of representation for the purpose of content based image retrieval and afterwards to make a comprehensive evaluation in the context of specialized AD-based CBIR. The evaluation will be performed for the single scan scenario, and even more important, in the case of multiple periodical scans available for the patients. This will contribute to better detect, predict, and understand the condition of the patient as well as the disease progression and/or treatment reaction.

Acknowledgement. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace,

Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

1. Mazziotta, J.C., Toga, A.W., Frackowiak, R.S. (eds.): Brain Mapping: The Disorders. Academic Press (2000)
2. Akgül, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B.: Content-Based Image Retrieval in Radiology: Current Status and Future Directions. J. Digit. Imag. 24(2), 208–222 (2011)
3. Chincarini, A., Bosco, P., Calvini, P., Gemme, G., Esposito, M., Olivieri, C., Rei, L., et al.: Local MRI Analysis Approach in the Diagnosis of Early and Prodromal Alzheimer's Disease. Neuroimage 58(2), 469–480 (2011)
4. Agarwal, M., Mostafa, J.: Content-Based Image Retrieval for Alzheimer's Disease Detection. In: 9th International Workshop on Content-Based Multimedia Indexing (CBMI), Madrid, Spain, pp. 13–18 (2011)
5. Agarwal, M., Mostafa, J.: Image Retrieval for Alzheimer's Disease Detection. In: Caputo, B., Müller, H., Syeda-Mahmood, T., Duncan, J.S., Wang, F., Kalpathy-Cramer, J. (eds.) MCBR-CDS 2009. LNCS, vol. 5853, pp. 49–60. Springer, Heidelberg (2010)
6. Yang, S.T., Lee, J.D., Huang, C.H., Wang, J.J., Hsu, W.C., Wai, Y.Y.: Computer-Aided Diagnosis of Alzheimer's Disease Using Multiple Features with Artificial Neural Network. In: Zhang, B.-T., Orgun, M.A. (eds.) PRICAI 2010. LNCS (LNAI), vol. 6230, pp. 699–705. Springer, Heidelberg (2010)
7. Moore, D.W., Kovanlikaya, I., Heier, L.A., Raj, A., Huang, C., Chu, K.W., Relkin, N.R.: A Pilot Study of Quantitative MRI Measurements of Ventricular Volume and Cortical Atrophy for the Differential Diagnosis of Normal Pressure Hydrocephalus. Neurology Research International 2012 (2011)
8. Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., et al.: Multidimensional Classification of Hippocampal Shape Features Discriminates Alzheimer's Disease and Mild Cognitive Impairment from Normal Aging. Neuroimage 47(4), 1476–1486 (2009)
9. Lötjönen, J., Wolz, R., Koikkalainen, J., Julkunen, V., Thurfjell, L., Lundqvist, R., Waldemar, G., Soininen, H., Rueckert, D.: Fast and Robust Extraction of Hippocampus from MR Images for Diagnostics of Alzheimer's Disease. Neuroimage 56(1), 185–196 (2011)
10. Leonardo, I.: Atrophy Measurement Biomarkers using Structural MRI for Alzheimer's Disease. In: The 15th Int. Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) (2012)

11. Nestor, S.M., Rupsingh, R., Borrie, M., Smith, M., Accomazzi, V., Wells, J.L., Fogarty, J., Bartha, R.: Ventricular Enlargement as a Possible Measure of Alzheimer's Disease Progression Validated using the Alzheimer's Disease Neuroimaging Initiative Database. Brain 131(9), 2443–2454 (2008)
12. Yuan, L., Wang, Y., Thompson, P.M., Narayan, V.A., Ye, J.: Multi-Source Feature Learning for Joint Analysis of Incomplete Multiple Heterogeneous Neuroimaging Data. NeuroImage 61(3), 622–632 (2012)
13. Sabuncu, M.R., Desikan, R.S., Sepulcre, J., Yeo, B.T.T., Liu, H., Schmansky, N.J., Reuter, M., et al.: The Dynamics of Cortical and Hippocampal Atrophy in Alzheimer Disease. Archives of Neurology 68(8), 1040–1048 (2011)
14. Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D.: Random Forest-Based Similarity Measures for Multi-Modal Classification of Alzheimer's Disease. NeuroImage 65, 167–175 (2013)
15. Gray, K.R., Wolz, R., Heckemann, R.A., Aljabar, P., Hammers, A., Rueckert, D.: Multi-Region Analysis of Longitudinal FDG-PET for the Classification of Alzheimer's Disease. NeuroImage 60(1), 221–229 (2012)
16. Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O.: Automatic Classification of Patients with Alzheimer's Disease from Structural MRI: A Comparison of Ten Methods using the ADNI Database. Neuroimage 56(2), 766–781 (2011)
17. Müller, H., Greenspan, H.: Overview of the Third Workshop on Medical Content–Based Retrieval for Clinical Decision Support (MCBR–CDS 2012). In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 1–9. Springer, Heidelberg (2013)
18. Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O.: Fully Automatic Hippocampus Segmentation and Classification in Alzheimer's Disease and Mild Cognitive Impairment Applied on Data from ADNI. Hippocampus 19(6), 579–587 (2009)
19. Heckemann, R.A., Keihaninejad, S., Aljabar, P., Gray, K.R., Nielsen, C., Rueckert, D., Hajnal, J.V., Hammers, A.: Automatic Morphometry in Alzheimer's Disease and Mild Cognitive Impairment. Neuroimage 56(4), 2024–2037 (2011)
20. FreeSurfer, `https://surfer.nmr.mgh.harvard.edu`
21. Accomazzi, V., Lazarowich, R., Barlow, C.J., Davey, B.: U.S. Patent No. 7,596,267. U.S. Patent and Trademark Office, Washington, DC(2009)
22. Cataldo, R., Agrusti, A., De Nunzio, G., Carlà, A., De Mitri, I., Favetta, M., Quarta, M., Monno, L., Rei, L., Fiorina, E.: Generating a Minimal Set of Templates for the Hippocampal Region in MR Neuroimages. J. Neuroim. 23(3), 473–483 (2013)
23. Casanova, R., Hsu, F.C., Espeland, M.A.: Alzheimer's Disease Neuroimaging Initiative: Classification of Structural MRI Images in Alzheimer's Disease from the Perspective of Ill-Posed Problems. PLoS One 7(10), e44877 (2012)
24. Qian, Y., Gao, X., Loomes, M., Comley, R., Barn, B., Hui, R., Tian, Z.: Content-Based Retrieval of 3D Medical Images. In: eTELEMED 2011, The Third International Conference on eHealth, Telemedicine, and Social Medicine, pp. 7–12 (2011)
25. Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B.: Within-Subject Template Estimation for Unbiased Longitudinal Image Analysis. Neuroimage 61(4), 1402–1418 (2012)

Method for Determination of the Protein Functions Based on the Global and Local Characteristics of the Structure

Georgina Mirceva

Ss. Cyril and Methodius University in Skopje,
Faculty of Computer Science and Engineering, Skopje, Macedonia
georgina.mirceva@finki.ukim.mk

Abstract. Protein molecules are very important since they take part in many processes in the organisms. Different proteins have different preferences to be involved in various processes, and these preferences determine their functions. The development of efficient and accurate computational methods for determination of the protein functions is of high importance, and therefore this research area is one of the hottest topics in bioinformatics. In this paper we present a method for functionally annotating protein structures. We consider the global characteristics of the protein structure, and also we take into consideration some local characteristics of the binding sites where the inspected protein structure get into interaction with another structure. After extracting the characteristics of the protein structure, then we induce prediction model by using the Binary Relevance method for multi-label learning. We present some experimental results of the evaluation of the method.

Keywords: Protein function prediction, protein structure, protein binding site, multi-label learning.

1 Introduction

Protein structures are one of the most important compounds in the organisms because they are involved in many essential processes in the organisms. The data gathered about the functions of the protein molecules are very important, since they could be used for designing drugs in order to regulate the processes in the organisms. Due to the importance of knowing the protein functions, various laboratory methods are applied. However, they are financially expensive, require intensive labor work and are time consuming. On the other hand, there are various methods for determining protein tertiary structures, and therefore there is a high number of known protein structures that are not functionally annotated. From this, the need for computational methods that would predict the protein functions in automated way is evident.

In the state of the art literature, a plethora of computational methods for protein function determination are proposed. Generally, the methods for determining protein functions can be grouped into four main categories. The first category contains the methods that annotate protein structures based on the annotations of their homologous structures [1]. The identification of the homologous proteins can be done by aligning

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_26

protein sequences or structures, and also by combining alignment of protein sequences and structures. Another possible approach is to determine the parts of the protein sequence or structure that are conserved and do not change during evolution, and then the annotation could be made based on analyzing these conserved parts of the proteins [2]. In this manner, the methods in the second category predict the function of the protein molecules. Instead of identifying the conserved parts of the protein molecules, the binding sites could be detected, which are the regions where the inspected structure interacts with some other structure. Therefore, the third category consists from the methods that determine protein functions by detecting the protein binding sites and analyzing their characteristics. In [3], the authors provide a wide overview of the existing tools and web servers for detecting protein's binding sites. The methods in the first three categories annotate the protein structures by analyzing the protein sequence or/and structure. The fourth category is composed of the methods that analyze the protein-protein interaction (PPI) networks [4]. The PPI networks contain data about the interacting structures that are involved in the known interactions. However, the acquisition of this knowledge in experimental manner is very expensive and time consuming, thus there is a limited knowledge gathered about the interacting pairs of protein structures.

In our research, we are focused on developing methods based on inspection of the structural information. Therefore, we concentrate on the first and the third categories of methods. In this paper, first we extract the global characteristics that contain information about the conformation of the protein tertiary structure in the three-dimensional space. Then, we extract several characteristics of the binding site for which we determine the functions that should be associated with the interaction that occurs at that place. After extraction of the global and local characteristics, then by using the Binary Relevance method for multi-label learning, we induce prediction model for determining protein functions.

The remaining of this research paper is organized in this way. The method for protein function prediction is described in Section 2. Then, in section 3 this method is evaluated and some experimental results are presented and discussed. Finally, in section 4 we give final conclusions and directions for further improvements.

2 Method for Determining Protein Functions

In this this research paper we present a method for annotating protein structures. The induction of the model made in the training phase is performed in three steps. In the first step, the global characteristics of the protein structure are extracted. Then, in the second step, for each binding site of the protein structure, several local characteristics are calculated. Finally, in the third step the prediction model is generated by using a classifier for multi-label learning. After building the model, next, the test (query) structures are annotated. For each test structure the global and local characteristics are extracted, and then according to these features and the prediction model, the functions of the query protein are determined. Figure 1 presents the training phase for generating the model for determining protein functions, as well as the testing phase for annotating the query structures.

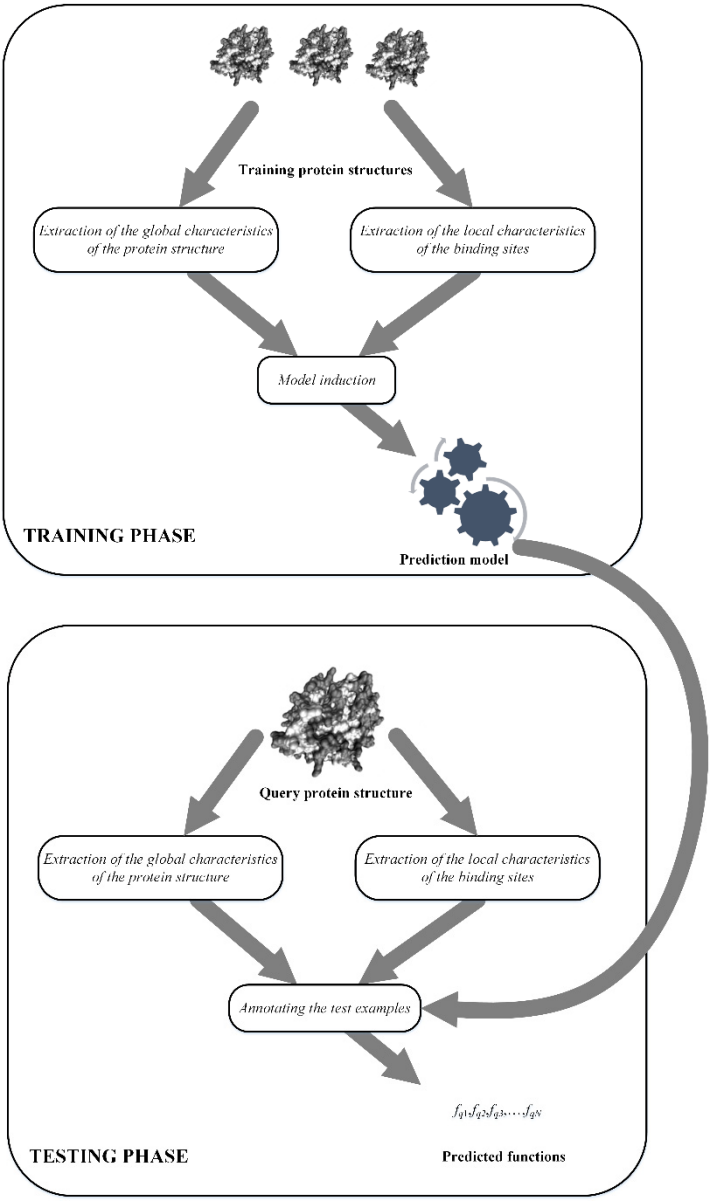

Fig. 1. The training and testing phases

2.1 Global Characteristics of the Protein Structures

In this research, as global characteristics we consider the features of the protein ray-based descriptor [5]. The algorithm for calculating the features of the descriptor is described in [5], and next we give a short description of this algorithm.

The protein molecule is composed of many different atoms, and one possible approach is to extract the features of its three-dimensional shape by taking into consideration all the atoms. However, in the literature there are many approaches that consider only the Cα atoms, and these approaches have shown as more accurate. For example in [6], various approaches for retrieving protein structures based on their global features are presented, and the experimental results made in [6] show that it is better to take into account only the Cα atoms that form the skeleton (protein backbone) of the protein structure.

First, the protein structure is scaled thus it is put into a sphere with a radius one. By this scaling, the scale invariance is provided. The idea of this descriptor is to extract features that represent how far the protein backbone is from the center of mass. One way to do this is to rise up rays from the center to some representative points of the skeleton. As representative points of the skeleton, the Cα atoms can be considered. However, the number of Cα atoms in the protein structures differs, so in this way the number of features in the descriptor will be different. To overcome this, the protein backbone is approximated with fixed number of interpolation points. In [6], two ways for backbone interpolation are presented, i.e. uniform and non-uniform interpolation, and it is shown that the uniform interpolation is more appropriate. Therefore, in this research we perform uniform interpolation of the backbone by approximating it with interpolation points that are equidistant along the protein backbone. In this research, the number of interpolation points is set to 64. After backbone interpolation, the feature extraction follows. As it was already mentioned, the features of the descriptor are the Euclidean distances from the representative (interpolation) points to the center of mass. By using such features, the exact position of the protein structure in the three-dimensional space does not have influence on the extracted features, thus the translation invariance is provided. Also, the rotation invariance is provided. Namely, if a given structure is rotated in the space, the same features will be obtained because the exact coordinates of the points are not important, but just the distances from these points to the center influence the extracted features.

2.2 Local Characteristics of the Protein Binding Sites

Besides the global characteristics, we also consider several local characteristics that describe the properties of the binding sites where an interaction occurs. The protein binding site is a region where the inspected protein structure get in contact with another structure, and it is composed of several amino acid residues. Therefore, first we calculate the characteristics of the individual residues, and then we aggregate the characteristics from all residues that constitute the inspected binding region. In this paper we extract the following amino acid residues features: Accessible Surface Area (ASA) [7], Relative Accessible Surface Area (RASA), [8], depth index (DPX) [9], protrusion index (CX) [10] and hydrophobicity [11].

ASA [7] is one of the most important characteristics of the amino acids residues because it gives evidence about the area of the residue that could be touched by another interacting structure. For that purpose, a probe sphere with some predefined radius is rolled over the surface of the protein, and the total accessible surface area is

estimated. Instead ASA, the RASA [8] characteristic could be used, and it is calculated as a ratio between the estimated ASA and the standard ASA for the corresponding amino acid. The calculation of the ASA and RASA is made by using the NACCESS program [8].

Another characteristic that is very important is the depth index (DPX) [9] that represents how far a given atom is from an atom that could be touched by the rolling sphere. The amino acid residues contain several atoms, so the depth index of a given residue is calculated as a mean value of the depth indices of its atoms.

In the literature, the protrusion index (CX) [10] is also very commonly used to represent the properties of the amino acid residues. This characteristic gives evidence about the propensity of the residue. In the process of calculating this characteristic, the surrounding around the inspected atoms is analyzed, and the ratio between the non-occupied and occupied volume is estimated by using the procedure described in [10]. Finally, the protrusion index of a given amino acid residue is calculated as an average of the protrusion indices of all atoms that form that residue.

In the bioinformatics society, it is widely known that different amino acids have different preferences to be located in different parts of the structure. These preferences are described by the hydrophobicity feature. Different research groups provide different scales for estimating the hydrophobic preferences, and in this research we use the scale presented in [11].

In this paper, our aim is to extract the characteristics of the binding regions that are composed of several residues, thus after calculating the features of the individual residues, next we aggregate these characteristics in order to estimate the characteristics of the binding site. Namely, we calculate the total, average, minimal and maximal value of a given characteristic over all residues that are part of the binding site, as well as the variance. In this way, we obtain 25 local characteristics. In order to make distinction between the larger and smaller binding sites in terms of number of constituting residues, additionally we consider the number of amino acid residues as 26-th local characteristic. The global characteristics are already in the interval [0, 1] due to the structure scaling. However, the local characteristics obtain values in different ranges. Therefore, we perform min-max normalization of the features in the interval [0, 1].

2.3 Multi-label Learning Method for Inducing Model

In the previous two sub-sections we described the first and the second steps that provide extraction of the global characteristics of the protein structure and the local characteristics of the binding sites, correspondingly. In the training phase, next, the third step follows where the prediction model is generated. As input for training the model, we provide the global and local characteristics as descriptive attributes, and the functions of the structures as predictive attributes. It is known that generally, the proteins have multiple functions since they are involved in several various interactions. Therefore, the task that we have to solve is to make a model for multi-label learning. In [12], the most commonly used methods for multi-label learning are presented. The most general categorization of these methods is to divide them into the following two groups: methods that transform the multi-learning problem into multi-class problems,

and methods that modify an existing multi-class learner for dealing with multi-label problems. In this research we use the Binary Relevance (BR) method [13] that transforms the multi-label learning problem into several binary multi-class problems. In this way, a separate problem is defined for distinguishing the examples that have same label from all remaining samples that do not have the corresponding label. In this way the number of binary problems that should be solved later is equal to the number of different labels (functions in this case).

For example, let we have four training binding sites x_1, x_2, x_3 and x_4, which have some of the functions f_1, f_2, f_3 and f_4 as presented in Table 1. As we already mentioned, the multi-label learning problem is transformed into separate binary problems for each label. For the first label f_1, we obtained a binary problem where all the samples that have this function are labeled with the positive class, while the remaining examples are considered to be in the negative class. After transforming the problem with the BR method, we obtain the transformed binary labels t_1, t_2, t_3 and t_4 presented in Table 1. We use the implementation of the BR method provided in the MULAN software [12], which is a software for multi-label learning.

Table 1. Details about the original and transformed labels of the training examples

Examples	Original labels	Transformed labels			
		t_1	t_2	t_3	t_4
x_1	$\{f_1, f_3\}$	f_1	┐f_2	f_3	┐f_4
x_2	$\{f_3, f_4\}$	┐f_1	┐f_2	f_3	f_4
x_3	$\{f_2\}$	┐f_1	f_2	┐f_3	┐f_4
x_4	$\{f_1, f_2, f_3\}$	f_1	f_2	f_3	┐f_4

After transformation of the problem, then individual models are generated for each label, so the number of models is equal to the number of different labels found in the training dataset. For this purpose, any multi-class classifier could be used to solve the individual binary problems. In this research we use the C4.5, Random Forest and Naïve Bayes methods for training the binary models.

3 Experimental Results

For evaluation of the method described in the previous section, we use a dataset obtained for the protein chains that were annotated according to the Gene Ontology (GO) on 12 July 2013. Since the number of protein chains is too large, therefore the prediction models are generated by using only the representative protein chains. This filtering is done not only for decreasing the time needed for generating the model, but also in order to reduce the redundancy since there are many similar protein structures that are obtained with some minor mutations obtained during their evolution. Therefore, in this research we filter the protein chains that have less than 100% similarity in their sequences by using the BLASTClust method, which clusters the protein chains based on the their distances determined by the BLAST method [14]. Then, the chains with less than 30%

similarity in their sequences are considered for evaluating the models, and the other chains are taken into consideration for training the models. As it was described, the method presented in this paper considers the global characteristics of the protein structures, as well as the local characteristics of their binding sites. Due to this, we can use only the chains for which we have data about the regions where the given structure binds with another structure. For that purpose, we filter the chains for which there are data about the binding regions in the Biomolecular Interaction Network Database (BIND) database [15]. Finally, we consider the chains that are annotated with the functions that are among the functions of the training chains. The number of training samples is 3167, while the number of test samples is 1449. The models that are built predict which of the 757 functions should be related with the inspected query samples.

As it was mentioned, in this research we aim to solve a multi-label learning problem, therefore we use several standard evaluation measures for estimating the prediction performances of the multi-label learning models. We use several examples based measures, as well as measures based on aggregating over the labels (functions in our case). The measures used in this research are described in [12].

The method was evaluated by using various sets of characteristics, and the results are given in Table 2, Table 3 and Table 4. It is evident that the local characteristics are not sufficient to make accurate predictions about the protein functions, while the global characteristics are much more relevant and provide better predictions. However, the best choice is to combine both global and local characteristics. We made experiments by using C4.5, Random Forest and Naïve Bayes classifiers for learning the binary models, and the results show that the models obtained by the Random Forest method generally are better than the models obtained by C4.5. Regarding Naïve Bayes, it is evident that with this classifier the recall is significantly higher, but the precision is drastically lower. It is worth to note that with the Naïve Bayes classifier, by using the global characteristics the macro measures are significantly higher than the micro measures, which means that the more specific functions are more accurately predicted than the global functions.

Table 2. The results obtained by using the global characteristics

Classifier	C4.5 Tree	Random Forest	Naïve Bayes
Precision	0.153	0.157	0.102
Recall	0.122	0.093	0.182
F_1	0.119	0.105	0.109
Accuracy	0.088	0.091	0.071
$Precision_{macro}$	0.161	0.679	0.097
$Recall_{macro}$	0.142	0.105	0.193
F_{1macro}	0.151	0.182	0.129
$Precision_{micro}$	0.083	0.224	0.146
$Recall_{micro}$	0.091	0.099	0.139
F_{1micro}	0.076	0.119	0.120
$AUC\text{-}ROC_{macro}$	0.660	0.716	0.837
$AUC\text{-}ROC_{micro}$	0.540	0.640	0.650

Table 3. The results obtained by using the local characteristics

Classifier	C4.5 Tree	Random Forest	Naïve Bayes
Precision	0.085	0.067	0.017
Recall	0.034	0.023	0.307
F_1	0.041	0.030	0.032
Accuracy	0.028	0.021	0.017
$Precision_{macro}$	0.156	0.175	0.016
$Recall_{macro}$	0.033	0.025	0.321
F_{1macro}	0.054	0.043	0.031
$Precision_{micro}$	0.011	0.024	0.013
$Recall_{micro}$	0.004	0.009	0.238
F_{1micro}	0.005	0.012	0.020
$AUC\text{-}ROC_{macro}$	0.771	0.656	0.756
$AUC\text{-}ROC_{micro}$	0.504	0.541	0.619

Table 4. The results obtained by using the global and local characteristics

Classifier	C4.5 Tree	Random Forest	Naïve Bayes
Precision	0.151	0.146	0.074
Recall	0.128	0.082	0.261
F_1	0.122	0.094	0.103
Accuracy	0.089	0.080	0.061
$Precision_{macro}$	0.160	0.688	0.073
$Recall_{macro}$	0.139	0.091	0.275
F_{1macro}	0.149	0.160	0.116
$Precision_{micro}$	0.078	0.181	0.120
$Recall_{micro}$	0.085	0.078	0.162
F_{1micro}	0.070	0.095	0.109
$AUC\text{-}ROC_{macro}$	0.665	0.721	0.832
$AUC\text{-}ROC_{micro}$	0.538	0.631	0.649

4 Conclusion

We presented a method for determining protein functions that take into consideration some global characteristics of the protein structure and several local characteristics of their binding sites. The induction of the prediction models is done by using the Binary Relevance method, which is one of the most basic methods for multi-label learning. The evaluation results demonstrate that the global characteristics are much more informative for making more accurate predictions about the proteins' functions. Regarding the classifier used for learning the binary models, it was shown that the Random

Forest classifier is more appropriate than C4.5, while the Naïve Bayes classifier makes more accurate predictions for the specific functions rather than the general functions.

Besides the characteristics used in this research, also some other characteristics could be included in the dataset in order to improve the models. Regarding the method used for learning the models, instead of the Binary Relevance method, we can also apply some other methods for multi-label learning.

Acknowledgments. This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University in Skopje", Skopje, R. Macedonia.

References

1. Todd, A.E., Orengo, C.A., Thornton, J.M.: Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. 307(4), 1113–1143 (2001)
2. Panchenko, A.R., Kondrashov, F., Bryant, S.: Prediction of functional sites by analysis of sequence and structure conservation. Protein Science 13(4), 884–892 (2004)
3. Tuncbag, N., Kar, G., Keskin, O., Gursoy, A., Nussinov, R.: A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. Briefings in Bioinformatics 10(3), 217–232 (2009)
4. Kirac, M., Ozsoyoglul, G., Yang, J.: Annotating proteins by mining protein interaction networks. Bioinformatics 22(14), e260–e270 (2006)
5. Mirceva, G., Kalajdziski, S., Trivodaliev, K., Davcev, D.: Comparative Analysis of three efficient approaches for retrieving protein 3D structures. In: 4th IEEE Cairo International Biomedical Engineering Conference (CIBEC 2008), pp. 1–4 (2008)
6. Mirceva, G., Cingovska, I., Dimov, Z., Davcev, D.: Efficient Approaches for Retrieving Protein Tertiary Structures. IEEE/ACM Trans. Comput. Biol. Bioinf. 9(4), 1166–1179 (2012)
7. Shrake, A., Rupley, J.A.: Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J. Mol. Biol. 79(2), 351–371 (1973)
8. Hubbard, S.J., Thornton, J.M.: NACCESS, Computer Program. Department of Biochemistry and Molecular Biology, University College London, London, UK (1993)
9. Pintar, A., Carugo, O., Pongor, S.: DPX: for the analysis of the protein core. Bioinformatics 19(2), 313–314 (2003)
10. Pintar, A., Carugo, O., Pongor, S.: CX, an algorithm that identifies protruding atoms in proteins. Bioinformatics 18(7), 980–984 (2002)
11. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157(1), 105–132 (1982)
12. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, 2nd edn., pp. 667–685. Springer (2010)
13. Luaces, O., Díez, J., Barranquero, J., José del Coz, J., Bahamonde, A.: Binary relevance efficacy for multilabel classification. Progress in Artificial Intelligence 1(4), 303–313 (2012)
14. Altschul, S., Gish, W., Miller, W., Myers, E.W., Lipman, D.: Basic local alignment search tool. J. Mol. Biol. 215(3), 403–410 (1990)
15. Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T., Hogue, C.W.: BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 29(1), 242–245 (2001)

Cooperation among Non-identical Oscillators Connected in Different Topologies

Miroslav Mirchev[1], Lasko Basnarkov[1], and Ljupco Kocarev[1,2,3]

[1] Facuty of Computer Science and Engineering, Ss. Cyril and Methodius University, 1000 Skopje, Macedonia
{miroslav.mirchev,lasko.basnarkov}@finki.ukim.mk
[2] Macedonian Academy of Sciences and Arts, 1000 Skopje, Macedonia
[3] BioCircuits Institute, University of California, San Diego, La Jolla, CA 92093 USA
lkocarev@ucsd.edu

Abstract. Various forms of oscillatory networks exist in our surrounding from neural cells to laser arrays. In many of these networks the nodes can go through a transient process of interaction and start oscillating in synchrony. Each of these nodes is characterized by its internal dynamics and changes its state accordingly. Using several forms of interactions, we numerically examine how the network dynamics is affected by network topology and potential random disturbances.

Keywords: Complex networks, Oscillators, Synchronization.

1 Introduction

Our environment is full of various networks of entities exhibiting periodic oscillations, like interconnected neural and heart cells in biology, wireless sensor networks, laser arrays and electronic oscillators in engineering, and so on. The nodes in these networks interact and after some time can agree to oscillate synchronously as studied in [15] and [16]. The synchronization could be desirable as in sensor networks and laser arrays, or undesirable as in epilepsy seizures in the brain. In all these networks, the nodes are characterized by states and if the network interactions result in all the nodes reaching the same state, it is said that the network is completely synchronized. On the other hand, in case of networks with weak interactions the nodes can agree on an equal frequency of oscillation without exact match of their individual states (amplitude values). This type of networks have been widely studied in the literature and many models for representation are known with the most famous example being the Kuramoto model [9] given as

$$\dot{\theta}_i = \omega_i + \frac{\gamma}{N} \sum_{j=1}^{N} \sin(\theta_j - \theta_i), \tag{1}$$

where θ_i is the phase of oscillation of each node of a population of N nodes, ω_i is the node's internal oscillating frequency, with which it would oscillate if

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_27

isolated, and γ is a general coupling strength typically larger then zero. Despite the differences in the internal frequencies a synchronous mode of oscillation is possible in this kind of networks.

In a previous paper [11] we have studied cooperation in networks of non-identical oscillators, particularly the case of non-identical interactions. In that paper convergence criteria toward frequency synchronization were provided for a specific type of nonnegative and symmetrical interactions. Moreover, the behavior of other types of systems were also examined, like asymmetric connections, external fields and frustration due to random disturbances. In this paper we further study these type of networks, particularly focusing on the effects of random disturbances with different forms of coupling functions and the effects of network topology on the dynamical behavior. All these issues have been part of a wider study of complex networks with imperfections in [10].

The paper continues with Section 2 where we introduce the notation of networks of non-identical oscillators. We numerically study the phenomenon of frequency synchronization, focusing in Section 3 on the effects of random disturbances and examining in Section 4 the network topology effects, while Section 5 provides some conclusions.

2 Networks of Non-identical Oscillating Nodes

We consider networks composed of N oscillating nodes whose dynamics can be represented by

$$\dot{x}_i = \omega_i + \gamma \sum_{j=1}^{N} a_{ij} f_{ij}(x_j - x_i), \tag{2}$$

where as in the Kuramoto model the phases lay on a unit circle S^1 ($x_i \in S^1$) and $x_i \in [0, 2\pi)$, the natural frequencies are $\omega_i \in \mathbb{R}$ and γ is a general coupling strength, while $\mathbf{A} = [a_{ij}]$, ($a_{ij} \geq 0$, $a_{ii} = 0$, $a_{ij} = a_{ji}$, $\forall i, j$) is an adjacency matrix representing the network's topology. The coupling functions f_{ij} are taken to be 2π-periodic and $f_{ij}(0) = 0$, $\forall i, j$.

One measure of network coherence is how close are the phases and to evaluate and visualize this type of network coherence an order parameter [9] can be used

$$re^{\mathrm{i}\Psi} = \frac{1}{N} \sum_{j=1}^{N} e^{\mathrm{i}x_j}, \tag{3}$$

where larger values mean larger coherence ($r \in [0, 1]$), while Ψ is the average phase of the population.

Another type of coherence is the network's synchrony and we use the following error function to evaluate it

$$e_\Omega(t) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\dot{x}_i(t) - \Omega)^2}, \tag{4}$$

which indicates how the oscillators' velocities are approaching the mean natural frequency $\Omega = (1/N)\sum_i \omega_i$. Sometimes the analysis require $< e_\Omega >$, the time average of $e_\Omega(t)$, which is calculated with excluded transient dynamics.

In our analysis we consider networks with random uniformly distributed initial phases and random natural frequencies following a triangular distribution in the range $[\omega_{min}, \omega_{max}] = [-0.5, 0.5]$ with a probabilistic density function's peak at $\omega_0 = 0$.

Typically the coupling functions are taken to be sinusoidal, as in the Kuramoto model [9]

$$f_{ij}(x_j - x_i) = \sin(x_j - x_i). \tag{5}$$

However, other types of coupling functions should also be analyzed as in reality the interactions are not exactly sinusoidal [4]. One simple case are linear coupling functions that are periodically repeated in the following way

$$f_{ij}(x_j - x_i) = (x_j - x_i - 2\pi k), \text{ for } -\pi + 2k\pi < x_j - x_i < \pi + 2k\pi, \tag{6}$$

for $k = 0, \pm 1, \pm 2, \ldots$. Another case that we study are periodically repeated cubic coupling function of the form

$$f_{ij}(x_j - x_i) = (x_j - x_i - 2\pi k)^3, \text{ for } -\pi + 2k\pi < x_j - x_i < \pi + 2k\pi, \tag{7}$$

for $k = 0, \pm 1, \pm 2, \ldots$.

In our study we numerically simulate networks consisting of $N = 100$ oscillating nodes. In Section 3 we use fully connected networks where $a_{ij} = 1, \forall i, j, i \neq j$ and $a_{ii} = 0, \forall i$, while in Section 4 we examine the effects of the network topology so not all adjacency elements have a value of one. The numerical integration of the equations of motion of the oscillators is performed using a fourth-order Runge-Kutta method with a fixed step $\Delta t = 0.001$.

3 Random Disturbances Effects

In reality the interactions among the nodes are prone to some environmental or internal random disturbances also called frustrations. In our paper [11] it was analytically and numerically shown that besides these frustrations, sinusoidally coupled oscillators can eventually agree on a common oscillation frequency, as previously observed in [5].

The random disturbances can be introduced by including elements $\phi_{ij} \in (-\pi/2, \pi/2)$, $\phi_{ij} \in \mathbb{R}$, $\forall i, j$ [14], thus the nodes dynamics takes the form

$$\dot{x}_i = \omega_i + \gamma \sum_{j=1}^{N} a_{ij} f_{ij}(x_j - x_i + \phi_{ij}). \tag{8}$$

As discussed in [11], if $\phi_{ji} = -\phi_{ij}$ and $\phi_{ii} = 0$ for all interactions, the oscillators can achieve frequency synchronization as the synchronized state is stable.

In this section we provide numerical results of the dynamics of systems of the form defined by (8) and their idealistic counter pairs as given by (2). The

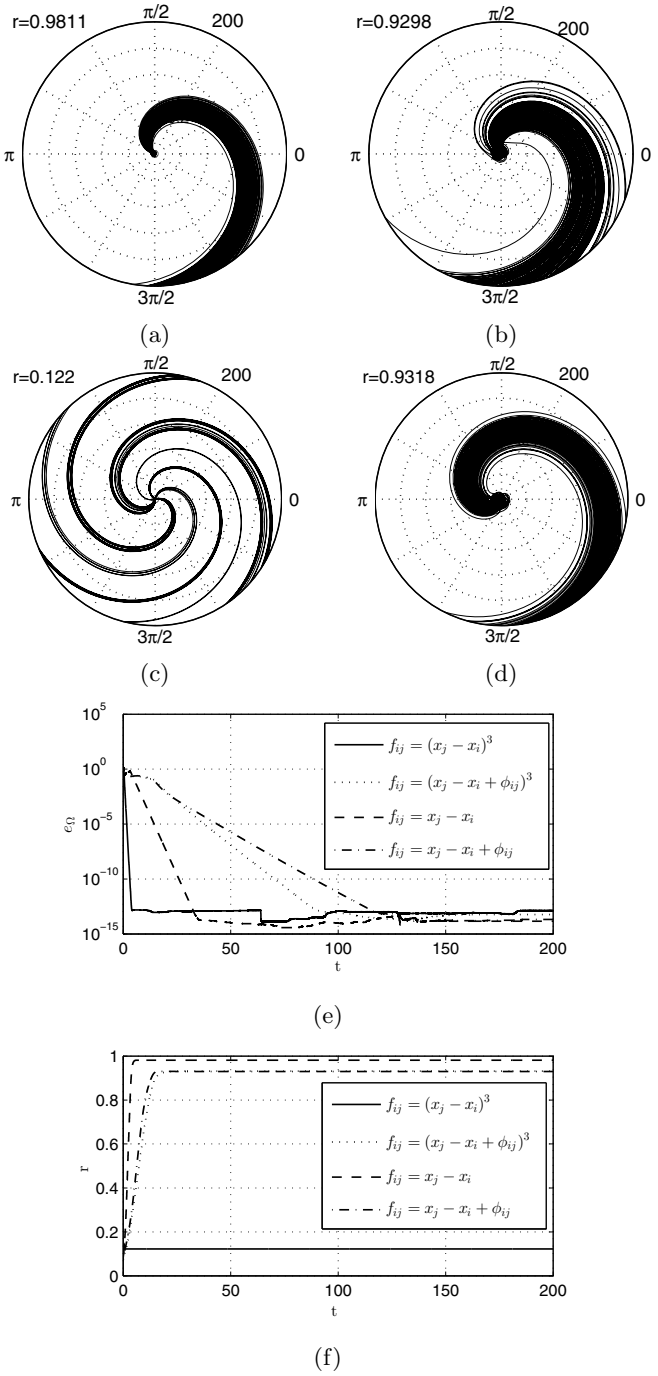

Fig. 1. Phase evolution in $t \in [0, 200]$ in a fully connected network with $\gamma = 0.01$ and (a) $f_{ij} = (x_j - x_i)$; (b) frustrated $f_{ij} = ((x_j - x_i) + \phi_{ij})$; (c) $f_{ij} = (x_j - x_i)^3$ and (d) frustrated $f_{ij} = ((x_j - x_i) + \phi_{ij})^3$; (e) synchronization error e_Ω and (f) coherence r.

network is taken to be fully connected in order to separate the disturbances effects from the topology effects that are considered in the following section. The general coupling strength is chosen to be $\gamma = 0.01$, which is big enough to allow network synchronization but not too large.

The first four sub-figures of Fig. 1 visualize the time evolution of the oscillators' phases in several different coupled networks. A polar coordinate system is used where the time flow is shown on the radial coordinate and the phases are on the angular coordinate. Thus, from the center of the circle toward the periphery the time t increases from 0 to 200 and the phase evolution of each oscillator is represented with a single continuous line. Fig. 1a shows the phase evolution in a linearly coupled network without frustration, where although the phases are not exactly matched ($r = 0.9811$) synchronization is achieved and the phases evolve at an equal rate. On the other hand, in Fig. 1b the linear interactions in the network are frustrated, which introduces a larger phase dispersion ($r = 0.9298$), though still allowing frequency synchronization among the oscillators. We also consider cubic coupling functions, first without frustration in Fig. 1c. This cubic coupling introduces clustered syncrhony that makes the order parameter low $r = 0.122$, as also observed previously in [11]. The introduction of the random disturbances in the cubic coupling in Fig. 1d prevents the network clustering and increases the coherence ($r = 0.9318$), while still allowing the oscillators to rotate their phases at an equal rate.

In Fig. 1e is shown how the synchronization error e_Ω reduces for the different types of coupling functions. As expected, the error reduces more rapidly with cubic coupling than with linear coupling, while the random disturbances though still allowing synchronization significantly slow down the convergence rate. The evolution of the order parameter r is shown in Fig. 1f and similar conclusions can be drawn as from the previous sub-figures. With linear coupling the frustration reduces the coherence, while with cubic coupling the coherence is increased in the presence of random disturbances due to the avoided clustering. The convergence rate of the order parameter is not as influenced as the synchronization error.

4 Topological Effects

In the networks studied in the previous section the oscillators were fully connected, however, in reality network interactions follow certain patterns that form the network topology. In this section, we study network dynamics in random networks generated using the Erdös-Rényi (ER) model [6] and scale-free networks generated using the Barabási-Albert (BA) model [1]. In all cases a care should be taken that the generated network is connected, i.e. there is a path among all node pairs, as otherwise synchronization is not achievable. Some analyses of the topology effects on the synchronization properties in a Kuramoto model with scale-free topology with standard sinusoidal couplings were done in [12, 8], and [13], using numerical simulations and different analytical approaches. However, a consensus have not been reached for the critical coupling gain at which synchronization occurs as also noted in [2]. In [7] and [3] both scale-free (BA)

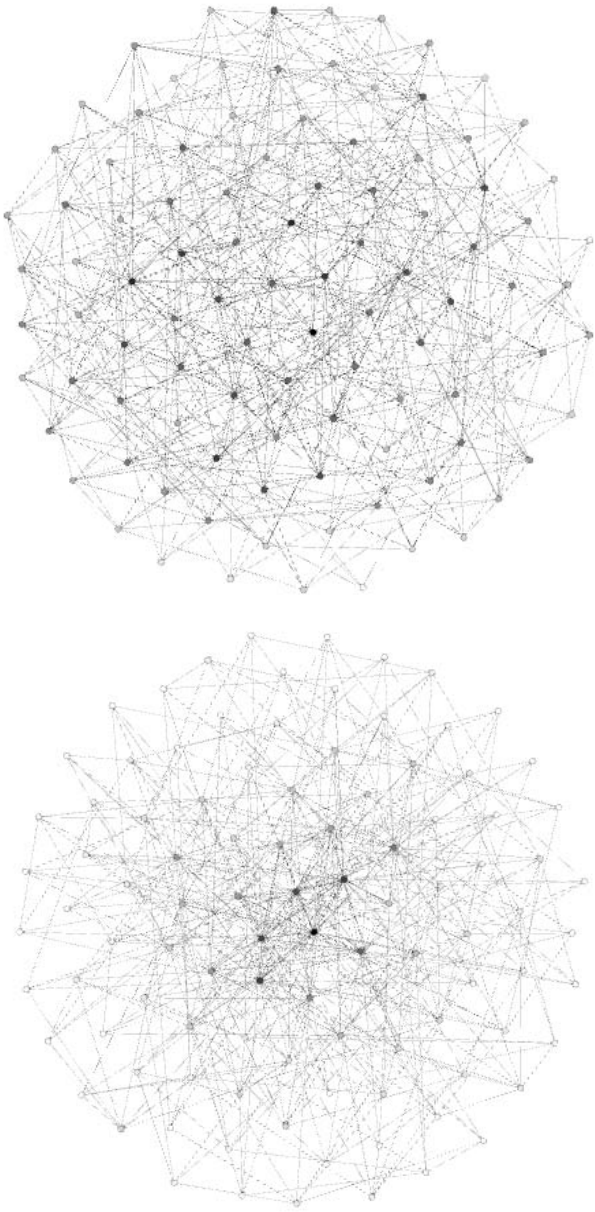

Fig. 2. Two example network topologies where the nodes with higher degree are colored darker and placed more centrally: (top) random network – Erdös and Rényi and (bottom) scale-free network – Barabási-Albert. Visualization is done in Gephi using the Fruchterman–Reingold algorithm.

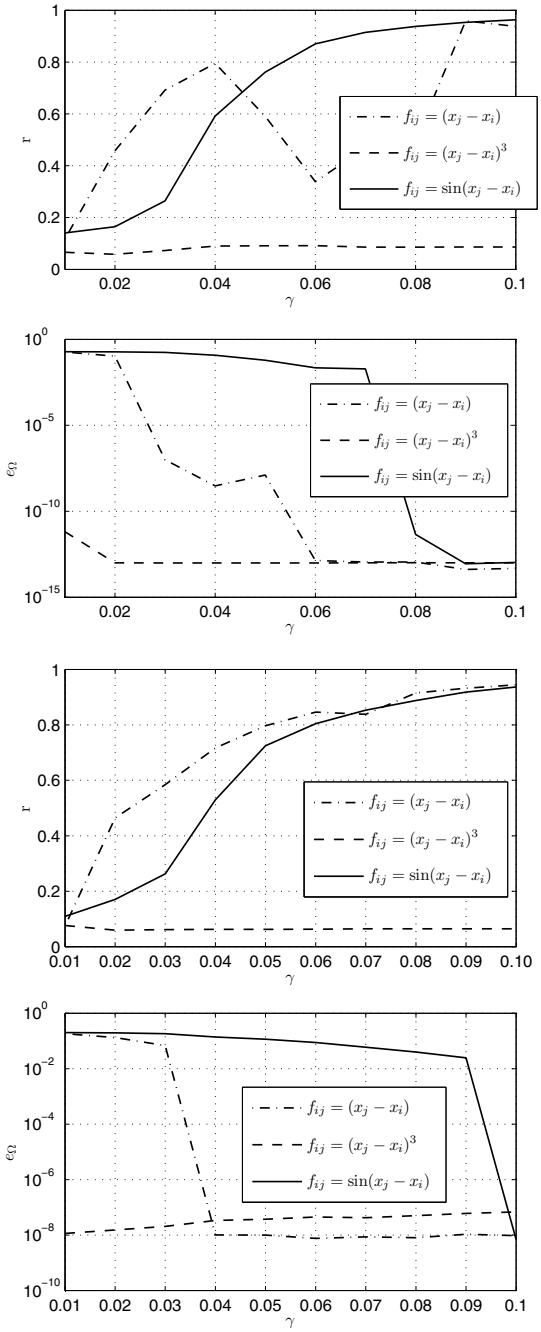

Fig. 3. The coherence r and the synchronization error e_Ω in networks of $N = 100$ nodes and 500 links generated using: (top two) the ER model, and (bottom two) the BA model for different coupling functions.

networks and random (ER) networks are examined in which the oscillator's natural frequencies are correlated to their degree of connectivity. In our study we do not assume any correlation among the degrees of the nodes and their internal frequencies.

The model for random network generation developed by Erdös and Rényi (ER) creates a graph $G(N, p)$ consisting of N nodes, where each of the possible links among the nodes exist with a probability p. For lower values of p this model can produce disconnected network parts, hence, if we need a connected network the connectivity should be checked and if the network is not connected the whole procedure could be repeated. Here we generate networks of $N = 100$ oscillators with link probability $p = 0.1$, which results in a network of about 500 links.

The Barabási-Albert model can be used for creating scale-free networks. The model requires an initial seed network to which gradually new nodes are added with L_N connections per node. This procedure, also known as preferential attachment, resembles a well known phenomenon where "the rich get richer", present for example in genetic networks, the World Wide Web, the Internet, social networks, etc. To assure that the networks with different topologies are comparable, the number of nodes and links is kept the same, so gradually to the seed network new nodes with $L_N = 5$ connections are added until we reach $N = 100$ nodes and 500 links.

Example networks with these topologies are given in Fig. 2. The comparison of the results in Fig. 3 show that in this case ER networks synchronize more easily than BA networks, because for both linear and sinusoidal coupling synchronization occur for lower values of γ. The coherence r is similar for both types of networks, and with the increase of the coupling strength r rises just slightly quicker in ER networks. It can be noticed by looking at the coherence r that for linear coupling clustering occurred in the random network for some values of γ, while in the scale-free network a step-wise increase was observed at some points due to the hierarchical structure. Similar results were reported in [3] for sinusoidally coupled networks in which node's degrees and natural frequencies are correlated.

5 Conclusion

This paper provides additional examination of the process of cooperation in networks of non-identical oscillators through several forms of interaction. Particularly, the focus is on the effects of possible random disturbances and the network topology on the network dynamics. Possible achievement of frequency synchronization was confirmed for linear and cubic coupling functions, in addition to the known results with the standard sinusoidal function. The study of the topology effects showed that frequency synchronization can be achieved more easily in random networks than in scale-free networks, although, typically complete synchronization happens more easily in scale-free networks.

Acknowledgements. The authors thank the Faculty of computer science and engineering at the Ss. Cyril and Methodius University in Skopje, under the DYSENE (Dynamical sensor networks) project for financial support.

References

1. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (1999)
2. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U.: Complex networks: Structure and dynamics. Phys. Rep. 424(4-5), 175–308 (2006)
3. Coutinho, B.C., Goltsev, A.V., Dorogovtsev, S.N., Mendes, J.F.F.: Kuramoto model with frequency-degree correlations on complex networks. Phys. Rev. E 8(3), 32106 (2013)
4. Daido, H.: Order function and macroscopic mutual entrainment in uniformly coupled limit-cycle oscillators. Prog. Theor. Phys. 88(6), 1213–1218 (1992)
5. Daido, H.: Quasientrainment and slow relaxation in a population of oscillators with random and frustrated interactions. Phys. Rev. Lett. 68(7), 1073–1076 (1992)
6. Erdös, P., Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hungar. Acad. Sci. 5, 17–61 (1960)
7. Gómez-Gardeñes, J., Gómez, S., Arenas, A., Moreno, Y.: Explosive synchronization transitions in scale-free networks. Phys. Rev. Lett. 106(12), 128701 (2011)
8. Ichonmiya, T.: Frequency synchronization in a random oscillator network. Phys. Rev. E 70(2), 26116 (2004)
9. Kuramoto, Y.: Chemical Oscillations, Waves, and Turbulence. Springer, Berlin (1984)
10. Mirchev, M.: Cooperative processes in complex networks with imperfections. Ph.D. thesis, Politecnico di Torino, Italy (2014)
11. Mirchev, M., Basnarkov, L., Corinto, F., Kocarev, L.: Cooperative phenomena in networks of oscillators with non-identical interactions and dynamics. IEEE Trans. Circuits Syst. I, Reg. Papers 61(3), 811–819 (2014)
12. Moreno, Y., Pacheco, A.F.: Synchronization of kuramoto oscillators in scale-free networks. Europhys. Lett. 68(4), 603–609 (2004)
13. Restrepo, J.G., Ott, E., Hunt, B.R.: Onset of synchronization in large networks of coupled oscillators. Phys. Rev. E 71(3), 036151 (2005)
14. Sakaguchi, H., Kuramoto, Y.: A soluble active rotator model showing phase transitions via mutual entrainment. Prog. Theor. Phys. 76(3), 576–581 (1986)
15. Wiener, N.: Nonlinear Problems in Random Theory. MIT Press, Cambridge (1958)
16. Winfree, A.T.: Biological rhythms and the behavior of populations of coupled oscillators. J. Theor. Biol. 16(1), 15–42 (1967)

Sentiment Analysis of Movie Reviews Written in Macedonian Language

Vasilija Uzunova and Andrea Kulakov

Computer Science and Engineering Department
University Sts. Cyril and Methodius
Skopje, Macedonia
vasilijauzunova@gmail.com, andrea.kulakov@finki.ukim.mk
http://www.finki.ukim.mk

Abstract. Identifying sentiments is a natural language processing problem that became popular lately with the advent of various forums and social networks on the Internet. In this paper the analysis will focus on opinions that can evoke either positive or negative feelings in people. Most of the existing researches on textual information processing focus on data mining and fact analysis such as information retrieval, web search, text classification, clustering and many other types of natural language processing, unlike opinion analysis, especially for texts written in Macedonian.

Keywords: Sentiment analysis, movie reviews, Macedonian, Naive Bayes.

1 Introduction

Plenty of time before the expansion of the Internet people asked for an advice from friends for various products or services. Today a source of such information is the Internet because it dramatically changed the way people convey their thoughts and opinions. They can now post reviews of products and express their views on almost every topic on the Internet through forums, discussion groups or blogs that are based on user generated content. There is a huge potential for processing such data that can provide additional value for both users and the companies that make public opinion analysis about any product or service [6]. However, finding sources of opinions and monitoring them can still be a difficult task because there are a number of different sources, and each source can also have a huge amount of reviews. That's why it is required to have a system that summarizes all these posts. This can save a lot of resources and time in search of this information online. In this paper, we will propose a solution for this problem for sentiment analysis of film reviews written in Macedonian, using a Naive Bayes classifier.

As a beginning of this analysis, we introduce the notion of sentiment polarity. Suppose we want to classify a given comment text whether it is positive or negative, based on the opinion of the author. Is it going to be an easy task? In

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_28

response to the question, we will take an example that consists of one sentence: "Човекот X Y беше многу нервозен поради лошиот проект" ("Person X Y was very nervous because of the bad project"). The theme of this segment can be identified with the phrase "X Y", but the presence of the words "нервозен" ("nervous") and лош ("bad") suggest a negative meaning. One would suppose that this task is really easy, and the polarity of opinions generally can be distinguished by a set of words.

However, the results of an analysis made by Pang and Lee for movie reviews point out that the suggestions coming through a set of keywords can be less trivial apart from the originally thought [9]. In order to get those keywords, opinion is taken by two human subjects to question whether what they think is positive or negative.

The main goal of our experiments is to confirm that the incorporation of some additional word processing into the polarity classification can significantly improve the results. In this paper, we first shortly describe some related work in the field of sentiment analysis and classification. Section 3 explains the most important challenges of making this analysis. Section 6 deals with sentiment analysis and text classification. In this section we talk about the use of a classifier and the data preprocessing as an important step in the sentiment analysis process. Finally we provide the obtained results using different machine learning algorithms and end up by reaching some conclusions, discussions and suggestions for further development.

2 Related Work

Sentiment analysis is currently receiving a lot of attention from the research community. Since 2001 till now there was rapid expansion and several papers along the subject because of the outstanding research and commercial potential [8]. The focus on this area is to solve the problem of computer processing of messages, sentiment and subjectivity in text.

Opinion mining is part of the area near to Web search and information retrieval. The opinion mining tool processes a set of search results for a given term or product, generates a number of product attributes (quality, features) and aggregates the opinions for each attribute e.g. bad, good [8].

Sentiment is a term used for the automatic evaluation of text and track predictions. Number of papers have placed their focus on sentiment analysis. Subjectivity analysis is a term also used for classification, in which documents are classified into two classes by their objectivity or subjectivity [12]. Thus, "sentiment analysis" and "opinion mining" can be considered as sub-areas of subjectivity analysis.

The rest of this section surveys previous work in sentiment analysis classification. In [7], the authors have focused on defining the polarity on Twitter posts by extracting a vector of weighted nodes from the graph of WordNet. For a supervised polarity they build a labeled corpus of tweets written in English. Therefore, they used positive and negative emoticons to label the tweets, ":)"

returns tweets with positive smileys, and ":(" with negative. They used total number of 376,296 tweets, and the reported accuracy level of this approach was 62%. Another interesting approach is presented by Turney, 2002. This paper presents a simple unsupervised algorithm for classifying text using sentiment orientation of the phrases [11]. The algorithm used different review domains and it achieved an average accuracy of 74%. In [9], the authors Pang, Lee and Vaithyanathan have used three machine learning methods (Naive Bayes, maximum entropy classification and support vector machines) for classifying reviews. As a data source they used the Internet Movie Database (IMDb) archive, and the reviews were collected by stars or some numerical value. The achieved results were very good using all methods. The Naive Bayes approach had a high classification rate of 82.9%.

3 Challenges

What other people think is always important information during the decision making process. Thus, this is the most important challenge for sentiment analysis. Another key challenges are:

1. *Named Entity Recognition* - it is an important stage for sentiment analysis, it is locating and classifying atomic elements in text into predefined categories [2]. (What is the person actually talking about in the sentence?)
2. *Sarcasm/Irony* - a statement with a certain structure, which, actually means the opposite of what that particular statement really means. Sarcasm could be wrongly interpreted as a positive sentiment.
3. *Metaphor* - it can be a replacement of the meaning of a word with another meaning.
4. *Language complexity, spelling and slang words*

4 Classification Based on Supervised Learning

Since sentimental analysis is a special case of text classification, we use algorithms that are used for classification. The classification is solved using supervised algorithms for machine learning.

In supervised machine learning there are training data on which the algorithm learns how to act when it gets new data, and how to classify it. All algorithms have two phases. The beginning phase is the learning phase, in which the algorithm learns how to classify the information. The second stage is the prediction. At this stage, the algorithm gets new, unfamiliar text and based on the training data and some other text analysis, predicts which class should be assigned. Sentiment analysis is a problem that can be solved by supervised learning with two classes, positive and negative. There are several algorithms that can be used for this analysis, of which the most common are Naive Bayes, Support Vector Machines (SVM), Entropy Classification etc [6].

5 Naive Bayes Classifier

Training any classifier requires labeled training examples and a model that will fit. In this paper, we describe a sentiment analysis solution using Naive Bayes classifier. The Naive Bayes classifier is a simple probabilistic classifier. A more descriptive term for this model would be "independent functional model". Under normal conditions, the Naive Bayes classifier assumes that the presence (or absence) of certain characteristic of a class is unrelated to the presence (or absence) of any other feature, so in this case the probability of a word appearing in the document does not affect the probability of another word. Probability of occurrence of words wi in class cj is equal to the frequency of appearance of word wi in class cj divided by the total numbers of words in class cj.

$$p(wi|c) = \frac{number\,of\,times\,wi\,occures\,in\,c}{total\,number\,of\,words\,in\,class\,c} \tag{1}$$

$$p(ci) = \frac{training\,documents\,in\,class\,ci}{total\,number\,of\,training\,documents} \tag{2}$$

$$p(ci) = \frac{Ni}{N} \tag{3}$$

If after the training the test data shows up a new word, which has not appeared before in the training data, it will result with a problem for calculating the probability. In this case, the cumulative probability is equal to 0. This problem is solved with Laplace smoothing. Laplace alignment introduces the assumption that a new word has appeared in the training set once. In order not to disrupt the possibility with this kind of assumption, the number of occurrences of all words must be increased by 1. The formula to calculate the probability of words in a class is given by [5]:

$$p(wi|cj) = \frac{1 + count(wi, cj)}{|V| + Ni} \tag{4}$$

In this way, all the words will have a certain probability greater than 0. Also, the probability of certain words will be reduced, however, their relationship is still going to remain the same. In order to achieve higher accuracy of the algorithm, it is necessary to introduce some additional processing. The first processing task that improves the results of sentiment analysis and many other language processing tasks is stemming [10]. It removes the suffixes in words in order that the words with same meaning and different inflection were treated as one. In our analysis, the stemming is avoided. The second processing is handling negation. This is an important concern in sentiment-related analysis [8].

6 Experimental Analysis

The application for sentiment analysis is developed using Groovy programming language powered by the Grails framework. The application consists of a service

that is doing the sentiment analysis for specified text. The learning algorithm as its warehouse uses a MySQL database in which are written and updated all the statistics for words obtained in the learning phase using Naive Bayes algorithm. The main table in the database is the table *Word* that contains information about the number of occurrences of all the words in particular class. The table contains the word, the class (positive/negative), the number of occurrences of a word in a given class (both positive and negative) and probability of that word to be in particular class.

The keywords, the important words that determine sentiment significantly, differ for positive and negative sentiment. These words are saved in two different tables, table *PositiveWord* for all the words that have a positive meaning, and table *NegativeWord* for all the words that have a negative meaning. Words that are not crucial to determine the sentiment in the test phase are stored in another table as neutral words - the table *NeutralWord*. Positive and negative words are obtained by conducting a survey of few people. Some of the words are shown in the table below.

Table 1. Word list

Positive	добар, одличен, убав, среќен, фантастичен, еуфоричен, живописен
Negative	лош, грозен, одвратен, вулгарен, грд, груб

We are using public movie review data from three Macedonian forums[1]. The data consists of 200 positive and 200 negative reviews, and they are divided in two categories, positive and negative. Neutral reviews are not included. The data is stored in two directories in the file system. The total number of words from these reviews is 13617 and this is the vocabulary |V| in equation (4). In order to enroll all the words with number of occurrences and calculated probability in the corresponding tables, first we are testing the Naive Bayes classifier on this dataset.

6.1 Handling Negation

The representation of these two sentences "Ми се допадна овој филм" ("I like this movie ") and "Не ми се допадна овој филм" ("I don't like this movie") are considered to be very similar, but in fact they have opposite meaning. The only different word is the negation term. To solve this problem we created simple algorithm that analyzes the words to the first punctuation mark. If the word "не" ("no") appears in the sentence, it means that the sentence will have opposite meaning, so we take the smaller value of possibility for all the words to the first

[1] `http://forum.femina.mk/filmovi/`
`http://forum.kajgana.com/forums/`
`http://forum.crnobelo.com/forums`

punctuation mark. This is not a perfect model, since there are negations that are not related with the text to the first punctuation mark, but it is good enough for this analysis.

Words that have no meaning and often emerge in the text have to be removed. Such words are called stop-terms. In our analysis, if there are such words in the text, they are saved in a different table for neutral words, e.g. conjunctions, exclamations, particles. . Such words in Macedonian language are: " без, во, врз, за, зад, и, кај, каде, бидејќи, ... " (" without, in, on top of, for, behind, and, where, when, because, ... "). These words are removed from the classification because we want the processed text to contain only the substantive words that have meaning for the sentiment analysis.

6.2 Spelling

Because of the fact that all the words in the database are in Cyrillic, and we want also to process text written in Latin, which is common on Internet forums, we use transliteration rules for mapping Cyrillic letters to Latin.

The spelling of the Macedonian language is phonetically based, which means that almost every word is written exactly as it is pronounced. Yet, on Internet forums, there are no clear rules for transliterating Cyrillic letters to Latin, but rather everybody uses different rules, even they can be changed in the same text. Like to write "sh" for "ш" and then later to write only "s" for the same Cyrillic letter. This created difficulties for identifying individual words and we have solved it by using two-steps transliteration, one of the most common and simple way of transliteration using the following table:

Table 2. Transliteration rules

Latin	Cyrillic	Transliteration process	
Gj gj	Ѓ ѓ	gjavol	ѓавол
Zh zh	Ж ж	zhivot	живот
Dz dz	S s	dzid	ѕид
Lj lj	Љ љ	ljubov	љубов
Nj nj	Њ њ	konj	коњ
Kj kj	Ќ ќ	kjumur	ќумур
Ch ch	Ч ч	chovek	човек
Dj dj	Џ џ	djudje	џуџе
Sh sh	Ш ш	shal	шал

Every word is preprocessed and in the process of transliteration each letter is mapped respectively, giving priority to the letters from the table above. Let's consider the word "sushtina" which should be mapped to "сушшина", in the first step of transliteration the letters "s" and "h" - "sh" are mapped into "ш" and the word becomes "suшtina", in the second step, all the other letters are mapped respectively, so the word finally becomes "сушшина".

7 Results with Different Approaches

7.1 Results with k-Fold Cross Validation

The k-fold cross validation is a process used for model selection and error estimation of classifiers [1]. In this analysis, we are using a 10 fold cross validation, so we are dividing the data into 10 sections, then train and test the classifier 10 times, each time choosing a different section as the test set and the other 9 sections as the training set. Since we have data for two different classes, we split the data from each class into 10 subsets.

With given 200 examples of each class we have successive blocks of 20 files that are used for cross-validation as test data. The results obtained in this analysis are shown in the table below.

Table 3. Results for every fold

Fold	Accuracy
0	0.475
1	0.875
2	0.975
3	1.0
4	1.0
5	1.0
6	1.0
7	1.0
8	1.0
9	0.925

The average Accuracy is:

$$Accuracy = 0.925 \tag{5}$$

7.2 Results with New Test Data from the Web Interface

All the words from the dataset are listed in the table Word with positive and negative occurrences for each class. When we enter new test data from the web interface, we are analyzing each word from a sentence and calculate the possibility whether it is positive or negative. In this case, we are also using a Naive Bayes classifier.

We are testing 30 new positive and 30 new negative movie reviews. Now the training set contains all the words from the collected movie reviews, 200 positive and 200 negative. The performance of this model is evaluated using the performance measures precision and recall [3]. The following table shows the confusion matrix for our analysis:

Table 4. Confusion matrix

		Predicted label	
		Positive	Negative
Known label	Positive	Tp = 25	Fp = 5
	Negative	Fn = 9	Tn = 21

$$Precision = 0,8333 \tag{6}$$

$$Recall = 0,7452 \tag{7}$$

$$Accuracy = 0,7666 \tag{8}$$

This is a small set of test data and in this case we are talking about very good results. If we have a larger test set, the probability of accuracy certainly would be less.

7.3 Testing with Other Classifiers

In the second experimental phase we used WEKA (Waikato Environment for Knowledge Analysis) to obtain the results from other classification techniques. WEKA is an open source data-mining tool [4]. WEKA has implementations of numerous classification and prediction algorithms. We are using the two decision tree algorithms J48 (C4.5) and ADTree with default values. The results are shown in the table below.

Table 5. Comparison results from each algorithm

Test scenario	Algorithm	Precision	Recall	Accuracy
Training set	J48	0,96	0,96	0,96
	ADTree	0,789	0,725	0,725
Cross-validation	J48	0,658	0,658	0,657
	ADTree	0,568	0,568	0,567

From the results in the table 5, we can assume that the Decision Tree classifier with accuracy rate around 60% in the two test scenarios does not perform the classification as well as the Naive Bayes classifier.

8 Discussion

Comparing the obtained results, we can assume that the Naive Bayes classifier works better with the collected movie reviews. We believe that we have good results because of the nature of our language and the size of our reviews. The algorithm is improved by processing new data as neutral words, which are excluded from the classification. In order to understand the misclassification of some reviews, we analyzed their content and found some of the following problems.

1. Neutral reviews are randomly classified according to the dominant sentiment of the contained words.
2. Some of the words that are positive are more frequent in negative reviews because of the negation.
3. Some of the words that are negative are more frequent in positive reviews because of the negation.
4. Irony is classified as negative sentiment.

There are specific cases in the Macedonian language for which this algorithm, most certainly does not make sense. Irony, Sarcasm and Metaphor are typical examples.

Movie reviews are great way of expressing an opinion of a movie. The reviews give enough details about the movie, and from it the reader can make an informed decision. In this paper, we selected the reviews that were expressed with numerical value or strength positive and negative meaning. The authors of the reviews are not public, and each of them is a member of the online forum from which reviews are extracted. The collected movie reviews are with extremely positive or negative sentiment and that's why the accuracy with k-fold validation is so high.

9 Conclusion

The increased interest in sentimental analysis is partly due to the potential use in applications available online. Equally important are the new intellectual challenges to the research community. There is a huge potential for processing data that make analysis of public opinions on a product or service. In this paper, a supervised approach to sentiment analysis of Macedonian movie reviews was described. Sentiment analysis using a Naive Bayes algorithm showed significant results, considering even small training data set. Despite the good results, there is a lot of space for improvements. The accuracy of the analyzer can be improved by providing a larger set of testing data. It is impossible for sentiment analysis to ever be 100 % accurate, but we will take new phrases and identify them, for example sarcasm and irony, as much as possible. As further improvement, we plan to add neutral sentiment as a separate category, and we suppose it will also affect the results.

References

1. Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S.: The “K"in K-fold Cross Validation. In: ESANN (2012)
2. Çelikkaya, G., Torunoğlu, D., Eryiğit, G.: Named Entity Recognition on Real Data: A Preliminary Investigation for Turkish. In: Proceedings of the 7th International Conference on Application of Information and Communication Technologies, AICT 2013. IEEE, Baku (2013)
3. Davis, J., Goadrich, M.: The relationship between Precision-Recall and ROC curves. In: ICML, pp. 233–240 (2006)
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations 11(1), 10–18 (2009)
5. Liang, A.: Rotten Tomatoes: Sentiment Classification in Movie Reviews. CS 229 (15 2006)
6. Liu, B.: Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing, 2nd edn. Taylor and Francis Group, Boca (2010)
7. Montejo-Ráez, A., Martìnez-Cámara, E., Martìn-Valdivia, T.M., Ureña-López, A.L.: Random Walk Weighting over SentiWordNet for Sentiment Polarity Detection on Twitter. In: Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, pp. 3–10. Association for Computational Linguistics (2012)
8. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2007)
9. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. CoRR cs.CL/0205070 (2002)
10. Smirnov, I.: Overview of Stemming Algorithms. Mechanical Translation (December 2008)
11. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: ACL, pp. 417–424 (2002)
12. Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., Patwardhan, S.: OpinionFinder: A System for Subjectivity Analysis. In: HLT/EMNLP (2005)

Efficient Attacks in Industrial Wireless Sensor Networks

Spase Stojanovski and Andrea Kulakov

Faculty of Computer Science and Engineering
University Sts. Cyril and Methodius
Skopje, Macedonia
spase.stojanovski@gmail.com, andrea.kulakov@finki.ukim.mk

Abstract. The need for applying WSN in Industrial Control Systems (ICS) leads to the development of the first open standard for wireless communications, *WirelessHART*, designed for wireless real-time monitoring and control of industrial processes. Sensor networks in ICS require high availability, since the consequences of abusing these systems might result in a catastrophic event. In this paper we analyze the *WirelessHART* protocol and examine how secure it is in terms of external attacks. We conduct our analysis on the Medium Access Control layer. Results show that systems based on the WirelessHART protocol are easily subjected to external jamming interference, disrupting the real-time communication in the industrial control system. Our main contribution is the proposed algorithm which shows the ability of a malicious sensor node to sniff the network traffic and abuse the learned parameters to disrupt the communication in an efficient manner.

Keywords: Wireless Sensor Networks, WSN, WirelessHART protocol, security, attacks, malicious node.

1 Introduction

In this paper we are investigating the security of wireless sensor networks deployed in industrial control systems. The medium used for communication makes sensor networks more susceptible to attacks [1]. External malicious nodes aim to decrease the lifetime of the networks' sensor nodes and might cause loss of availability of the entire industrial control system. Therefore, research in this area is of high importance. In particular, we analyze the WirelessHART protocol [2], running in a virtual sensor network and we focus on the data link layer.

The authors in [3] analyze several MAC protocols based on their timing requirements, and propose attack scenarios for energy efficient jamming exploiting semantics of the data link layer. In [4] the authors initiate jamming attacks following radio activity detection by periodically reading Radio Signal Strength Indicator (RSSI), Clear-Channel Assessment (CCA) and after capturing the Start of Frame Delimeter (SFD) sequence. In [5] the authors analyse how long nodes are kept awake when sending random messages, unathenticated messages and valid replayed messages.

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_29

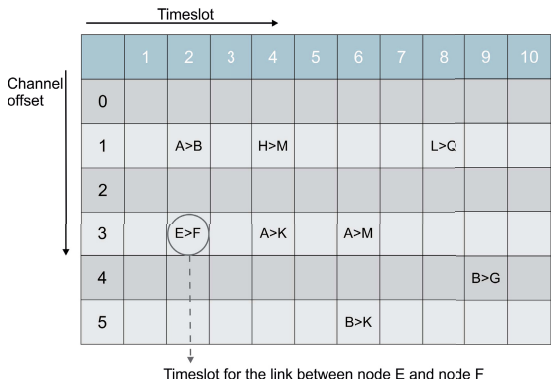

Fig. 1. Slot and channel matrix

We suggest an algorithm where malicious node performs an efficient attack in the sensor network. We showed on a few tests that using this algorithm, the attacker can significantly decrease the availability of the network communication by causing lost packets and delay in the network communication.

Results show that systems based on the WirelessHART protocol are easily subjected to external jamming interference, disrupting the real-time communication in the industrial control system. Our main contribution is the proposed algorithm which shows the ability of a malicious sensor node to sniff the network traffic and abuse the learned parameters to disrupt the communication in an efficient manner.

2 Time Slotted Channel Hopping (TSCH) in WirelessHART

The main concepts the WirelessHART protocol stack is based on are derived from the Time Synchronized Mesh Protocol [6]. TSMP is a media access protocol developed by DustNetworks designed for low power sensor networks. As a subset of TSMP the concept of Time Slotted Channel Hopping (TSCH) represents a MAC scheme that allows robust communication through channel hopping. TSCH defines time-slotted communication where messages are sent and received based on a schedule which clearly states on what slot and which channel the sensor node receives/sends the message from/to a neighbor node. The main idea behind the TSCH concept is that the wireless channel is divided into frequency and time. Time is splitted into discrete *timeslots*. The approach that TSCH takes to model the Radio Frequency (RF) space is based on a matrix with slot-channel cells.

Fig. 1 shows the TSCH matrix. The presented figure illustrates the TSCH concept of a *timeframe* which represents a collection of cells that repeat at constant intervals. The presented *timeslot* is with length of $10ms$ and repeats once every $100ms$. In this case the *timeframe* consists of 10 slots. TSCH defines

the notion of *link*. A *link* specifies the message exchange that occurs in one cell between nodes. In a simple scenario a *link* defines the message exchange between two sensor nodes at either end of the link that communicate periodically once in every *timeframe*. The *links* in TSCH hop based on a pseudorandom scheme over predefined channels, one packet at a time.

Each message is transmitted over a certain frequency which is the physical channel. For every pair of nodes that communicate to each other, a fixed value is defined denoted as *channel offset*. This value is used to determine the physical channel with Equation 1:

$$(\text{AbsoluteSlotNumber} + \text{ChannelOffset}) \; mod \; \text{NumberChannels}. \quad (1)$$

In Equation 1, the value $AbsoluteSlotNumber(ASN)$ represents the number of *timeslots* since the network was formed, $ChannelOffset$ is the offset for the channel for a specific *link* in the *slot* and *channel* matrix, and $NumberChannels$ denotes the number of active channels. In each *timeframe* a *link* for a pair of nodes is always defined in the same *timeslot* and in the same channel offset. The only difference is the value for the physical channel of every *link*.

3 Efficient Attacks

The requirements for an attacker to execute efficient attacks in the WirelessHART sensor network are: i) Least amount of energy spent; ii) Increased number of dropped packets in the network and iii) Decreased reliability of the network.

In this paper we try to reduce the reliability of a WirelessHART sensor network by executing efficient attack of the communication between two sensor nodes. Further we explain the attack scenario executed by external malicious node on existing WirelessHART network. We define the following assumptions around which we build our attack scenario: i) An external node performs an attack in a sensor network in which the network traffic is based on the WirelessHART protocol; the node is capable of sending and receiving messages based on this protocol. ii) The attacking node is hidden and placed centrally relative to the geographical location of the sensor nodes from the existing WirelessHART network. iii) The WirelessHART sensor network is already formed and active.

3.1 Attack Scenario

Our main approach is to induce collision between packets with valid messages by injecting malicious requests into the network traffic. To achieve this goal efficiently, the attacker must send messages in the network at the same time when the valid messages between two nodes are exchanged. Therefore, the attacker must know the exact *time slots* in which two nodes communicate and the exact *frequency* (physical channel) used for a specific message exchange. Once the *time slots* are known, the attacker is able to build different attack strategies.

WirelessHART initially determines in which *time slots* the nodes communicate during the network formation. This data is transmitted as part of the encrypted

application layer of the WirelessHART protocol stack and thus, is not easily retrievable. In the following steps we present a novel algorithm for eavesdropping the network communication and calculating the exact *time slots* in which two nodes communicate.

Step 1 (Synchronization). Initially, every sensor node needs to be synchronized with the network. This is a common procedure for every node including the malicious node, and part of the WirelessHART protocol. Synchronization means that each node has knowledge when a *time slot* begins and when it ends.

Step 2 (Determine Active Channels). The next step is to identify the active channels on which there is communication. We assume that the attacker can listen only in one channel at a time. The adversary listens the communication for a certain period in each channel starting from channel 11 until channel 26. The result from this step is the amount of active channels, denoted as N_A.

Step 3 (Calculate Channel Offset). The adversarial node chooses two nodes A and B, for which the communication will be disrupted. The value of the channel offset is the logical value of the channel in the communication between two nodes. We denote this value as CH_L. The value of the logical channel in which two nodes communicate is constant but the value for the physical channel CH_F is different in different *timeframes*. The physical channel is calculated as:

$$(ASN + CH_L)\ mod\ N_A = CH_F \tag{2}$$

The attacker node needs to calculate the value CH_L for both nodes A and B. Later with the same equation the malicious node needs to be able to calculate on which physical channel the next message is sent. This is of great importance because the attacker needs to inject malicious packets on the exact physical channel where the normal communication between the nodes A and B happens.

The value CH_L can be calculated if the attacker chooses an active physical channel to listen the communication for a period of time. Later the attacker node will check which are the two nodes that communicate between each other since this information is located in the unencrypted layer of the exchanged messages. If the attacker encounters a message that is exchanged between A and B, he retrieves the ASN value from the DLPDU header [2]. Now it is easy to calculate the CH_L value using the equation 2.

Step 4 (Following the Communication between Two Nodes). Once the CH_L value is calculated for the nodes A and B, the attacker may start following the communication between both nodes. After a message is heard on a physical channel between A and B, the attacker already knows the value ASN for that specific message. Then the attacker needs to switch to a different channel in which the next message between A and B would arrive. The value of the physical channel for the next message, ASN_{next}, depends on the ASN of the new message.

The distance to the next sent message is still not known and the attacker can not calculate the value AS_{next}. This is why we assume that a message will be sent in each following *timeslot* for which it holds: $ASN_{next} = ASN + 1$. The latter indicates that the attacker needs to switch to a new physical channel in each new *timeslot*, on every $10ms$, and check if a message has been sent between A and B.

During this traversal of physical channels the value for ASN for each captured message is saved. At the same time, the attacker calculates the time intervals between two consecutive *timeslots* in which a message is sent from A to B. We call these time intervals (distances) as *inter-arrival times*. After sufficient long listening of the communication the attacker has collected a list of such *inter-arrival times*, $r_1, r_2, ..., r_n$, and a list of ASN values, $ASN_1, ASN_2, ...ASN_n$.

Step 5 (Calculating the Length of the *timeframe*). In this step we show how the attacker can predict the length of the *timeframe*, R_L, during the network communication. The following rules are imposed by the WirelesssHART protocol:

1. The time in which there was a communication between the nodes in the period $(0, T]$, is divided in *timeframes* with equal length R_L (with the exception of the last *timeframe* where the length can be arbitrary). These *timeframes* are the following intervals: $(0, R_L]$, $(R_L, 2 * R_L]...(k * R_L, T]$;
2. The length of the *timeframe* R_L is always divisible by 10;
3. In each *timeframe* there are only a few exact *timeslots* in which a message can be exchanged. The protocol defines the exact values $p_1 < p_2 < ... < p_m$, and in any *timeframe* $R_k = (t, t + R_L]$, a message can be exchanged only in the moments $t + p_1, t + p_2, ..., t + p_m$, $R_L > p_m$.

In Fig. 2 a segment of the communication between two eavesdropped nodes A and B is shown. The length of the *timeframe* is 20, while the values $p_1, p_2, ..., p_m$ that denote the possible *timeslots* used for message exchange are: 4, 7, 16. Thus in the first *timeframe* a message is sent in the *timeslots* 4 16, in the second *timeframe* no message is sent and in the third *timeframe* a message is sent in every possible *timeslot*: 44, 47, 56. While eavesdropping the communication in the example from Fig. 2, after executing Step 4, the attacker has identified 4 *inter-arrival times* in which a message has been received: $r_1 = 12$, $r_2 = 28$, $r_3 = 3$ and r_4. The exact *timeslots* for sending a message $(4, 16, 44, 47, 56)$ are unknown for the attacker since he is not aware of the exact moment when the communication between the nodes started.

Messages sent during the same *timeslot* p_i from different *timeframes* are called *compatible messages*. Two *compatible messages* sent in two adjacent *timeframes* are *adjacent compatible messages*. In Fig. 2 the messages sent in the *timeslots* 4 and 44 are *compatible messages*, as well as 16 and 56, while there exist no *adjacent compatible messages* in all three *timeframes*. In case a message was sent in the *timeslot* 24, it would be *adjacent compatible message* with the messages in the *timeslots* 4 and 44.

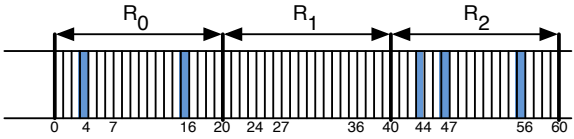

Fig. 2. Communication between two nodes in the period of the first 3 *timeframes*

Further, we note that if a message P is sent in the moment t, a new message P' is expected in the moment $t + k * R_L$, a message which is *compatible message* with P. If $k = 1$, then the message is *adjacent compatible message*. The difference between the time of arrival of the messages P and P', as well as between any two *compatible messages* is $k * R_L$.

The input data for the attacker are only the differences between the *timeslots* in which two consecutive messages arrive. The distance between two *timeslots* when two *compatible messages* P and P' arrive is actually sum of several consecutive *inter-arrival times*, which are the *inter-arrival times* of messages that arrive between P and P'. In the example in Fig. 2 the distance between two *compatible messages* sent in the *timeslots* 4 and 44 is the sum of two *inter-arrival times* 12 and 28 which occur due to the message sent in the *timeslot* 16.

Further we explain the main idea behind our proposed algorithm. For each message, we identify the distance rk to the next *compatible* from the sniffed network communication. Later we will describe how these values are calculated. Next we analyze the set of these distances $S = \{rk_1, rk_2, ..., rk_n\}$. We are certain that each of these distances rk_i is a value in the form: $rk_i = k * R_L$. This will allow us to calculate the exact value for the length R_L which is the *greatest common divisor (GCD)* of all distances:

$$R_L = GCD(rk_1, rk_2, ..., rk_n). \tag{3}$$

Intuitively calculating the value for R_L in the proposed manner will be correct even after several calculations, rk_i: in the following sections we will describe why it is important to calculate great number of these distances while analyzing the frequency of occurrence of each distance.

In our example on Fig. 2, the set of distances is $S = \{40, 40\}$ since there are only two pairs of *compatible messages* shown with length of 40 in between. The result from our algorithm would be $R_L = 40$. Nonetheless in a real environment in which the attacker would listen the network traffic for a long time it is highly probable to expect the correct result 20, in which case we would expect at least two *compatible messages* with distance of 20 or 60.

Still the main challenge is the attacker to calculate the values from the set S. According to the rule 2 defined above, the value R_L is divisible by 10. That is why our idea is to calculate the sum of several *inter-arrival times* for a given message, until the sum is divisible by 10.

In the communication we might encounter two messages P and P' occurred at the same moment in time t_1 t_2, so that $t_2 - t_1 = 10 * k$ holds, but still they

are *compatible messages*. We call these *false compatible messages*; they need to be eliminated from the calculation of the value R_L, see Equation 3.

When analyzing real data it was shown clearly that the *false compatible messages* are easy to be noticed because the frequency of their occurrence is very small in comparison with the frequency of occurrence of real *compatible messages*. Thus it is very important to determine the frequency of occurrence of every value rk_i from the set S of distances between the closest *compatible messages*. From the analysis that we have conducted, we noticed that the frequency of occurence of the distance between two *false compatible messages* is at least 2.5 times smaller than the frequency of occurrence of each of the distances between true *compatible messages*. As a result, we modify the Equation 3 and shown below:

$$\begin{aligned} R_L = GCD(rk'_1, ..., rk'_k) \; where \; & SV = \{rk'_1, ..., rk'_k\} \subseteq S && (i) \\ & \forall i, j = 1..k, i \neq j \Rightarrow rk'_i \neq rk'_j && (ii) \\ & \forall rk_m \in S \setminus SV, \; \forall rk_n \in SV \frac{f(rk_n)}{f(rk_m)} \geq 2.5 && (iii) \end{aligned} \tag{4}$$

Equation 4 shows that the result from the algorithm is the *greatest common divisor (GCD)* from the distances from the set SV. The idea is that this set will store the distances from the set S for which we expect they are the distances between true *compatible messages* (i). The condition (ii) points to the difference between all values in the set SV. Moreover, the condition (iii) shows that the distances that are not considered for the calculation $S \setminus SV$, are expected to be distances between *false compatible messages*: all distances from SV have occurred at least 2.5 times more than the distances between *false compatible messages*. The function $f(rk_i)$ denotes the frequency of occurrence of the difference rk_i.

Step 6 (Calculating *timeslots*) The next step is to determine in which *timeslots* the nodes communicate. In Step 4 we show how an attacker is able to calculate the value of the *Absolute Slot Number, ASN* for each message. Using the values, $ASN_1, ASN_2, ...ASN_n$, as well as the value for the length of *timeframe* R_L that we calculate in Step 5, we can calculate the *timeslot* in which the message was sent. The *timeslot* TSN_i for a message with *Absolute Time Slot* ASN_i is calculated as: $TSN_i = ASN_i \; mod \; R_L$.

We expect to have most of the calculated values repeated during the calculation of the *timeslots*. The final result from this step is a set of only few different values TSN of *timeslots* in which communication might occur in any *timeframe*.

Step 7 (Injecting Malicious Packets). Lats step from the attack scenario is to inject malicious packets into the normal operating sensor network. The simplest strategy to inject packets in each of the calculated *timeslots* from Step 6. The procedure to inject packets is the following: as soon as the attacker detects a message between A and B, he already knows the ASN value and then the TSN value for the detected message. The former and latter values are used to calculate

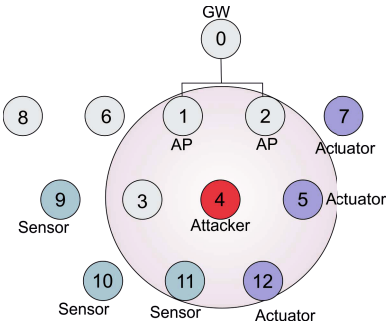

Fig. 3. Simulated attack of a WirelessHART sensor network

the values for TSN_{next} and ASN_{next} for the next message. The value for the physical channel where the new message will be sent is calculated using Equation 2. The attacker is switching to this channel, sleeps for period of $TSNnext-TSN$ *timeslots* and then injects packets in this channel. The procedure after the latter action is repeated.

4 Results

In this section we will show the effects of the implemented attack scenario in a simulated WirelessHART sensor network in the network simulator NS-2 [7] [8]. In Fig. 3 the simulated network including a malicious node consists of:

- GW 0 - Gateway node for aggregating sensor data from the WirelessHART sensor network and distributing this data into the remaining automation network;
- AP1, AP2 - Access point nodes for collecting and forwarding sensor data from different parts of the sensor network;
- Sensor nodes 9, 10 and 11, which emit sensor data on a predefined period;
- Actuator nodes 12, 5 and 7, which process sensor data and execute control commands;
- Malicious attacker node 4;
- Remaining sensor nodes 3, 8 and 6, that forward sensor data throughout the network.

A typical application of an WirelessHART sensor network in an industrial control system defines *sensor nodes* and *actuator nodes*. The *sensor nodes* emit sensor data from the industrial process towards the *actuator nodes* periodically. Based on the received information, the *actuator nodes* calculate the next control command to maintain the process in the desired state.

The simulated sensor network contains the attacker node, placed centrally with respect to the nodes from the sensor network. The attacker node is running

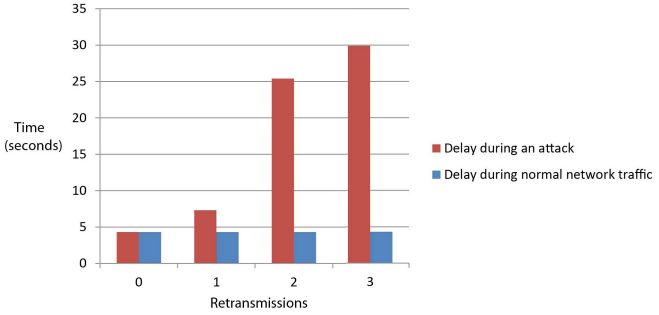

Fig. 4. Delay in exchanged messages between sensor and actuator nodes

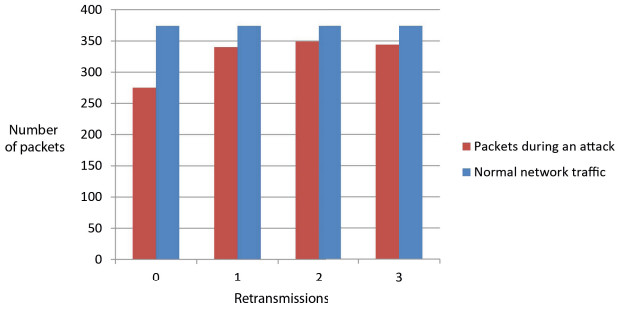

Fig. 5. Packets that arrive at the actuator node

a modified WirelessHART protocol stack including the implementation of the proposed algorithm.

In this simulation, an attack is performed on the communication between the sensor node 3 and sensor node 1. To be able to determine the effect of this attack on the reliability of the sensor network we analyze the following metrics: i) *end to end delay* - the delay of messages sent from the senor nodes to the actuator nodes; ii) *packet delivery ratio* - the ratio of the lost packets which did not reach to the actuator nodes.

WirelessHART protocol defines a mechanism for retransmission, where a message is sent repeatedly for the defined amount of retries until the source node receives confirmation for reception on time. We analyze the attacks in several cases with different values for *retransmission*.

Fig. 4 shows the delay of packets in the sensor network. The measurements are performed by comparing the arrival time of the packets in the actuator nodes, during normal traffic and during an attack.

Fig. 5 shows the comparison of the arrived packets to the actuator nodes in normal network traffic and in case where the attacker node performs efficient disruption of the communication between node 3 and node 1.

Fig. 4 and Fig.5 show that without any retransmission attempt, the arrival time of the messages is the same whether there is an attack or there is normal

traffic flow. With retransmission of failed packets, the probability that a message will reach the destination actuator node increases. In this case the attack interferes the communication and delays the reception of the message. The delay in the communication is due to the collision of normal traffic with injected malicious traffic form the attacker node, thus the messages need to be retransmitted.

5 Conclusion

In this paper we analyzed the security of wireless sensor networks in industrial systems. In particular we studied WirelessHART, the first open standard for industrial sensor networks. In order to show weaknesses in WirelessHART, we proposed an algorithm for an attacker to find out concrete parameters in the network communication, which are the key for executing an efficient attack. We showed that the result of such an attack might cause large number of lost packets and therefore, decreased reliability in the network communication.

Our algorithm shows certain weaknesses in WirelessHART, protocol used in industrial control systems. For future work, we plan to use the results from the paper and further investigate techniques for mitigating vulnerabilites.

Acknowledgments. We thank Pouria Zand for his useful feedback.

References

1. Wood, A.D., Stankovic, J.A.: Denial of service in sensor networks. IEEE Computer 35(10), 54–62 (2002)
2. Staff, B.S.I.: Industrial Communication Networks. Fieldbus Specifications. WirelessHART Communication Network and Communication Profile. B S I Standards (2009), http://books.google.nl/books?id=OIyGPgAACAAJ
3. Law, Y.W., Van Hoesel, L., Doumen, J., Hartel, P., Havinga, P.: Energy-efficient link-layer jamming attacks against wireless sensor network mac protocols. In: Proceedings of the 3rd ACM Workshop on Security of Ad hoc and Sensor Networks, pp. 76–88. ACM (2005)
4. Wood, A.D., Stankovic, J.A., Zhou, G.: Deejam: Defeating energy-efficient jamming in ieee 802.15.4-based wireless networks. In: SECON, pp. 60–69 (2007)
5. Raymond, D.R., Marchany, R.C., Brownfield, M.I., Midkiff, S.F.: Effects of denial-of-sleep attacks on wireless sensor network mac protocols. IEEE T. Vehicular Technology 58(1), 367–380 (2009)
6. Pister, K., Doherty, L.: Tsmp: Time synchronized mesh protocol. In: IASTED Distributed Sensor Networks, pp. 391–398 (2008)
7. Proceedings of 2012 IEEE 17th International Conference on Emerging Technologies & Factory Automation, ETFA 2012, Krakow, Poland, September 17-21. IEEE (2012)
8. The network simulator ns-2 (2014), http://www.isi.edu/nsnam/ns

Robustness of the Gray Code Arrangements of the Genetic Code in Mitochondria

Dragan Bošnački, Hubertus M.M. ten Eikelder,
Marieke Maanders, and Peter A.J. Hilbers

Faculty of Biomedical Engineering, Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
{D.Bosnacki,H.M.M.t.Eikelder,M.Maanders,P.A.J.Hilbers}@tue.nl

Abstract. The genetic code determines how the genetic information is translated into proteins, which are building blocks of the living organisms. Genetic code can be related to Gray codes, a class of error resisting codes. In this paper we investigate the robustness of the Gray code representations of the genetic code in mitochondria, small cell organelles that contain genetic information separate from the rest of the cell. Mitochondria use a slightly modified version of the standard code which defines the main genetic information. Our result show that despite the fact that the mitochondrial codes seemingly show more regularities than the standard code, they are less robust with regard to the Gray code arrangement criterion. This could be a result of the lighter evolutionary pressure that the mitochondrial codes have endured compared to the standard code.

Keywords: computational biology, genetic code, Gray code, mitochondria, error resisting codes.

1 Introduction

The genetic code [1] governs the translation of the genetic information to proteins, which is one of the most basic genetic processes taking place in the cells of all living beings. There are patterns in the genetic code which indicate that during the evolution the code has been optimized in order to reduce the errors in the translation of the genetic information. In that context, the idea to relate the genetic code to a special kind of codes, called Gray codes [2] was originally presented by Rosemary Swanson as a "unifying concept for the amino acid code" [3].

In a previous work [4], by reducing the arrangement forming to the well known Traveling Salesman Problem, we showed that cyclic arrangements of the code different from the one proposed by Swanson. These arrangements are still based on the principle of minimal change used in the Gray code and yield a better grouping of the amino acids by similarity.

Most of the genetic information resides in the DNA in the nucleus of the cell (more precisely, this holds for the eucariotes, the organisms that have a

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_30

cell nucleus). This information is governed by the standard genetic code. It is well known that the genetic code is slightly different in mitochondria, small intracellular organelles which are responsible for energy production and cellular respiration, and which also contain part of the genetic information.

We compare the robustness of the mitochondrial and the standard genetic codes, in the view of the cyclic code arrangements. The goal is to investigate a possible role of the evolution in the shaping of the genetic code. To this end we introduce the idea that the quality of grouping can be used as a measure of the robustness of the code. The more the groups are compact, the more the code is resistant to errors during the translation process.

Since at first sight the mitochondrial code exhibits more regularity than the standard code, one might expect that this would entail a greater robustness of the code. However, our results show that, on the contrary, the mitochondrial code is less robust than the standard code, possibly because the evolutionary pressure on the mitochondrial code is less emphasized than on the standard code. Changes in the mitochondria are usually not lethal to the organism which allows a greater margin of error during the translation from mRNA to protein.

2 Preliminaries

Genetic Code In all organisms a gene is usually constructed of several hundreds of codons. A codon is a sequence of three letters (called bases) that codes for an amino acid. The amino acids are building blocks of the proteins - the proteins can be considered as strings of amino acids. Over an alphabet consisting of four letters, A, G, C, and U, in total 64 codons are present, encoding for 20 amino acids. In the textbooks the genetic code is usually represented as a table (see Fig. 1). A notable feature of the code is that not all amino acids are encoded with the same number of codons. For instance, three amino acids, leucine, serine and arginine are represented by six codons, whereas the amino acid methionine is encoded with only one codon. This same codon, AUG, has a double role since it encodes also the start of the gene, therefore it is called the start codon. Three codons are reserved as stop codons, indicating the end of transcription.

Classification of the Amino Acids Following [4,3,5], we classify the amino acids occurring in the proteins (and therefore in the code) in four groups. Their identification is based on size and hydrophobicity. {A,T,G,P,S}, {D,N,E,Q,K}, {H,R,W,Y,F} and {L,M,I,V,C} in one letter code. Four groups can roughly be characterized as small (size), external (hydrophilic), large (size) and internal (hydrophobic), respectively. We consider a cyclic ordering of the groups. Thus, in a natural way we consider successive groups as neighboring, e.g., small and external, but also, because we assume cyclicity, small and internal. Similarly, we say that the small and large groups, as well as for the hydrophilic and hydrophobic, are opposite groups.

Amino acids which are similar by their characteristics and have substituted for one another during the evolution of the living organisms, are encoded with

UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC Phe	UCC Ser	UAC Tyr	UGC Cys
UUA Leu	UCA Ser	UAA END	UGA END
UUG Leu	UCG Ser	UAG END	UGG Trp
CUU Leu	CCU Pro	CAU His	CGU Arg
CUC Leu	CCC Pro	CAC His	CGC Arg
CUA Leu	CCA Pro	CAA Gln	CGA Arg
CUG Leu	CCG Pro	CAG Gln	CGG Arg
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC Ile	ACC Thr	AAC Asn	AGC Ser
AUA Ile	ACA Thr	AAA Lys	AGA Arg
AUG Met	ACG Thr	AAG Lys	AGG Arg
GUU Val	GCU Ala	GAU Asp	GGU Gly
GUC Val	GCC Ala	GAC Asp	GGC Gly
GUA Val	GCA Ala	GAA Glu	GGA Gly
GUG Val	GCG Ala	GAG Glu	GGG Gly

Fig. 1. The genetic code

similar codons. We build on this observation and we further investigate what the possible influence the evolution could have had on the code.

Gray Codes The Gray codes [2] were invented with the idea that, in order to minimize the errors during information transfer, similar entities should be encoded with similar code words. To illustrate the concept, consider the standard binary encodings of the numbers 0, 1, 2, and 3, i.e., 00, 01, 10, 11. The goal is that the encodings of successive numbers should differ in exactly one position. With the standard encoding, this is true for the pairs 0 and 1, and 2 and 3, but it does not hold for the pair 1 and 2, since their encodings differ in both positions. A better, Gray code, encoding is possible: 00, 01, 11, 10, in which indeed all neighboring numbers differ in only one position. Moreover, this code is a cyclic one, since also the first and the last code words differ in only one position. So, if we consider a modulo 4 arithmetic, i.e., if 3 and 0 are also neighbors, then they are encoded with code words differing in only one position.

By extrapolating this idea from binary codes, we we consider cyclic orderings of the 64 three letter codons such that each pair of adjacent codons differs in exactly one position. The corresponding letters which are in this position of difference do not matter. In other words, we assume that all letters are equally (di)similar.

The Code in Mitochondria Mitochondria, small intracellular organelles which are responsible for energy production and cellular respiration, often have different codes. Each cell contains hundreds of mitochondria, so the chance that a codon corruption (mutation) would cause a meltdown of the cell and its descendants is much smaller compared to a mutation in the main genetic information, usually stored in the cell nucleus. Therefore, one would expect that the chance for a change in the mitochondrial code would be much greater since the evolutionary pressure is much smaller than in a complex genomes of thousands of genes.

Mitochondria can be divided in two categories:

- *Plant*, using the universal genetic code in Figure 1.
- *Non-plant*, the non universal genetic codes in mitochondria are derived from the universal code. The genetic code variations found in non-plant mitochondria are presented in Table 1.

Table 1. Genetic code variations found in non-plant mitochondria [1]

Organism	UGA Stop	AUA Ile	AAA Leu	AGN Arg	CUN Leu	UAA Stop
Vertebrates	Trp	Met	Lys	Stop	Leu	Stop
Arthropods	Trp	Met	Lys	Ser*	Leu	Stop
Echinoderms	Trp	Ile	Asn	Ser	Leu	Stop
Molluses	Trp	Met	Lys	Ser	Leu	Stop
Platyhelminths	Trp	Ile	Asn	Ser	Leu	Tyr
Yeast	Trp	Met	Lys	Arg	Thr	Stop
Euascomycetes	Trp	Ile	Lys	Arg	Leu	Stop
Protozoa	Trp	Ile	Lys	Arg	Leu	Stop

* *AGA only*

3 Methods

Along the line of the methods from [4], which are based on the Traveling Salesman Problem (TSP), cyclic Gray code arrangements are made of the mitochondrial codes of the non-plant organisms given in Table 1. This means that we consider only cyclic arrangements of all 64 codons such that each codon with each of its two neighbors differs in exactly one position. We define a distance between the codons to be the same as the distance between the amino acids as defined in the Pet91 matrix [6]. In the context of the TSP problem the codons are considered as cities, therefore the goal is to minimize the sum of all distances between the neighboring codons, i.e., to minimize the length of the cyclic route of the traveling salesman.

1 000 000 iterations are used to derive the cyclic Gray codes. This program calculates the TSP-solution of the given codon distribution. In an ideal, most robust, arrangement of the codons these groups should remain intact with no exceptions. That is, the codons encoding for amino acids of the same group should remain as close as possible in the arrangement. In reality this is not achievable, so to asses the quality of the arrangements, we introduce a ranking based on penalties. Two cases of group mismatch can be considered: an amino acid is in the neighbor group or in the opposite group, where the latter case is the worst. This is reasonable because, when instead of a small amino acid a large one is translated in a protein, the folding might be blocked. Similar arguments apply to the internal and external amino acids. For each occurrence of a (codon of an) amino acid outside its group a penalty is added which is defined as follows:

$$P(a) = \begin{cases} 0, \text{ if } a \text{ is in the correct group} \\ 1, \text{ if} a \text{ is in the neighboring group} \\ 2, \text{ if} a \text{ is in the opposite group} \end{cases} \quad (1)$$

The penalty values are chosen rather arbitrarily. Of course, a more refined penalty system is possible. e.g., based on further physico-chemical analysis of the main properties of the amino acid groups.

4 Results

Cyclic gray code arrangements are made of all mitochondrial codes of the above mentioned non-plant organisms. The robustness results are summarized in Table 2.

Table 2. Penalty points of the cyclic Gray code arrangements of the genetic code of non-plant mitochondria

Organism	Da
Vertebrates	8
Arthropods	10
Echinoderms	8
Molluses	8
Platyhelminths	8
Yeast	9
Euascomycetes	12
Protozoa	12

The cyclic Gray code of the mitochondrial code of vertebrates is shown in Figure 4. Most amino acids are located in the clusters size and hydrophobicity, except for glycine. Thereby, the stop codons are not positioned in one cluster. The mitochondrial code of arthropods scores ten penalty points when arranged in cyclic Gray code based on the above mentioned distance. Three quarter of the arrangement in the cycle is faultless and only in the cluster with large amino acids three amino acids are positioned in a different location than expected, as visualized in Figure 3. The small amino acid glycine are wrongly incorporated, but also external lysine and the internal amino acid cysteine. So, the mitochondrial code of arthropods does not show robustness for mutation and translation errors.

Only eight penalty points are given to the cyclic Gray code of the mitochondrial code of echinoderms. Especially the middle base is arranged very regular, visualized in Figure 4. The line-up of U, C and G is not disturbed and the line-up of A is only disturbed by cysteine and tryptophan. However, the large amino acid histidine is placed in the between external amino acids.

The cyclic Gray code of the molluses is very similar to the cyclic Gray code of the echinoderms, only methionine is encoded by an additional codon, see

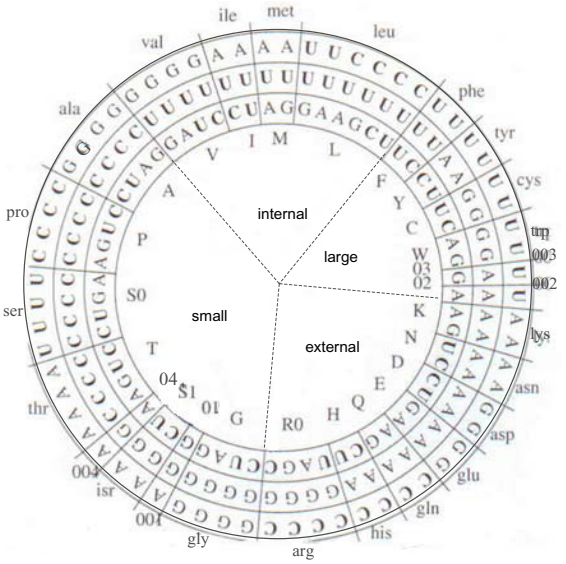

Fig. 2. Cyclic Gray code of the mitochondrial code of vertebrates

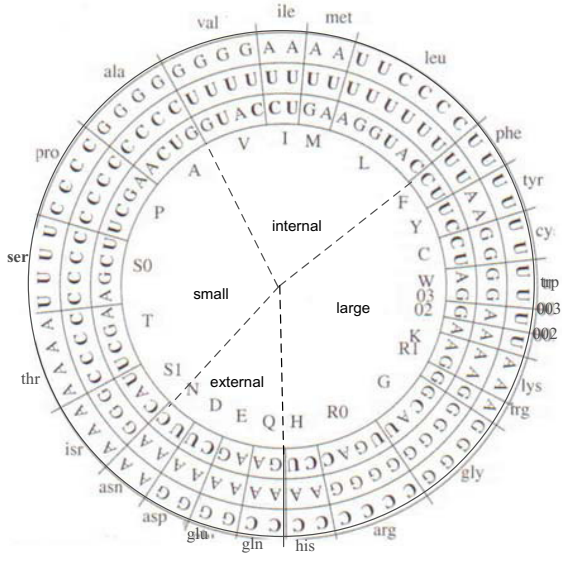

Fig. 3. Cyclic Gray code of the mitochondrial code of arthropods

Figure 5. The string of middle bases U, C and G is not disturbed and the line-up of A is only disturbed by cysteine and tryptophan. However, the large amino acid histidine is placed in the between external amino acids.

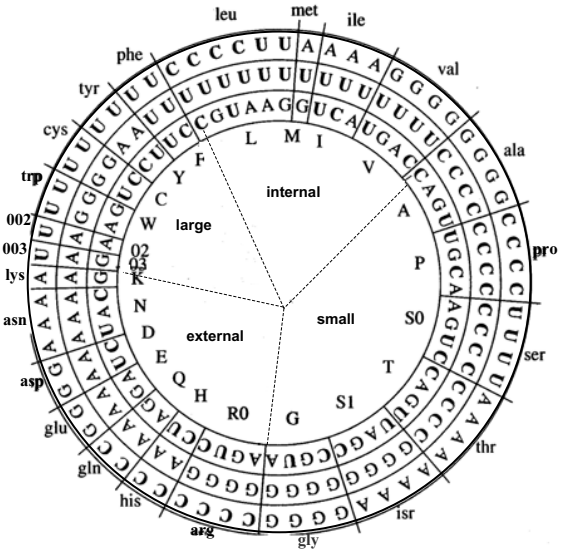

Fig. 4. Cyclic Gray code of the mitochondrial code of echinoderms

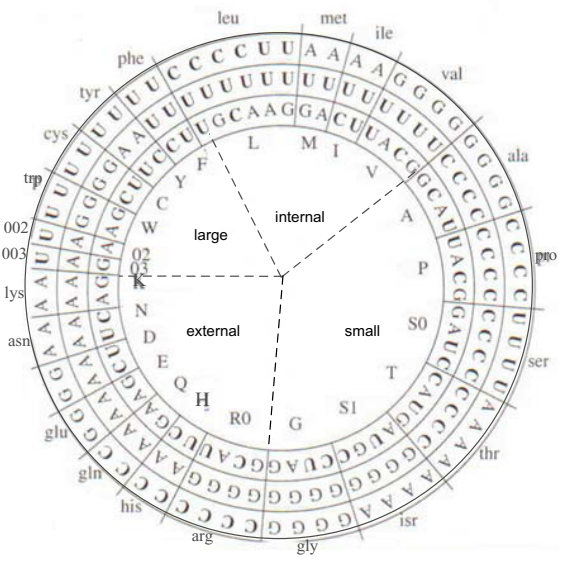

Fig. 5. Cyclic Gray code of the mitochondrial code of molluses

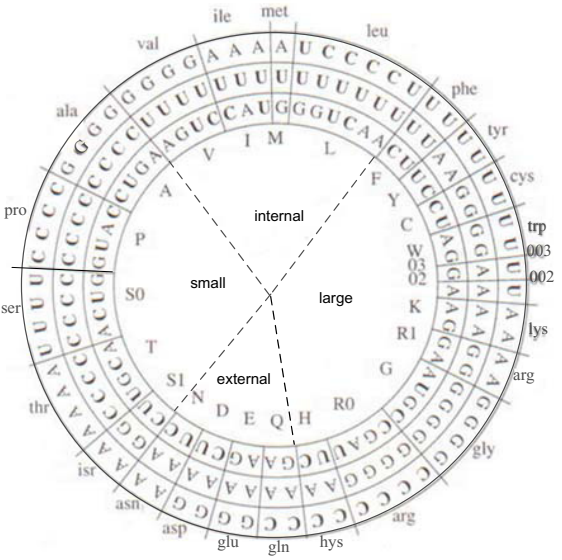

Fig. 6. Cyclic Gray code of the mitochondrial code of easacomycetes

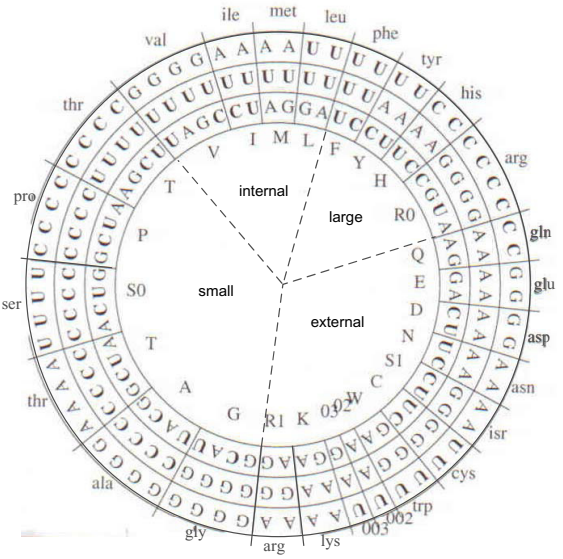

Fig. 7. Cyclic Gray code of the mitochondrial code of yeast

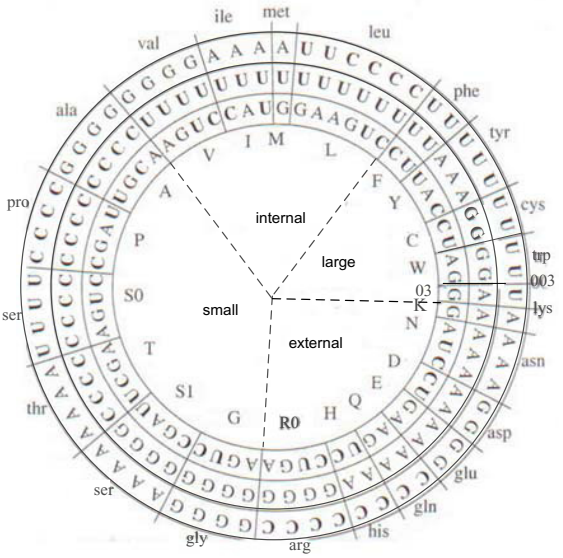

Fig. 8. Cyclic Gray code of the mitochondrial code of platyhelminths

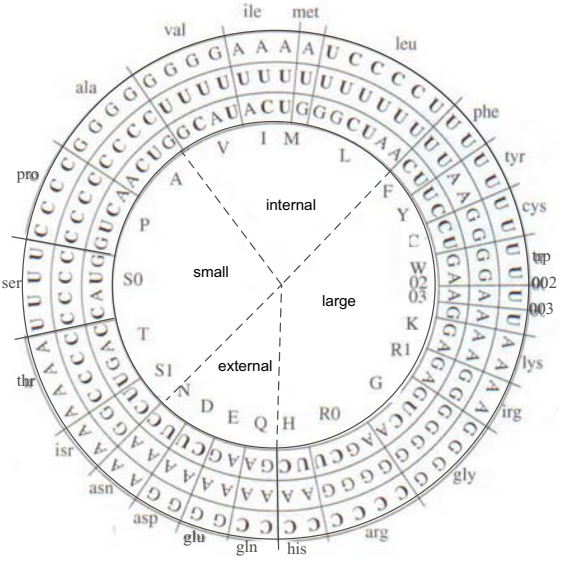

Fig. 9. Cyclic Gray code of the mitochondrial code of protozoa

The cyclic Gray code of yeast is visualized in Figure 7. The groups with small, internal and large amino acids are fairly well distributed. However, the group with external amino acids does not fit the arrangement that well. Two codons of arginine, tryptophan and cysteine and two codon of serine are all located in the group with external amino acids, resulting in 9 penalty points.

Only eight penalty points are given to the cyclic Gray code of the mitochondrial code of platyhelminths. Again, the middle base is arranged very regular, visualized in Figure 8. The lineup of U, C and G is unperturbed and the lineup of A is only perturbed by cysteine and tryptophan. Nonetheless, the large amino acids histidine and arginine are positioned between external amino acids. Furthermore, cysteine is also located wrongly, between the large amino acids. However, classifying cysteine is notoriously difficult [4,3].

The cyclic Gray code of the mitochondrial code of protozoa does not show a robustness for mutations and translation errors, see Figure 9. Twelve penalty points are scored by this arrangement, caused by a misplacement of lysine, glycine and cysteine.

5 Discussion

The outcome of calculating the robustness of the mitochondrial codes based on size and hydrophobicity suggest that the mitochondrial code has not evolved in a more robust code. Actually, only the robustness of the mitochondrial code of vertebrates, echinoderms, molluses and platyhelminths is not decreased in comparison with the cyclic gray code of [4]. In [4] it was shown that this Gray Code arrangement was most robust using as criteria the size and hydrophobicity. A decrease in robustness of the cyclic mitochondrial Gray code is found for the organisms, arthropods, yeast, euascomycetes and protozoa. The reason for that can be the smaller evolutoinary pressure which requires less robustness. In that case, the appearance of less robust distributions could be explained.

References

1. Osawa, S., Jukes, T., Watanabe, K., Muto, A.: Recent evidence for evolution of the genetic code. Microbiological Reviews 56, 229–264 (1992)
2. Gray, F.: Pulse code communication, March 17, 1953 (filed November 1947). u.s. patent 2,632,058 (1953)
3. Swanson, R.: A unifying concept for the amino acid code. Bulletin of Mathematical Biology 2, 187–203 (1984)
4. Bošnački, D., ten Eikelder, H.M.M., Hilbers, P.A.J.: Genetic code as a gray code revisited. In: Proceedings of The 2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, METMBS (2003)
5. Taylor, W.R.: The classification of amino acid conservation. Journal of Theoretical Biology 119, 205–218 (1986)
6. Jones, D.T., Taylor, W.R., Thornton, J.M.: The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences, 275–282 (1992)

Error-Detecting Code Using Linear Quasigroups

Nataša Ilievska[1] and Danilo Gligoroski[2]

[1] Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University, Skopje, Republic of Macedonia
natasa.ilievska@finki.ukim.mk
[2] Department of Telematics,
Norwegian University of Science and Technology, Trondheim, Norway
danilog@item.ntnu.no

Abstract. In this paper we consider an error-detecting code based on linear quasigroups of order 2^q defined in the following way: The input block $a_0a_1...a_{n-1}$ is extended into a block $a_0a_1...a_{n-1}d_0d_1...d_{n-1}$, where redundant characters $d_0d_1...d_{n-1}$ are defined with $d_i = a_i * a_{i+1} * a_{i+2}$, where $*$ is a linear quasigroup operation and the operations in the indexes are modulo n. We give a proof that the probability of undetected errors is independent from the distribution of the characters in the input message. We also calculate the probability of undetected errors, if quasigroups of order 8 are used. We found a class of quasigroups of order 8 that have smallest probability of undetected errors, i.e. the quasigroups which are the best for coding. We explain how the probability of undetected errors can be made arbitrary small.

Keywords: error-detecting codes, linear quasigroup, noisy channel, binary symmetric channel, probability of undetected errors.

1 Introduction

In this paper we consider an error-detecting code based on linear quasigroups of order 2^q.

Definition 1. *Quasigroup is algebraic structure* $(Q, *)$ *such that*

$$(\forall u, v \in Q)(\exists! x, y \in Q) \quad (x * u = v \ \& \ u * y = v) \tag{1}$$

Definition 2. *The quasigroup* $(Q, *)$ *of order* 2^q *is linear if there are non-singular binary matrices* A *and* B *of order* $q \times q$ *and a binary matrix* C *of order* $1 \times q$*, such that*

$$(\forall x, y \in Q) \quad x * y = z \Leftrightarrow \boldsymbol{z} = \boldsymbol{x}A + \boldsymbol{y}B + C \tag{2}$$

where $\boldsymbol{x}$*,* $\boldsymbol{y}$ *and* $\boldsymbol{z}$ *are binary representations of* x*,* y *and* z *as vectors of order* $1 \times q$ *and* $+$ *is binary addition.*

In what follows when we say that $(Q, *)$ is a quasigroup of order 2^q, than we take $Q = \{0, 1, ..., 2^q - 1\}$.

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_31

Lemma 1. *Let Q be a linear quasigroup of order 2^q and let A, B and C are matrices such that*

$$(\forall x, y \in Q) \quad x * y = z \Leftrightarrow \boldsymbol{z} = \boldsymbol{x}A + \boldsymbol{y}B + C$$

Then, for all $a_0, a_1, a_2 \in Q$, if $\boldsymbol{z}$ is the binary representation as $1 \times q$ matrix of the product $a_0 * a_1 * a_2$, then

$$\boldsymbol{z} = \boldsymbol{a_0}A^2 + \boldsymbol{a_1}BA + \boldsymbol{a_2}B + CA + C \tag{3}$$

2 Definition of the Code

We consider the following code: Let $(Q, *)$ be a quasigroup of order 2^q. Then, each input block $a_0a_1...a_{n-1}$, where $a_i \in Q$, is extended into a block $a_0a_1...a_{n-1}$ $d_0d_1...d_{n-1}$. The redundant characters d_i, are defined with the following equation:

$$d_i = a_i * a_{i+1} * a_{i+2}, \ i \in \{0, 1, ..., n-1\} \tag{4}$$

In the definition of this code, all operations in the indexes are modulo n. Finally, the extended block $a_0a_1...a_{n-1}d_0d_1...d_{n-1}$, turned into a binary form is transmitted through the binary symmetric channel.

When receiver receives the block, it checks whether the equations (4) are satisfied for the received block. If the equation is not satisfied for some $i \in \{0, 1, ..., n-1\}$, it concludes that there are errors in transmission and ask the sender to send the block once again.

3 The Probability of Undetected Errors

Since the redundant characters d_0, d_1,..., d_{n-1} are transmitted trough the binary symmetric channel, the noise affects them too. This means that some of the redundant characters may be incorrectly transmitted thus making the following situation possible: the equations (4) are satisfied, although some of the information characters a_0, a_1,..., a_{n-1} are incorrectly transmitted. Consequently, it is possible to have undetected errors in transmission. For this reason, when we consider an error-detecting code it is important to know the probability of undetected errors. Of course, it is good this probability to be as small as possible.

For small values of p (such as the probability of bit-error p in a binary symmetric channel), the probability of undetected errors is inconsiderably small if 4 or more characters are incorrectly transmitted. For this reason, when we use term probability of undetected errors, we mean about the probability that at most 3 characters of the input block are incorrectly transmitted and the error is not detected.

With $P\{i \to j\}$ we denote the probability that i is transmitted into j and we use the following lemma:

Lemma 2. *For all binary vectors* $\boldsymbol{a}$, $\boldsymbol{b}$ *and* $\boldsymbol{c}$ *holds:*

$$P\{\boldsymbol{a}+\boldsymbol{b}\rightarrow\boldsymbol{a}+\boldsymbol{c}\}=P\{\boldsymbol{b}\rightarrow\boldsymbol{c}\} \tag{5}$$

Theorem 1. *The probability of undetected errors is independent from the distribution of the characters in the input message and from the matrix* C.

Proof. Recall that under the probability of undetected errors we mean the probability that at most three characters are incorrectly transmitted and the error is not detected.

It is sufficient to show that the claim of the theorem is true for every possible choice of at most three incorrectly transmitted information characters such that the sets of redundant characters on which they affect are not disjoint. Actually, we should show that the probabilities of undetected errors of the following random events are independent from the distribution of the characters in the input message and from the matrix C:

A_k - exactly k consecutive characters $a_i, a_{i+1}, \ldots, a_{i+k-1}$ from the initial message $a_0a_1\ldots a_{n-1}$ are incorrectly transmitted and the error is not detected, $k=1,2,3$;
C_1 - exactly two characters: a_i and a_{i+2} of the initial message $a_0a_1\ldots a_{n-1}$ are incorrectly transmitted and the error is not detected;
C_2 - exactly three characters: a_i, a_{i+1} and a_{i+3} of the initial message $a_0a_1\ldots a_{n-1}$ are incorrectly transmitted and the error is not detected;
C_2' - exactly three characters: a_i, a_{i+2} and a_{i+3} of the initial message $a_0a_1\ldots a_{n-1}$ are incorrectly transmitted and the error is not detected;
C_3 - exactly three characters: a_i, a_{i+2} and a_{i+4} of the initial message $a_0a_1\ldots a_{n-1}$ are incorrectly transmitted and the error is not detected.

We will give a general proof for the above random events, but for smaller values of the block length n, some special cases arise, so the claim should be proved for the following random events:

R_{ij} - exactly i consecutive characters from a block with length j are incorrectly transmitted and the error is not detected, when 1) $i=2, j=3$, 2) $i=3, j=3$ and 3) $i=3, j=4$;
S_{24} - exactly two nonconsecutive characters - a_i and a_{i+2} from a block with length 4 - $a_0a_1a_2a_3$, are incorrectly transmitted, and the error is not detected;
S_{35} - exactly three characters - a_i, a_{i+1} and a_{i+3} from a block with length 5 - $a_0a_1a_2a_3a_4$, are incorrectly transmitted and the error is not detected;
S_{36} - exactly three characters - a_i, a_{i+2} and a_{i+4} from a block with length 6 - $a_0a_1a_2a_3a_4a_5$, are incorrectly transmitted and the error is not detected.

Note that all operations in the indexes are modulo n.

Why these special cases occur? Let say that the two consecutive characters a_0 and a_1 are incorrectly transmitted. The information character a_0 affects the redundant characters d_{n-2}, d_{n-1} and d_0, while a_1 affects d_{n-1}, d_0 and d_1. If the block length is greater or equal then 4, then a_0 and a_1 have two common redundant characters and both of them affect d_{n-1} and d_0. But if the block length is

equal to 3, then the characters a_0 and a_1 have three common redundant characters that are affected: d_0, d_1 and d_2. For this reason, the probability $P(A_2)$ for the blocks with length three is different than the probability for the blocks with length greater than three. Therefore, this case should be considered separately from the general one, and requesting R_{23} to be introduced. A similar situation is for the other special cases.

In order to show that $P(A_1)$ (the probability that the character a_i is incorrectly transmitted and the error is not detected) is independent from the distribution of the characters in the input message and the matrix C, we introduce the following random events:
H_j: the true value of the incorrectly transmitted character a_i is j, $j \in \{0, 1, 2, ..., 2^q - 1\}$.

From the formula for total probability, we have:

$$v_1 = P(A_1) = \sum_{j=0}^{2^q-1} P(A_1|H_j)P(H_j) \tag{6}$$

$$P(A_1|H_j) = \sum_{\substack{k=0 \\ k \neq j}}^{2^q-1} B_k \tag{7}$$

where

$$\begin{aligned} B_k &= P\{a_i \to k\}P\{d_{i-2} \to a_{i-2} * a_{i-1} * k\}P\{d_{i-1} \to a_{i-1} * k * a_{i+1}\}\cdot \\ &\quad \cdot P\{d_i \to k * a_{i+1} * a_{i+2}\} \\ &= P\{j \to k\}P\{a_{i-2} * a_{i-1} * j \to a_{i-2} * a_{i-1} * k\}\cdot \\ &\quad \cdot P\{a_{i-1} * j * a_{i+1} \to a_{i-1} * k * a_{i+1}\}P\{j * a_{i+1} * a_{i+2} \to k * a_{i+1} * a_{i+2}\} \\ &= P\{\boldsymbol{j \to k}\}P\{\boldsymbol{a_{i-2}}A^2 + \boldsymbol{a_{i-1}}BA + \boldsymbol{j}B + CA + C \to \boldsymbol{a_{i-2}}A^2 + \boldsymbol{a_{i-1}}BA \\ &\quad + \boldsymbol{k}B + CA + C\}P\{\boldsymbol{a_{i-1}}A^2 + \boldsymbol{j}BA + \boldsymbol{a_{i+1}}B + CA + C \to \boldsymbol{a_{i-1}}A^2 \\ &\quad + \boldsymbol{k}BA + \boldsymbol{a_{i+1}}B + CA + C\}P\{\boldsymbol{j}A^2 + \boldsymbol{a_{i+1}}BA + \boldsymbol{a_{i+2}}B + CA + C \to \\ &\quad \to \boldsymbol{k}A^2 + \boldsymbol{a_{i+1}}BA + \boldsymbol{a_{i+2}}B + CA + C\} \end{aligned}$$

Using Lemma 2, we have

$$B_k = P\{\boldsymbol{j \to k}\}P\{\boldsymbol{j}B \to \boldsymbol{k}B\}P\{\boldsymbol{j}BA \to \boldsymbol{k}BA\}P\{\boldsymbol{j}A^2 \to \boldsymbol{k}A^2\}$$

and again Lemma 2:

$$B_k = P\{\mathbf{0} \to (\boldsymbol{k} + \boldsymbol{j})\}P\{\mathbf{0} \to (\boldsymbol{k} + \boldsymbol{j})B\}P\{\mathbf{0} \to (\boldsymbol{k} + \boldsymbol{j})BA\}P\{\mathbf{0} \to (\boldsymbol{k} + \boldsymbol{j})A^2\} \tag{8}$$

We substitute (8) in (7) and introduce replacement $l = k + j$. For a given j, l gets all values from $Q \setminus \{0\}$ when k runs over all values of $Q \setminus \{j\}$. We get:

$$P(A_1|H_j) = \sum_{l=1}^{2^q-1} P\{\mathbf{0} \to \boldsymbol{l}\}P\{\mathbf{0} \to \boldsymbol{l}B\}P\{\mathbf{0} \to \boldsymbol{l}BA\}P\{\mathbf{0} \to \boldsymbol{l}A^2\} \tag{9}$$

As we can see from (9), $P(A_1|H_j)$ does not depend on j. So, replacing (9) in (6), we obtain:

$$v_1 = P(A_1) = \sum_{l=1}^{2^q-1} P\{\mathbf{0} \to \boldsymbol{l}\}P\{\mathbf{0} \to \boldsymbol{l}B\}P\{\mathbf{0} \to \boldsymbol{l}BA\}P\{\mathbf{0} \to \boldsymbol{l}A^2\} \tag{10}$$

It is obvious from (10) that $P(A_1)$ does not depend on the true value of the incorrectly transmitted character, neither on any other character from the input message. This means that $P(A_1)$ does not depend on the distribution of the characters in the input message. Also, $P(A_1)$ does not depend on the matrix C.

To calculate $P(A_2)$ (the probability that characters a_i and a_{i+1} are incorrectly transmitted and the errors are not detected) we introduce the following random events:
H_{jk}: the true values of the incorrectly transmitted characters a_i and a_{i+1} are j and k, respectively, $j, k \in \{0, 1, 2, ..., 2^q - 1\}$.

$$v_2 = P(A_2) = \sum_{j=0}^{2^q-1}\sum_{k=0}^{2^q-1} P(A_2|H_{jk})P(H_{jk}) \tag{11}$$

$$P(A_2|H_{jk}) = \sum_{\substack{s=0 \\ s\neq j}}^{2^q-1}\sum_{\substack{t=0 \\ t\neq k}}^{2^q-1} B_{st} \tag{12}$$

where

$$\begin{aligned}
B_{st} &= P\{a_i \to s\}P\{a_{i+1} \to t\}P\{d_{i-2} \to a_{i-2} * a_{i-1} * s\}P\{d_{i-1} \to a_{i-1} * s * t\} \\
&\quad \cdot P\{d_i \to s * t * a_{i+2}\}P\{d_{i+1} \to t * a_{i+2} * a_{i+3}\} \\
&= P\{j \to s\}P\{k \to t\}P\{a_{i-2} * a_{i-1} * j \to a_{i-2} * a_{i-1} * s\} \cdot \\
&\quad \cdot P\{a_{i-1} * j * k \to a_{i-1} * s * t\}P\{j * k * a_{i+2} \to s * t * a_{i+2}\} \cdot \\
&\quad \cdot P\{k * a_{i+2} * a_{i+3} \to t * a_{i+2} * a_{i+3}\} \\
&= P\{\boldsymbol{j} \to \boldsymbol{s}\}P\{\boldsymbol{k} \to \boldsymbol{t}\}P\{\boldsymbol{j}B \to \boldsymbol{s}B\}P\{\boldsymbol{j}BA + \boldsymbol{k}B \to \boldsymbol{s}BA + \boldsymbol{t}B\} \cdot \\
&\quad \cdot P\{\boldsymbol{j}A^2 + \boldsymbol{k}BA \to \boldsymbol{s}A^2 + \boldsymbol{t}BA\}P\{\boldsymbol{k}A^2 \to \boldsymbol{t}A^2\}, \qquad \text{for } n \geq 4
\end{aligned}$$

$$\begin{aligned}
B_{st} &= P\{\mathbf{0} \to \boldsymbol{s} + \boldsymbol{j}\}P\{\mathbf{0} \to \boldsymbol{t} + \boldsymbol{k}\}P\{\mathbf{0} \to (\boldsymbol{s} + \boldsymbol{j})B\} \cdot \\
&\quad \cdot P\{\mathbf{0} \to (\boldsymbol{s} + \boldsymbol{j})BA + (\boldsymbol{t} + \boldsymbol{k})B\}P\{\mathbf{0} \to (\boldsymbol{s} + \boldsymbol{j})A^2 + (\boldsymbol{t} + \boldsymbol{k})BA\} \\
&\quad \cdot P\{\mathbf{0} \to (\boldsymbol{t} + \boldsymbol{k})A^2\}, \qquad \text{for } n \geq 4
\end{aligned} \tag{13}$$

We substitute (13) in (12) and introduce replacement $l = s + j$ and $m = t + k$. For given j and k, l and m receive all values from $Q \setminus \{0\}$ when s runs over all values of $Q \setminus \{j\}$ and t runs over all values of $Q \setminus \{k\}$. We get:

$$\begin{aligned}
P(A_2|H_{jk}) = \sum_{l=1}^{2^q-1}\sum_{m=1}^{2^q-1} & P\{\mathbf{0} \to \boldsymbol{l}\}P\{\mathbf{0} \to \boldsymbol{m}\}P\{\mathbf{0} \to \boldsymbol{l}B\}P\{\mathbf{0} \to \boldsymbol{l}BA + \boldsymbol{m}B\} \cdot \\
& \cdot P\{\mathbf{0} \to \boldsymbol{l}A^2 + \boldsymbol{m}BA\}P\{\mathbf{0} \to \boldsymbol{m}A^2\}, \qquad \text{for } n \geq 4
\end{aligned} \tag{14}$$

From (14) we see that $P(A_2|H_{jk})$ does not depend on j and k. From (14) and (11), we obtain:

$$v_2 = P(A_2) = \sum_{l=1}^{2^q-1}\sum_{m=1}^{2^q-1} P\{\mathbf{0}\to \boldsymbol{l}\}P\{\mathbf{0}\to \boldsymbol{m}\}P\{\mathbf{0}\to \boldsymbol{l}B\}P\{\mathbf{0}\to \boldsymbol{l}BA+\boldsymbol{m}B\}\cdot$$
$$\cdot P\{\mathbf{0}\to \boldsymbol{l}A^2+\boldsymbol{m}BA\}P\{\mathbf{0}\to \boldsymbol{m}A^2\}, \qquad \text{for } n\geq 4 \tag{15}$$

The probabilities for the other random events are calculated in a similar manner:

$$v_3 = P(A_3) = \sum_{h=1}^{2^q-1}\sum_{r=1}^{2^q-1}\sum_{t=1}^{2^q-1} P\{\mathbf{0}\to \boldsymbol{h}\}P\{\mathbf{0}\to \boldsymbol{r}\}P\{\mathbf{0}\to \boldsymbol{t}\}P\{\mathbf{0}\to \boldsymbol{h}B\}\cdot$$
$$\cdot P\{\mathbf{0}\to \boldsymbol{h}BA+\boldsymbol{r}B\}P\{\mathbf{0}\to \boldsymbol{h}A^2+\boldsymbol{r}BA+\boldsymbol{t}B\}\cdot$$
$$\cdot P\{\mathbf{0}\to \boldsymbol{r}A^2+\boldsymbol{t}BA\}P\{\mathbf{0}\to \boldsymbol{t}A^2\}, \qquad \text{for } n\geq 5 \tag{16}$$

$$c_1 = P(C_1) = \sum_{l=1}^{2^q-1}\sum_{m=1}^{2^q-1} P\{\mathbf{0}\to \boldsymbol{l}\}P\{\mathbf{0}\to \boldsymbol{m}\}P\{\mathbf{0}\to \boldsymbol{l}B\}P\{\mathbf{0}\to \boldsymbol{l}BA\}\cdot$$
$$\cdot P\{\mathbf{0}\to \boldsymbol{l}A^2+\boldsymbol{m}B\}P\{\mathbf{0}\to \boldsymbol{m}BA\}P\{\mathbf{0}\to \boldsymbol{m}A^2\}, \text{ for } n\geq 5 \tag{17}$$

$$c_2 = P(C_2) = \sum_{h=1}^{2^q-1}\sum_{r=1}^{2^q-1}\sum_{t=1}^{2^q-1} P\{\mathbf{0}\to \boldsymbol{h}\}P\{\mathbf{0}\to \boldsymbol{r}\}P\{\mathbf{0}\to \boldsymbol{t}\}P\{\mathbf{0}\to \boldsymbol{h}B\}\cdot$$
$$\cdot P\{\mathbf{0}\to \boldsymbol{h}BA+\boldsymbol{r}B\}P\{\mathbf{0}\to \boldsymbol{h}A^2+\boldsymbol{r}BA\}P\{\mathbf{0}\to \boldsymbol{r}A^2+\boldsymbol{t}B\}\cdot$$
$$\cdot P\{\mathbf{0}\to \boldsymbol{t}BA\}P\{\mathbf{0}\to \boldsymbol{t}A^2\}, \qquad \text{for } n\geq 6 \tag{18}$$

$$c_2' = P(C_2') = \sum_{h=1}^{2^q-1}\sum_{r=1}^{2^q-1}\sum_{t=1}^{2^q-1} P\{\mathbf{0}\to \boldsymbol{h}\}P\{\mathbf{0}\to \boldsymbol{r}\}P\{\mathbf{0}\to \boldsymbol{t}\}P\{\mathbf{0}\to \boldsymbol{h}B\}\cdot$$
$$\cdot P\{\mathbf{0}\to \boldsymbol{h}BA\}P\{\mathbf{0}\to \boldsymbol{h}A^2+\boldsymbol{r}B\}P\{\mathbf{0}\to \boldsymbol{r}BA+\boldsymbol{t}B\}\cdot$$
$$\cdot P\{\mathbf{0}\to \boldsymbol{r}A^2+\boldsymbol{t}BA\}P\{\mathbf{0}\to \boldsymbol{t}A^2\}, \qquad \text{for } n\geq 6 \tag{19}$$

$$c_3 = P(C_3) = \sum_{h=1}^{2^q-1}\sum_{r=1}^{2^q-1}\sum_{t=1}^{2^q-1} P\{\mathbf{0}\to \boldsymbol{h}\}P\{\mathbf{0}\to \boldsymbol{r}\}P\{\mathbf{0}\to \boldsymbol{t}\}P\{\mathbf{0}\to \boldsymbol{h}B\}\cdot$$
$$\cdot P\{\mathbf{0}\to \boldsymbol{h}BA\}P\{\mathbf{0}\to \boldsymbol{h}A^2+\boldsymbol{r}B\}P\{\mathbf{0}\to \boldsymbol{r}BA\}\cdot$$
$$\cdot P\{\mathbf{0}\to \boldsymbol{r}A^2+\boldsymbol{t}B\}P\{\mathbf{0}\to \boldsymbol{t}BA\}P\{\mathbf{0}\to \boldsymbol{t}A^2\}, \text{ for } n\geq 7 \tag{20}$$

$$r_{23} = P(R_{23}) = \sum_{l=1}^{2^q-1}\sum_{m=1}^{2^q-1} P\{\mathbf{0}\to \boldsymbol{l}\}P\{\mathbf{0}\to \boldsymbol{m}\}P\{\mathbf{0}\to \boldsymbol{l}A^2+\boldsymbol{m}BA\}\cdot$$
$$\cdot P\{\mathbf{0}\to \boldsymbol{m}A^2+\boldsymbol{l}B\}P\{\mathbf{0}\to \boldsymbol{l}BA+\boldsymbol{m}B\} \tag{21}$$

$$r_{33} = P(R_{33}) = \sum_{h=1}^{2^q-1}\sum_{r=1}^{2^q-1}\sum_{t=1}^{2^q-1} P\{\mathbf{0}\to\boldsymbol{h}\}P\{\mathbf{0}\to\boldsymbol{r}\}P\{\mathbf{0}\to\boldsymbol{t}\}\cdot$$
$$\cdot P\{\mathbf{0}\to\boldsymbol{h}A^2+\boldsymbol{r}BA+\boldsymbol{t}B\}P\{\mathbf{0}\to\boldsymbol{r}A^2+\boldsymbol{t}BA+\boldsymbol{h}B\}\cdot$$
$$\cdot P\{\mathbf{0}\to\boldsymbol{t}A^2+\boldsymbol{h}BA+\boldsymbol{r}B\} \quad (22)$$

$$r_{34} = P(R_{34}) = \sum_{h=1}^{2^q-1}\sum_{r=1}^{2^q-1}\sum_{t=1}^{2^q-1} P\{\mathbf{0}\to\boldsymbol{h}\}P\{\mathbf{0}\to\boldsymbol{r}\}P\{\mathbf{0}\to\boldsymbol{t}\}\cdot$$
$$\cdot P\{\mathbf{0}\to\boldsymbol{h}A^2+\boldsymbol{r}BA+\boldsymbol{t}B\}P\{\mathbf{0}\to\boldsymbol{r}A^2+\boldsymbol{t}BA\}\cdot$$
$$\cdot P\{\mathbf{0}\to\boldsymbol{t}A^2+\boldsymbol{h}B\}P\{\mathbf{0}\to\boldsymbol{h}BA+\boldsymbol{r}B\} \quad (23)$$

$$s_{24} = P(S_{24}) = \sum_{l=1}^{2^q-1}\sum_{m=1}^{2^q-1} P\{\mathbf{0}\to\boldsymbol{l}\}P\{\mathbf{0}\to\boldsymbol{m}\}P\{\mathbf{0}\to\boldsymbol{l}A^2+\boldsymbol{m}B\}\cdot$$
$$\cdot P\{\mathbf{0}\to\boldsymbol{m}BA\}P\{\mathbf{0}\to\boldsymbol{m}A^2+\boldsymbol{l}B\}P\{\mathbf{0}\to\boldsymbol{l}BA\} \quad (24)$$

$$s_{35} = P(S_{35}) = \sum_{h=1}^{2^q-1}\sum_{r=1}^{2^q-1}\sum_{t=1}^{2^q-1} P\{\mathbf{0}\to\boldsymbol{h}\}P\{\mathbf{0}\to\boldsymbol{r}\}P\{\mathbf{0}\to\boldsymbol{t}\}\cdot$$
$$\cdot P\{\mathbf{0}\to\boldsymbol{h}A^2+\boldsymbol{r}BA\}P\{\mathbf{0}\to\boldsymbol{r}A^2+\boldsymbol{t}B\}\cdot$$
$$\cdot P\{\mathbf{0}\to\boldsymbol{t}BA\}P\{\mathbf{0}\to\boldsymbol{t}A^2+\boldsymbol{h}B\}P\{\mathbf{0}\to\boldsymbol{h}BA+\boldsymbol{r}B\} \quad (25)$$

$$s_{36} = P(S_{36}) = \sum_{h=1}^{2^q-1}\sum_{r=1}^{2^q-1}\sum_{t=1}^{2^q-1} P\{\mathbf{0}\to\boldsymbol{h}\}P\{\mathbf{0}\to\boldsymbol{r}\}P\{\mathbf{0}\to\boldsymbol{t}\}\cdot$$
$$\cdot P\{\mathbf{0}\to\boldsymbol{h}A^2+\boldsymbol{r}B\}P\{\mathbf{0}\to\boldsymbol{r}BA\}P\{\mathbf{0}\to\boldsymbol{r}A^2+\boldsymbol{t}B\}\cdot$$
$$\cdot P\{\mathbf{0}\to\boldsymbol{t}BA\}P\{\mathbf{0}\to\boldsymbol{t}A^2+\boldsymbol{h}B\}P\{\mathbf{0}\to\boldsymbol{h}BA\} \quad (26)$$

Expressions (15) - (26) show that none of A_2, A_3, C_1, C_2, C_2', C_3, R_{23}, R_{33}, R_{34}, S_{24}, S_{35}, S_{36}, depend neither on the distribution of the characters in the input message nor on the matrix C. Thus, the theorem is proven.

Using the fact that the probability of undetected errors is independent from the distribution of the characters in the input message the following theorem holds (will be proven elsewhere):

Theorem 2. *Let $f(n,p)$ be the probability of undetected errors in a transmitted block with length n through the binary symmetric channel where p is the probability of incorrect transmission of a bit. If a linear quasigroup of order 2^q is used for coding, than the probability of undetected errors is given by the following formulas:*

$$f(3,p) = 3v_1v_0^2 + 3r_{23}v_0 + r_{33}$$
$$f(4,p) = 4v_1v_0^4 + 4v_2v_0^2 + 2s_{24}v_0^2 + 4r_{34}v_0$$
$$f(5,p) = 5v_1v_0^6 + 5v_2v_0^4 + 5c_1v_0^3 + 5v_3v_0^2 + 5s_{35}v_0^2$$
$$f(6,p) = 6v_1v_0^8 + 6v_2v_0^6 + 6c_1v_0^5 + 3v_1^2v_0^4 + 6v_3v_0^4 + 6c_2v_0^3 + 6c_2'v_0^3 + 2s_{36}v_0^3$$
$$f(7,p) = 7v_1v_0^{10} + 7v_2v_0^8 + 7c_1v_0^7 + 7v_1^2v_0^6 + 7v_3v_0^6 + 7c_2v_0^5 + 7c_2'v_0^5 + 7v_2v_1v_0^4 + 7c_3v_0^4$$
$$f(8,p) = 8v_1v_0^{12} + 8v_2v_0^{10} + 8c_1v_0^9 + 12v_1^2v_0^8 + 8v_3v_0^8 + 8c_2v_0^7 + 8c_2'v_0^7 + 16v_2v_1v_0^6 + 8c_3v_0^6 + 8c_1v_1v_0^5$$
$$f(n,p) = nv_1v_0^{2n-4} + nv_2v_0^{2n-6} + nc_1v_0^{2n-7} + \frac{n(n-5)}{2}v_1^2v_0^{2n-8} + nv_3v_0^{2n-8} + nc_2v_0^{2n-9} + nc_2'v_0^{2n-9} + n(n-6)v_2v_1v_0^{2n-10} + nc_3v_0^{2n-10} + n(n-7)c_1v_1v_0^{2n-11} + \frac{n(n^2-15n+56)}{6}v_1^3v_0^{2n-12}, \qquad \textit{for } n \geq 9.$$

where parameters v_1, v_2, v_3, c_1, c_2, c_2', c_3, r_{23}, r_{33}, r_{34}, s_{24}, s_{35}, s_{36} are the ones obtained in Theorem 1 and v_0 is the probability of correct transmission of a character.

In order to calculate the probability of undetected errors for a given linear quasigroup of order 2^q, first the probabilities v_1, v_2, v_3, c_1, c_2, c_2', c_3, r_{23}, r_{33}, r_{34}, s_{24}, s_{35} and s_{36} should be calculated using formulas (10), (15), (16), (17), (18), (19), (20), (21), (22), (23), (24), (25) and (26) and then to substitute these values in Theorem 2.

4 The Error-Detecting Code When Linear Quasigroups of Order 8 Are Used for Coding

4.1 The Probability of Undetected Errors

We considered the case when linear quasigroups of order 8 are used for coding. The goal is to find linear quasigroups of order 8 which give the smallest probability of undetected errors and the corresponding probability function. From Theorem 1 it follows that instead of working with all 225792 linear quasigroups of order 8, we should work only with all pairs of non-singular binary matrices of order 3×3. There are 168 non-singular binary matrices of order 3×3, which means that there are $168^2 = 28224$ pairs of non-singular binary matrices, i.e. 28224 linear quasigroups of order 8 for which the probability of undetected errors should be calculated. Using Theorem 1 and Theorem 2, we calculated the probability of undetected errors for these 28224 linear quasigroups of order 8 and obtained the smallest probability of undetected errors:

$$
\begin{aligned}
f(3,p) &= p^4(9 - 126p + 862p^2 - 3696p^3 + 10886p^4 - 22884p^5 + 35346p^6 \\
&\quad -40464p^7 + 34252p^8 - 21000p^9 + 8844p^{10} - 2256p^{11} + 231p^{12} \\
&\quad +18p^{13} - p^{14}) \\
f(4,p) &= 2p^5(1-p)^6(6 - 65p + 346p^2 - 1152p^3 + 2496p^4 - 2812p^5 - 2854p^6 \\
&\quad +22851p^7 - 63444p^8 + 118644p^9 - 166120p^{10} + 180049p^{11} - 152546p^{12} \\
&\quad +100669p^{13} - 50952p^{14} + 19204p^{15} - 5120p^{16} + 880p^{17} - 80p^{18} + 2p^{19}) \\
f(5,p) &= 5p^6(1-p)^{10}(2 - 19p + 86p^2 - 232p^3 + 372p^4 - 197p^5 - 679p^6 + 2212p^7 \\
&\quad -3596p^8 + 3824p^9 - 2759p^{10} + 1316p^{11} - 388p^{12} + 60p^{13} - p^{14}) \\
f(6,p) &= p^6(1-p)^{15}(12 - 150p + 918p^2 - 3568p^3 + 9696p^4 - 19056p^5 + 26960p^6 \\
&\quad -25800p^7 + 12498p^8 + 5642p^9 - 16491p^{10} + 15591p^{11} - 8782p^{12} + 3054p^{13} \\
&\quad -519p^{14} + 11p^{15}) \\
f(7,p) &= 7p^6(1-p)^{21}(2 - 25p + 153p^2 - 596p^3 + 1629p^4 - 3235p^5 + 4656p^6 - 4590p^7 \\
&\quad +2436p^8 + 640p^9 - 2563p^{10} + 2527p^{11} - 1482p^{12} + 570p^{13} - 111p^{14} + 3p^{15}) \\
f(8,p) &= 4p^6(1-p)^{26}(4 - 54p + 356p^2 - 1498p^3 + 4450p^4 - 9728p^5 + 15786p^6 \\
&\quad -18516p^7 + 14152p^8 - 3856p^9 - 5971p^{10} + 9790p^{11} - 7911p^{12} + 4268p^{13} \\
&\quad -1531p^{14} + 278p^{15} - 9p^{16}) \\
f(n,p) &= \frac{1}{6}np^6(1-p)^{6(n-4)} \times \Big[12 - 186p + 1404p^2 - 6792p^3 + 23406p^4 - 60378p^5 \\
&\quad +12(n + 9915)p^6 - 24(4n + 7445)p^7 + 6(76n + 32877)p^8 - 12(122n \\
&\quad +11693)p^9 + 3(1063n + 7389)p^{10} - (4572n - 90204)p^{11} + 2(4n^2 + 1923n \\
&\quad -65833)p^{12} - 120(11n - 835)p^{13} - 6(2n^2 + 165n + 7541)p^{14} + 72(23n \\
&\quad +133)p^{15} + 6(n^2 - 151n + 96)p^{16} + 24(7n - 19)p^{17} - (n^2 - 6n + 11)p^{18}\Big], \\
&\qquad \text{for } n \geq 9. \qquad (27)
\end{aligned}
$$

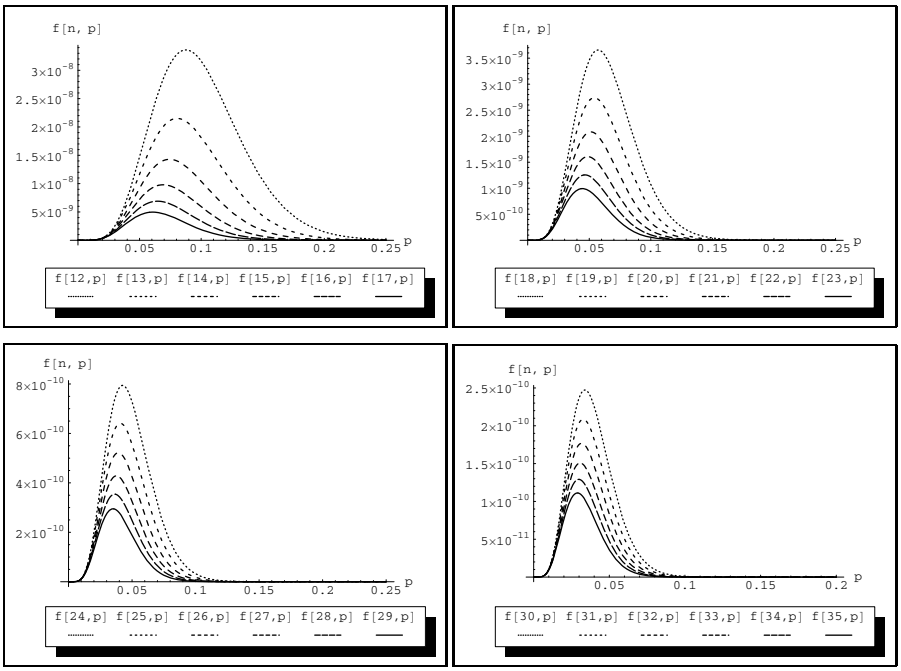

Fig. 1. The probability of undetected errors for the best class of linear quasigroups of order 8. Note that the scaling on y axis is different and is with different resolution.

The graphic of this function, for different values of the block length n, is given in the Figure 1.

4.2 Controlling the Probability of Undetected Errors

As we can see from Figure 1, the probability of undetected errors decreases as block length increases. The maximum of this function converges to zero when the block length tends to infinity. Using this property, we can control the probability of undetected errors. Namely, suppose that we want the probability of undetected errors to be smaller than ε. From the fact that the series of maximums of $f(n,p)$ converges to zero when $n \to \infty$, it follows that there is some $n_0 \in N$, such that for all $n \geq n_0$, the maximum of $f(n,p)$ is smaller then ε. Since for all $n \geq n_0$, the maximum of $f(n,p)$ is smaller than ε, it follows that for all $n \geq n_0$, the value $f(n,p)$ will be smaller than ε for all values of p. Now, we choose the block length n to be exactly n_0, we split the message into blocks with length n and encode each block separately. The probability of undetected errors will be smaller then ε. Practically, we choose the block length to be the smallest number n for which the maximum of the function $f(n,p)$ is smaller than ε.

4.3 The Best Class of Quasigroups of Order 8 for Coding

We define a class of quasigroups of order 8 by the criterium that they are giving the smallest probability of undetected errors. The following 6 pairs of non-singular

binary matrices A and B of order 3×3 define the quasigroups of order 8 which are best for coding:

$$\begin{array}{cc} A & B \\ \begin{bmatrix} 0&1&1 \\ 0&0&1 \\ 1&1&1 \end{bmatrix} & \begin{bmatrix} 0&1&1 \\ 1&1&0 \\ 1&1&1 \end{bmatrix} \\ \begin{bmatrix} 0&0&1 \\ 1&0&1 \\ 1&1&1 \end{bmatrix} & \begin{bmatrix} 1&1&0 \\ 1&0&1 \\ 1&1&1 \end{bmatrix} \end{array} \qquad \begin{array}{cc} A & B \\ \begin{bmatrix} 0&1&1 \\ 1&1&1 \\ 0&1&0 \end{bmatrix} & \begin{bmatrix} 0&1&1 \\ 1&1&1 \\ 1&0&1 \end{bmatrix} \\ \begin{bmatrix} 1&1&1 \\ 1&0&0 \\ 1&1&0 \end{bmatrix} & \begin{bmatrix} 1&1&1 \\ 0&1&1 \\ 1&1&0 \end{bmatrix} \end{array} \qquad \begin{array}{cc} A & B \\ \begin{bmatrix} 0&1&0 \\ 1&1&1 \\ 1&1&0 \end{bmatrix} & \begin{bmatrix} 1&0&1 \\ 1&1&1 \\ 1&1&0 \end{bmatrix} \\ \begin{bmatrix} 1&1&1 \\ 1&0&1 \\ 1&0&0 \end{bmatrix} & \begin{bmatrix} 1&1&1 \\ 1&0&1 \\ 0&1&1 \end{bmatrix} \end{array}$$

For any of these 6 pairs, any of the 8 binary matrices C of order 1×3 can be chosen. It follows that there are only 48 linear quasigroups of order 8 in this class of quasigroups.

5 Conclusion

In this paper we analyzed an error-detecting code based on linear quasigroups. We gave a proof that the probability of undetected errors is independent from the distribution of the characters in the input message. Using this property, we gave a formula for the probability of undetected errors if for the coding we use an arbitrary linear quasigroup of order 2^q. Additionally, we found the best class of linear quasigroups of order 8 for such coding and we computed the corresponding probability of undetected errors. Finally, we explained how the probability of undetected errors can be made arbitrary small.

Acknowledgments. This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss.Cyril and Methodius" University.

References

1. Bakeva, V., Ilievska, N.: A probabilistic model of error-detecting codes based on quasigroups. Quasigroups and Related Systems 17, 135–148 (2009)
2. Ilievska, N., Bakeva, V.: A Model of error-detecting codes based on quasigroups of order 4. In: Proc. Sixth International Confer. Informatics and Information Technology, Bitola, Republic of Macedonia, pp. 7–11 (2008)
3. Chen, Y., Niemenmma, M., Han Vinck, A.J., Gligoroski, D.: On the error detection capability of one check digit. IEEE Transactions on Information theory, 261–270 (2014)
4. The On-line Encyclopedia of Integer Sequences, `http://oeis.org/search?q=nonsingular+binary+matrices&sort=&language=&go=Search`
5. Gligoroski, D., Dimitrova, V., Markovski, S.: Quasigroups as Boolean functons, their equation systems and Gröbner bases. In: Mora, T., Perret, L., Sakata, S., Sala, M., Traverso, C. (eds.) Groebner, Coding, and Cryptography. short-note for RISC Book Series, pp. 415–420. Springer (2009)
6. Koscielny, C.: Generating quasigroups for cryptocraphic applications. Int. J. Appl. Math. Comput. Sci. 12, 559–569 (2002)
7. McKay, B.D., Wanless, I.M.: On the number of Latin squares. Annals of Combinatorics 9, 335–344 (2005)

Automatic Movie Posters Classification into Genres

Marina Ivasic-Kos, Miran Pobar, and Ivo Ipsic

Department of Informatics,
University of Rijeka,
Rijeka, Croatia
{marinai,mpobar,ivoi}@uniri.hr

Abstract. A person can quickly grasp the movie genre (drama, comedy, cartoons, etc.) from a poster, regardless of short observation time, clutter and variety of details. Bearing this in mind, it can be assumed that simple properties of a movie poster should play a significant role in automated detection of movie genres. Therefore, visual features based on colors and structural cues are extracted from poster images and used for poster classification into genres.

A single movie may belong to more than one genre (class), so the poster classification is a multi-label classification task. To solve the multi-label problem, three different types of classification methods were applied and described in this paper. These are: ML-kNN, RAKEL and Naïve Bayes. ML-kNN and RAKEL methods are directly used on multi-label data. For the Naïve Bayes the task is transformed into multiple single-label classifications. Obtained results are evaluated and compared on a poster dataset using different feature subsets. The dataset contains 6000 posters advertising films classified into 18 genres.

The paper gives insights into the properties of the discussed multi-label classification methods and their ability to determine movie genres from posters using low-level visual features.

Keywords: multi-label classification, data transformation method, movie poster.

1 Introduction

One of the goals of a poster is to convey information about a movie (genre, etc.) to potential moviegoers without them paying a lot of attention. With just a cursory glance at a poster while driving along or looking shortly while passing by, a person can grasp the movie genre (drama, comedy, cartoons, etc.) from variety of perceptual and semantic information on the poster. Taking this phenomenon [1] into account one can suppose that relevant information for determining the genre could be contained in global low-level features such as dominant color, spatial structure, color histogram, texture, etc.

Keeping this in mind our goal was to develop a method that would automatically determine the movie genres using mostly global low-level features of movie posters.

We used data from the TMDB [2] and realized that the problem we are dealing with is a multi-label problem since most of the movies belong to more than one genre.

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_32

For example, “Delivery Man” belongs to Comedy, "The Wolf of Wall Street“ belongs to Crime, Drama and Comedy genres and „The LEGO Movie“ belongs to Adventure, Fantasy, Animation, Comedy, Action and Family genres. The problem is even more complex as the number of possible genres is large and there is no limit to the number of genres a film can be classified into.

The issue of classifying a film into genres from their supporting promotional material (trailers) has recently attracted some attention. In the paper [3], low-level features are extracted from movie trailers and used to classify 100 movies into 4 genres (drama, action, comedy, horror). In [4] GIST, CENTRIST and W-CENTRIST scene features are obtained from a collection of temporally-ordered static key frames. These feature representations are used as visual vocabulary for genre classification and their discriminate ability is tested on 1239 movie trailers.

In [5] the same visual features were used as in [3]. Movies were classified into three genres (action, drama, and thriller) which were selected because of their frequency among movies that were played in Taiwan from 2004 to 2006. Some additional genres were grouped together and presented as those three (e.g. drama included comedy and romance while thriller included horror).

All these approaches [3-5] consider only a single genre per movie in order to reduce the problem to the single-label classification case and apply the classic methods for single-label classification.

However, many different approaches have lately been developed to solve multi-label classification problems. These methods were primarily focused on text classification (news, web pages, e-mails etc.), but lately there are more and more domains in which they are applied, such as functional genomics classification (gene and protein function), music and song categorization into moods and genres [6], scene classification [7], video annotations, poster classification [8], etc. Comparison of methods for multi-label learning is given in [9].

In our approach, we treat the poster classification into movie genres as a multi-label classification task.

In Section 2, two methods for multi-label problem adaptation are explained. Both methods were applied to the poster classification problem, in an experiment as detailed in Section 3. The obtained results are compared and presented in Section 4. The paper ends with a conclusion and directions for future work.

2 Adaptation of the Multi-label Problem

The aim of our work is to develop a method that will automatically provide a list of relevant labels (movie genres) for a given, previously unseen poster, based on extracted low-level features. A movie can belong to more than one genre; therefore the task of poster classification into movie genres is a multi-label classification problem.

Multi-label classification of an example e_j can be formally expressed as:

$$\exists\, e_j \in E : \varphi(e_j) = \{C_l, C_m\} \cup Z,\ Z \subseteq \bigcup_{i=1}^{k} C_i,\ l, m \in 1..k, l \neq m, \varphi: E \to C, \quad (1)$$

where E is a set of samples, C is a set of class labels and function φ is a classifier so that exists at least one example e_j that is mapped into two or more classes C_l and C_m.

Methods most commonly used to tackle a multi-label classification problem can be divided into two different approaches [10]. These are problem transformation methods (referred as P1 in the following) and algorithm adaptation methods (referred to as P2 in the following).

In the P1 multi-label transformation approach, the multi-label classification problem is transformed into more single-label classification problems [11]. In the single-label classification problem an example e_j is classified to a single class label from the set of C. The aim is to transform the data so that any classification method, designed for single-label classification, can be applied. We applied the P1 approach to transform the multi-label problem into single-label problems in two ways.

In the first case, binary relevance method [10] is applied, referred to as P1.1. One binary classifier is independently trained for each label. Each classifier then decides whether to assign one class label to an example or not. The overall classification result contains all class labels assigned to that instance. Therefore, each instance with multiple labels was transformed into a set of ordered pairs so that the first element of each pair is the instance and the second one is the class label. Thus, if an instance $e_j \in E$ can be classified into $Y_j = \{C_l, C_m, \dots, C_r\}$, $Y_j \subseteq C$ then that instance is replaced with $|Y_j|$ ordered pairs $(e_j, C_l), (e_j, C_m) \dots (e_j, C_r)$. For example, the movie „Non-Stop“ belongs to Action, Thriller and Mystery genres, and would be transformed into the set of ordered pairs containing individual genres: {(Non-Stop, Action), (Non-Stop, Thriller), (Non-Stop, Mystery)}.

In the second case, a label power-set method is used to create one classifier for every possible label combination (referred to as P1.2). That is, entire label set $\mathrm{Y_j} = \{\mathrm{C_l}, \mathrm{C_m}, \dots, \mathrm{C_r}\}$ is transformed into a new combined class $\mathrm{C_{l,m,\dots,r}}$ that is assigned to the example $\mathrm{e_j}$. In that way a set C is expanded with new combined classes into the set C', so that applies $\mathrm{C}' \supseteq \mathrm{C}$. Using this problem transformation method “Non-Stop” example would be transformed into the ordered pair containing one combined genre (Non-Stop, Action&Thriller&Mystery).

In both cases, a standard classification algorithm can be applied to assign a first element of the ordered pair (instance) to a second element of the pair (class label). On the other hand, algorithm adaptation methods (P2) extend specific learning algorithms in order to handle multi-label data directly.

3 Experiments

The experiments performed on the poster dataset obtained from the TMDB are presented below. The aim is to provide a list of relevant labels for the unknown poster using low-level features.

3.1 Data and Preprocessing Step

Our dataset consists of 6739 movie posters dated from 1990 onwards. We have selected 18 genres (Action, Adventure, Animation, Comedy, Crime, Disaster, Documentary, Drama, Fantasy, History, Horror, Mystery, Romance, Science Fiction, Suspense, Thriller, War and Western) and for each we picked 20 most popular movies for each year in the range. The total number of movies was smaller, because some movies were among the most popular in more than one genre.

Also, since each movie can have multiple genres, additional genre labels (e.g. Film Noir, Indie and Sport) were present in the data, so the total number of genres was 35.

The maximum number of films with a certain label is 2610, but one genre had only one instance which is not enough data to define a model for that class.

To prevent data scarcity for rare genres, we transformed the data in two ways. In the first, we joined the genres with few examples with similar genres according to our judgment, such as Neo-Noir with Crime and Road Movie with Adventure. We also joined some genres that commonly appeared together, e.g. Mystery and Crime with Thriller. After this transformation, the number of genres was reduced to 11. We refer to this data set as JG. In the second case we have simply discarded the additional genres in the data that were not among the selected 18. We refer to this data set as DG.

The resulting data distribution is more suitable for learning of classification models because sufficient examples per most genres are obtained (approximately more than 1000 posters per genre), although the width of the classes are rather uneven (std. dev is 631 in case with discarded genres and even less favorable, std. dev equals 916, in case with joined genres). The more detailed statistics of data before and after transformation is presented in the Table 1.

Table 1. Original and transformed data set statistics

Statistic	*Original data*	*Transformed data*	
		joined genres (JG)	*discarded genres (DG)*
No. of classes in set C	35	11	18
Max examples per genre	2610	3209	2610
Min examples per genre	1	279	64
Mean examples per genre	558	1497	982
Std. dev. per genre	641	916	631

Data transformed in such way was directly handled by methods proposed in approach P2. For the P1 approach an additional pre-processing step was needed to enable the use of single label classification methods. In the case of the P1.1 approach the subset of multi-label data is used more than once, in fact as many times as there are labels into which the poster is classified. In the case of P1.2 approach, for each multi-label set a new combined class was created, so the set C' is built. However, with this approach a significant problem of sparse classes appears. The relation among the number of class labels in the original sets and the number of class labels contained in the set C' is presented in the Table 2 for the cases of original and transformed data sets.

Table 2. Number of classes in original and transformed data sets when a label power-set method is used

Statistic	*Original data*	*Transformed data*	
		joined genres (JG)	*discarded genres (DG)*
No. classes in set C'	1480	387	891
Max examples per genre	417	605	607
Min examples per genre	1	1	1
Mean examples per genre	4.6	17.41	7.5
Std. dev. per genre	14.87	47.88	26.04

Considering the statistics presented in the Table 2, problem of data scarcity becomes even more obvious. In all cases many new classes are formed (e.g. 387 vs. 11 in case of joined genres, 891 vs. 18 in case of discarded genres) with a small number of examples, e.g. in cases of joined genres 44.44% genres have less than 3 elements. Due to an insufficient number of examples per most genres this kind of data transformation is not used for learning the classification model. The solution could be to form new genres that will include mostly triplets of genres but such reduction of set is not applied here.

3.2 Features

Motivated by the way people capture relevant information about the movie with just a glance at billboards we wanted to examine if low-level features that can be easily noticed on the poster, such as dominant colors and structure, have discriminative ability in terms of genre classification.

Before the low-level features were extracted, each poster was proportionally sized to so that it is 100 pixels wide and converted to HSV color space. Then, the image color histogram was calculated on hue (H), saturation (S) and value (V) channels of the whole image. Subsequently, histogram bins with the highest values for each channel are selected. Obtained features correspond to dominant colors (referred to as DC). We have experimentally tested different numbers of dominant colors (3, 6, 8, 12, 16, 24 and 36) and have determined that in our task 12 dominant colors per channel yield the best classification results. Thus, 36-dimensional DC vectors were used (12 dominant colors per three channels).

To preserve the information about the color layout of a poster, we have computed 5 local HSV histograms from which 12 dominant colors per channel were selected. These are referred to as DC1 to DC5, with total size of 180. DC1, DC2 and DC3 were computed from 3x1 grids, DC4 was computed on the central part of the poster that would probably contain the object and DC5 on the surrounding part that would probably contain the background. The central part was of the same proportions as the whole image, but 1/4 of the diagonal size. The arrangement of image grids from which the local dominant colors were computed is given in Fig. 1.

Fig. 1. The arrangement of image grids from which the local color histograms were computed

Additionally, we have computed the statistics and color moments (CM) for each HSV channel, such as mean, standard deviation, skew and kurtosis. The size of CM feature vector is 12.

Also, the GIST image descriptor, available at [12], was used. It is a structure-based image descriptor created for recognition of similar scenes, like mountains, streets, etc. This descriptor refers to the dominant spatial structure of the scene characterized by properties of its boundaries (e.g., the size, degree of openness, perspective) and its content (e.g., naturalness, roughness) [13]. Spatial properties are estimated using global features computed as a weighted combination of Gabor-like multi scale-oriented filters. The dimension of GIST descriptor is n x n x k where n x n is the number of samples used for encoding and k is the number of different orientation and scales of image components. GIST descriptor of each genre is implemented with 8x8 encoding samples obtained by projecting the averaged output filter frequency within 8 orientations per 8 scales. The size of GIST feature vector is 512.

3.3 Classification Methods

We have used three classification methods for classifying unknown posters into movie genres. Naïve Bayes classifier was used on data adapted according to the P1.1 approach described above. RAKEL [14], a kind of problem transformation method (P1) and ML-kNN [15], an algorithm adaptation method (P2), can be directly used on multi-label data. All methods are tested with both transformed data sets JG and DG (with joined or with discarded genres).

When using the Naïve Bayes (NB), binary relevance method was used and a single NB classifier was trained per each genre to distinguish that genre from all other genres. To classify an unseen poster sample all NB classifiers are applied.

RAKEL (random k-label sets) was run using the nearest neighbor (1-NN) classifier based on the Bhattacharyya distance (2) as base classifier,

$$d_B(p,p') = \sum_{i=1}^{N} \sqrt{p(i)p'(i)}, \tag{2}$$

where p and p' are histograms and N is the number of bins. The RAKEL subset size was 3 and the number of models was 12.

ML-kNN is a variant of the lazy learning algorithm derived from the traditional k-Nearest Neighbor (kNN) algorithm. For each unseen poster instance, its k nearest neighbors are firstly identified in the training set. Then, based on the statistical

information gained from the genre label sets of these neighboring instances, maximum a posteriori (MAP) principle is utilized to determine the genre set for the unseen poster instance.

4 Experimental Results

From 6739 collected posters, 80% were used for training and 20% for testing. We have used the feature set that includes GIST features, dominant color (DC) features, local dominant color features (DC1 to DC5) and color moments (CM). The size of feature vector is 740. Since the size of feature vector is large in proportion to the number of posters, we have also tested the classification performance using five subsets.

We have used accuracy, precision, recall and F1 score as instance-based and label-based evaluation measures. The instance-based evaluation measures are based on the average differences of the actual and the predicted sets of genres over all posters in the test dataset. The label-based evaluation measures assess the predictive performance for each genre separately and then average the performance over all genres [10].

The Table 3 shows the label-based evaluation results obtained using NB on data transformed with joined genres (JG) and on data transformed by discarding additional genre labels beyond the selected 18 (DG), for different feature subsets. The results obtained using all features are only slightly better than with other subsets with much smaller number of features for both data transformation methods JG and DG, and accuracy was even better when using only the feature subset DC+CM.

Thus, in further experiments, we only tested the subset with the smallest number of features (DC+CM) that performed similarly and the set with all features. The idea was to keep the number of features low enough to emphasize the information that is relevant for classification and adequately train the classifier.

The results obtained with the DG data transformation are significantly lower than with JG method, which is not surprising due to much larger number of classes.

Table 3. Label (genre) based evaluation results for datasets JG with joined genres (11) and dataset DG with 18 genres, using NB

Label-based evaluation measure	***All features***		***GIST***		***DC+CM***		***DC1..DC5 + CM***		***DC + DC1..DC5 + CM***	
	JG	**DG**	**JG**	**DG**	**JG**	**DG**	**JG**	**DG**	**JG**	**DG**
Accuracy	0.62	0.62	0.63	0.65	**0.65**	**0.70**	0.61	0.59	0.61	0.59
Precision	**0.30**	**0.21**	0.29	**0.21**	0.28	**0.21**	0.28	0.19	0.28	0.19
Recall	**0.61**	**0.60**	0.57	0.54	0.48	0.39	0.56	0.55	0.56	0.56
F1 score	**0.38**	**0.29**	0.37	**0.29**	0.34	0.21	0.36	0.27	0.36	0.27

Instance-based classification results obtained using NB are presented in the Table 4 with both data transformation methods, for different feature subsets. Instance-based results are lower than genre based results for all evaluation measures. Also, all feature

subsets perform similarly with respect to F1 score. This suggests that most of the features are interdependent. The F1 score is a measure of classification accuracy that considers both precision and recall. It can be interpreted as a weighted average of the precision and recall. The F1 score reaches its best value at 1 and worst score at 0.

Table 4. Instance (movie poster) based evaluation results for datasets JG and DG, using NB

Instance-based evaluation measure	***All features***		***GIST***		***DC+CM***		***DC1..DC5 + CM***		***DC + DC1..DC5 + CM***	
	JG	**DG**	**JG**	**DG**	**JG**	**DG**	**JG**	**DG**	**JG**	**DG**
Accuracy	0.55	0.57	0.56	0.60	**0.61**	**0.68**	0.57	0.56	0.57	0.56
Precision	**0.25**	-	**0.25**	**0.17**	-	-	-	-	-	-
Recall	0.47	0.44	0.45	0.41	0.44	0.33	**0.49**	**0.46**	**0.49**	**0.46**
F1 score	0.31	**0.23**	0.31	0.23	**0.32**	0.21	**0.32**	**0.23**	**0.32**	**0.23**

Label-based classification results with RAKEL and ML-kNN are presented in Table 5. Results with RAKEL are significantly better for all evaluation measures than with ML-kNN, but actually slightly worse than with NB. As with NB, DC+CM feature set performs similarly as all features for RAKEL.

Table 5. Label based evaluation results for datasets JG and DG, using RAKEL and ML-kNN

Label-based evaluation measure	***RAKEL***				***ML-kNN***			
	All features		***DC+CM***		***All features***		***DC+CM***	
	JG	**DG**	**JG**	**DG**	**JG**	**DG**	**JG**	**DG**
Precision	0.33	0.26	0.33	0.31	0.53	0.52	0.44	0.57
Recall	0.32	0.25	0.32	0.31	0.06	0.03	0.03	0.12
F1 score	0.33	0.26	0.32	0.31	0.1	0.05	0.05	0.20

The results for instance based evaluation, shown in Table 6, show similar relationships between classification methods, feature sets and data transformation methods as for genre based evaluation.

Table 6. Instance (movie poster) based evaluation results for datasets JG and DG, using RAKEL and ML-kNN

Instance-based evaluation measure	***RAKEL***				***ML-kNN***			
	All features		***DC+CM***		***All features***		***DC+CM***	
	JG	**DG**	**JG**	**DG**	**JG**	**DG**	**JG**	**DG**
Accuracy	0.23	0.19	0.22	0.24	0.06	0.07	0.03	0.15
Precision	0.33	0.26	0.32	0.32	0.13	0.09	0.07	0.23
Recall	0.33	0.25	0.32	0.32	0.06	0.07	0.03	0.16
F1 score	0.31	0.24	0.30	0.30	0.08	0.07	0.04	0.17

Overall the best results are obtained using the Naive Bayes classification algorithm with all features on data transformed with joined genres, however only slightly worse performance was observed with only a small number of color-based features DC+CM.

5 Conclusion and Future Work

In this paper, automated detection of movie genres from posters was modeled as a multi-label classification task, where a single movie may belong to more than one genre. The experiment was conducted on a dataset containing 6739 movie posters, classified into one or more of 18 genres. Since some genres had too few examples to effectively train the classifier, the performance of classification was compared with the same dataset where some genre labels were merged, yielding 11 genres.

As the usual single-label classification algorithms can't directly be used to solve the multi-label problem, either the problem or the algorithms must be adapted in some way. Two different methods for problem transformation were applied and use of appropriated classifiers is described in this paper. These are: ML-kNN, Naïve Bayes and RAKEL. ML-kNN and RAKEL methods are directly used on multi-label data. For the Naïve Bayes the task is transformed into multiple single-label classifications. The features used in the classification were low-level features based on color histograms and color moments combined with the GIST descriptor. Obtained results are evaluated and compared on a poster dataset using different subsets of color and structural features.

The best result considering the F1 score was about 0.38 for the case of Naive Bayes classifier on the complete feature set and for 11 genres. Reducing the number of features from 740 to only 48 features related to the dominant colors of the HSV histogram didn't significantly impact the results, yielding the F1 score of 0.34. This suggests that few dominant colors indeed carry discriminative part of information about the movie genres, and that other tested features might be largely interdependent.

In the future work, we plan to test the dense SURF [16], other visual features used for scene representation [13] as well as features for text recognition. We also plan to test the classification on a much larger dataset and with different classification methods.

Also, a subjective test will be conducted to determine human ability to detect genres from poster images, and the results will be used for comparison with automatic detection.

References

1. Potter, M.C.: Short-term conceptual memory for pictures. Journal of Experimental Psychology: Human Learning and Memory 2(5), 509 (1976)
2. The movie database, `http://www.themoviedb.org/`

3. Rasheed, Z., Sheikh, Y., Shah, M.: On the use of computable features for film classification. IEEE Transactions on Circuits and Systems for Video Technology 15(1), 52–64 (2005)
4. Zhou, H., Hermans, T., Karandikar, A.V., Rehg, J.M.: Movie genre classification via scene categorization. In: ACM Proceedings of the International Conference on Multimedia, pp. 747–750 (2010)
5. Huang, H.-Y., Shih, W.-S., Hsu, W.-H.: A film classifier based on low-level visual features. In: IEEE 9th Workshop on Multimedia Signal Processing, MMSP 2007, pp. 465–468 (2007)
6. Allmusic, http://www.allmusic.com/
7. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition 37(9), 1757–1771 (2004)
8. Ivašić-Kos, M., Pobar, M., Mikec, L.: Movie Posters Classification into Genres Based on Low-level Features. In: IEEE Proceedings of International Conference MIPRO, pp. 1148–1153 (2014)
9. Madjarov, G., Kocev, D., Gjorgjevikj, D., Džeroski, S.: An extensive experimental comparison of methods for multi-label learning. Pattern Recognition 45(9) (2012)
10. Tsoumakas, G., Katakis, I.: Multi-Label Classification: An Overview. International Journal of Data Warehousing & Mining 3(3) (2007)
11. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier Chains for Multi-label Classification. Machine Learning Journal 85(3) (2011)
12. GIST, http://people.csail.mit.edu/torralba/code/spatialenvelope/
13. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
14. Tsoumakas, G., Vlahavas, I.: Random k-label sets: An ensemble method for multi-label classification. In: Machine Learning: ECML, pp. 406–417. Springer (2007)
15. Zhang, M.L., Zhou, Z.H.: ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40(7), 2038–2048 (2007)
16. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)

Migraine Diagnosis Support System Based on Classifier Ensemble

Konrad Jackowski[1,2], Dariusz Jankowski[1], Dragan Simić[3], and Svetlana Simić[4]

[1] IT4Innovations, VSB – Technical University of Ostrava,
17. listopadu 15/2172, 708 33 Ostrava - Poruba, Czech Republic
[2] Wroclaw University of Technology, Department of Systems and Computer Networks,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{konrad.jackowski,dariusz.jackowski}@pwr.edu.pl
[3] University of Novi Sad, Faculty of Technical Sciences,
Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
drdragansimic@gmail.com
[4] University of Novi Sad, Faculty of Medicine,
Hajduk Veljkova 1-9, 21000 Novi Sad, Serbia

Abstract. A valid diagnosis of migraine is a non-trivial decision problem. This is due to the fact that migraine can manifest wide range of varied symptoms. Thus, designing a computer aided diagnosis system for that problem remains still a very interesting topic. In this paper we present an ensemble classifier system designed for headache diagnosis. We assumed that the system should make fast initial diagnosis based on an analysis of data collected in the questionnaire only. Such an assumption eliminated possibility of application of most classical classification algorithms as they could not obtain decent level of accuracy. Therefore, we decided to apply an ensemble solution. Although it is clear that ensemble should consists of complementary classifiers, there is no guidance on how to choose ensemble size and ensure its diversity. Thus, we applied two stages strategy. Firstly, large pool of elementary classifiers were prepared. Its diversity was ensured by selecting algorithms of different types, structures, and learning algorithms. Secondly, we determined optimal size of the ensemble and selected its constituents using exhaustive search approaches. Results of experiments, which were carried on dataset collected in University of Novi Sad, shows that proposed system significantly outperformed all classical methods. Additionally we present analysis of diversity and accuracy correlation for tested systems.

Keywords: ensemble classifier systems, medical diagnosis support system, headache diagnosis.

1 Introduction

Machine learning algorithm are widely used in areas where traditional programming methods do not allow us to design effective solutions. This usually applies to tasks related with designing decision-making systems, where it is not possible to define

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_33

decision rules explicitly, or it is extremely difficult. The advantage of machine learning algorithms due to the fact that these systems are able to automatically create decision rules or decision algorithm. Therefore, it should not be surprising that up to 21% of research publications on practical application of machine learning algorithms applies to medicine [1].

The first crucial question is which algorithm should be chosen. There are a plethora of options such as Bayesian classifiers, neural networks or decision trees to mention just a few. Each of them has different characteristics that make it more or less appropriate for a given task. Many of classical algorithms are not able to effectively model the problem in hands. In this case, it may be helpful to use an ensemble classifier system. Their functioning relies upon collective decision-making by a committee composed from a number of elementary classifiers. It has been proved that, in certain cases, such an approach allows for a significant improvement of decision accuracy. Unfortunately, there are no clear rules that answer a question how to effectively create a committee.

In this article we present results of our work on computer aided decision support system dedicated to headache diagnosis. Unfortunately, the precise diagnosis of the headache type is very complex and usually imprecise, thus a high quality classification system is desirable diagnostic tool.

While designing our classification algorithm we made the following three assumptions. Decision of the system should be made base on analysis of patients' answers for several questions gathered in questionnaire. This assumptions, along with fact that there are 11 predefined migraine classes, makes application of classical classification algorithm very difficult. According to our previous experiments, they cannot obtain satisfying level of accuracy. Therefore, we decided to apply ensemble approach in our algorithm which consists of two phases. Firstly, a large pool of elementary classifiers is created. In order to maintain their diversity we used several different classifiers, i.e. such that have different structure, training algorithm, and decision making methods. All of them are trained to recognize 11 predefined classes of migraine. In the second phase, ensemble classifier system is created based on selected from the pool subset of classifiers. A size of the ensemble and its constituents are determined by exhaustive search manner, i.e each possible combination of classifiers are evaluated based on available learning set gathered by the team from the University of Novi Sad [2].

The content of the work is as follows. Section 2 introduces the medical problem.

In the next section, we describe a mathematical model ensemble classifier system. Then we present the experimental evaluation of the selected classifiers for the problem under consideration. The last section concludes the paper.

2 Headache Diagnosis

Headache or cephalalgia is defined as a continuous pain in the head or neck region. The brain itself has no pain receptors. The pain originates from the tissues and pain-

sensitive structures surrounding them. Treatment of a headache depends on many non-specific symptoms.

Our knowledge and understanding of headache is still growing. New diagnosis systems and treatments are available for headache disorders. Most popular classification system of all headaches is organized by the International Headache Society, and published in the International Classification of Headache Disorders (ICHD). The current version, the ICHD-2, was published in 2004. This classification system is accepted by the World Health Organization (WHO) [3] and become a standard for headache diagnosis and clinical research. Headaches are classified by the ICHD-2 into two broad categories: the primary headache disorders (without organic cause) and the secondary headache disorder (where etiological cause can be determined). Other classification systems also exist [4,5].

Approximately 90% of people have a headache at some point in their lives. The most common among the general population are tension-type headache (20.8%) and migraine (15%) [6]. Only a very small percent of the population have secondary headaches. Most secondary headaches can be easily diagnosed, while tension-type headache and migraine recognition can be a problematic even for an experience physician.

Tension-type headache (TTH) is the most common type of primary headache. The pain can radiate from forehead to the occiput. Often described as a band-like nonpulsatile ache or tightness in frontal, temporal and occipital regions. TTH can be episodic or chronic. Episodic tension-type headaches occurring fewer than 15 days a month, whereas chronic tension headaches occur 15 days or more in a month for at least 6 months. Various triggers may cause tension-type headache: stress, sleep deprivation, hunger, uncomfortable position, bad posture or eyestrain. Episodic TTH generally respond well to popular medicines such as ibuprofen, paracetamol and aspirin. Analgesic drugs for chronic TTH are amitriptyline, topiramate or sodium valproate. Tension headaches are more common in women than men (23% to 18% respectively) [6].

Migraine is disorder characterized by repeated from moderate to severe attacks of headache. Typically the pain is placed on one half of the head and pulsating in nature. Associated symptoms may include nausea and vomiting, increased sensitivity to sound/light/odors and pain, which is worsened by any physical activity. The two common forms of migraine are called migraine without aura, and migraine with aura. An aura is accompanied by visual and/or sensory and/or speech symptoms. Aura appear gradually and takes no longer than one hour. Migraine without aura, or "common migraine", involves migraine headaches that are not accompanied by visual disturbances. The underlying causes of migraines are unknown. Migraines becomes more common among women and have their own specifics. Research conducted by MacGregor et al. [7] proved that migraine in women can be divided to: menstrual and non-menstrual. Additionally there are two types of menstrual migraine: pure menstrual migraine and menstrually-related migraine.

3 Ensemble Classifier System Model

In machine learning algorithms [8] it is assumed that the model of a decision making is created based on empirical data collected in the form of learning set (1). It consists of set of pairs, i.e. set of attributes (results of medical examinations) denoted by x, which describe the objects (patient), and corresponding class label j (medical diagnosis given by an expert).

$$LS=\{(x_1, j_1),(x_2, j_2),\ldots,(x_N, j_N)\}, \tag{1}$$

A classification algorithm Ψ assigns object to one of M predefined classes.

$$\Psi : X \rightarrow \mathbf{M} \tag{2}$$

The quality of the system is usually assessed by calculating its accuracy, i.e. a fraction of correctly classified object from testing set.

3.1 Ensemble of Classifiers

As it was stated before, there are many classification algorithms which can be chosen for creating decision making system. Nonetheless, there are very few hints how to make the selection. Therefore, common approach is to test several options and chose the best one. Nevertheless, in practice it might happened that the performance of the best classifier still does not meet our expectation. One can point out many reasons: (1) Insufficient and not representative learning set; (2) Presence of noise which spoils the data; (3) Too high complexity of the problem, which cannot be approximated by simple classifiers; (4) Not efficient learning algorithm.

In this case, it may be helpful to use an ensemble classifier system [8], i.e. algorithm in which a decision is made collectively by committee of elementary classifiers (3).

$$\Pi^{\Psi} = \{\Psi_1, \Psi_2, \ldots, \Psi_K\}, \tag{3}$$

where Ψ_k is k-th elementary classifier in the set.

There are many variants of ensemble decision making formula. Most popular one is Majority Voting, which is based on counting votes casted by classifiers for each class [9]. More advanced methods fuse the voices of classifiers on a level of their discriminating function [10], i.e. the function which represent the support of the classifier for each class. In this case, aggregating strategy of discriminating function can be used such as: sum, maximum, and average. We use simple summary operator what leads to the following formula for ensemble decision making (4).

$$\overline{\Psi}(x) = \underset{i=1}{\overset{M}{\arg\max}} \sum_{k=1}^{K} d_{k,i}(x), \tag{4}$$

where $d_{k,i}(x)$ is an support given by k-th classifier for i-th class, and $\bar{\Psi}$ is ensemble.

Although there exists even more sophisticated and effective methods of fusion such as weighted fusion [11], we decided to apply simple aggregating algorithm because it does not require any additional training for setting the weights. This is essential features because exhaustive search is applied for searching optimal subset of classifiers.

3.2 Ensemble Diversity

The committee should be filled by possibly different components. There are several ways to enforce the diversity, among them: (1) Train each individual classifier to recognize a subset of selected classes and then choose a fusion method that recovers the whole set of classes; (2) Train individual classifiers based on different classifier models; (3) Differentiate elementary classifier inputs.

In our approach we decided to use the second and third options. The former one was apply explicitly by creating pool consisting of 11 different classifiers. The last one was applied implicitly as we used backward feature selection algorithm for each classifier separately.

The other question is how to measure the diversity of the ensemble. In [12] selected proposition can be found. Most of them were inspired by methodology of designing reliable software systems [13]. For our purposes, the entropy measure was used (5).

$$E = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{(M-\left(\frac{M}{2}\right)-1} min\left\{\sum_{k=1}^{K} y_{j,k}, M - \sum_{k=1}^{K} y_{j,k}\right\}, \quad (5)$$

where $y_{j,m}$ is an variable which can takes two values: 1 when k-th classifier misclassifies j-th object, and 0 otherwise.

Entropy measure can be easily interpret. It varies between 0 and 1, where 1 indicates the highest diversity and 0 indicates no difference among the classifiers.

In our experiments apart from evaluating ensemble accuracy we examined the entropy measure to evaluate correlation between the two.

3.3 Selecting Ensemble Members

Having a large pool of classifiers does not guarantee achieving the highest accuracy of classification because it is impossible to predict how classifiers complement each other. It is likely, that the smaller subset of selected classifiers can perform much better then entire pool. While there are many heuristic approaches of classifier selection, we decided to apply in our algorithm exhaustive search of optimal, i.e. all possible subsets of classifiers were evaluated in term of their accuracy and diversity. This approach, although time consuming, has several advantages. Firstly, we are sure that we select the best possible subset. Secondly, we are able to analyse a relation between the accuracy, diversity, and size of the committee.

4 Experiments

In this section we are going to present results of experimental analysis of proposed algorithm. There are objectives of the analysis: (1) Verification if ensemble approach allows to improve classification accuracy and outperform simple classification algorithms; (2) Examination how size of ensemble and its diversity affect the accuracy of the system; (3) Examination which classical algorithms gathered in the ensemble allows to create the most effective ensemble for problem in hand.

4.1 Empirical Material

Our research were conducted on 579 patients (in an age between 20 to 67) on the area of Novi Sad (Republic of Serbia). Table 1 presents a distribution of the diagnosis in 11 previously described migraine classes.

Table 1. Data distribution among the migraine types

Class Id	Migraine type	# samples
1	Migraine without aura in men	16
2	Migraine without aura and pure menstrual migraine	7
3	Migraine without aura and menstrual related migraine	49
4	Migraine without aura and non-menstrual migraine	31
5	Migraine with aura in men	5
6	Migraine with aura and menstrual related migraine	35
7	Migraine with aura and non-menstrual migraine	26
8	Rare episodic tension type headache	116
9	Frequent episodic tension type headache	99
10	Chronic tension type headache	9
11	Other headache type	186

Classification was made based on attributes formed based on responses collected in the questionaries' and originally consisted of 30 questions. Provided answers can be one of three types: true/false, select one of the answers or typing the appropriate value. All the answers were converted into numerical values for classification. The full version of the questionnaire is available on-line (http://www.kssk.pwr.wroc.pl/wp-content/uploads/downloads/2014/02/QUESTIONARE_headache.pdf).

4.2 Experimental Framework

All experiments were carried on in KNIME framework using as a base classifiers implemented in Weka. The pool of available classifiers consisted of 11 algorithms,

namely: (BFTree) best first decision tree, (IBk) implementation of k-Near Neighbor, (J48) implementation of C4.5 decision tree, (LAD) implementation of decision tree using the LogitBoost strategy, (LWL) localy weighted instance, (MLP) multilayer perceptron, (NB) naïve Bayes, (NBTree) decision tree with naive Bayes classifiers at the leaves, (RandTree) random decision tree, (RBF) radial basis neural network, (SVM) support vector machine,

In order to ensure the reliability of the tests all the experiments was carried on using 10 fold cross validation methods.

4.3 Results

Table 2 presents classification results of elementary classifiers collected in the pool ordered according to mean accuracy obtained over 10 repetition of cross validation procedure.

Table 2. Classification accuracy of elementary classifier in the pool

	Accuracy	
Classifier	**Mean**	**StdDev**
LADTree	0,721	0,026
NBTree	0,703	0,033
SVM	0,700	0,018
NB	0,698	0,020
BFTree	0,675	0,034
J48	0,674	0,024
RBF	0,636	0,043
RandTree	0,577	0,033
MLP	0,572	0,023
LWL	0,553	0,021
IBk(3)	0,546	0,023
Min	0,546	
Mean	0,641	
Max	0,721	

Classification accuracy of tested algorithms varies significantly in range from 54% to 72%. It shows that selecting algorithms for given decision problem in hand is not easy task as we cannot predict in advance which algorithm would be the most appropriate without having extensive knowledge on characteristic of data.

Experimental evaluation of available option can help to select the single best options if its quality gains acceptable level.

In our experiment the highest effectiveness was reached by LAD decision tree. Accuracy on the level of 72% of correct classification would be assessed as acceptable considering that our classification problems consists of 11 classes and decision is made based on analysis of simple questionnaires without any additional medical examinations.

In next experiments we investigated how application of ensemble would help to improve the results. Authors put questions: how to choose the size of the ensemble and how to select ensemble members. Table 3 presents average accuracy and entropy vs committee size.

Table 3. Classification accuracy and entropy of ensemble vs committee size

Committee size	Accuracy		Entropy		No. of combinations
	Mean	StdDev	Mean	StdDev	
1	0,641	0,067	0,000	0,000	11
2	0,690	0,038	0,252	0,058	55
3	0,710	0,023	0,252	0,036	165
4	0,721	0,017	0,297	0,033	330
5	0,728	0,013	0,297	0,026	462
6	0,732	0,011	0,316	0,023	462
7	0,734	0,009	0,316	0,018	330
8	0,737	0,008	0,326	0,015	165
9	0,739	0,006	0,326	0,011	55
10	0,741	0,005	0,332	0,008	11
11	0,741	0,000	0,332	0,000	1

The best committee consisting of 11 classifiers allows to elevate classification accuracy on 10 percent points what is very encouraging result.

The accuracy increases along with committee size. Additionally, analysis of the entropy which also is proportional to the size, seems to confirm conclusion that the larger the committee is the more diversified and complementary knowledge is collected in the ensemble. Nonetheless, it has to be remembered, that in the Table 3. we present average value for different combination of the elementary classifiers. For instance, there are 462 combinations of committee in the case of ensemble consisting of 5 members. Therefore closer analysis must be done on the level of each combination. Of course it is not possible to report results for each tested combinations as there are 2047 of them.

Therefore, in the Table 4 we present scores for best ensemble in each size group. Now, we can observe that there is no straight relation between the size of the ensemble and its accuracy. The best accuracy was gain by committee consisting 4 and 5 members. To understand this phenomena we have to look closer inside the ensembles. Two winners consisted of LADTree, NB, NBTree, RBFNetwork and LADTree, LWL, NB, NBTree, RBFNetwork classifiers respectively. The following explanation can be made. Both ensembles eliminated the worst elementary classifiers (i.e. IBk, LWL, RandTree) what allowed to avoid spoiling decision by voices of irrelevant voters. In contrary, ensemble consisting all classifiers, although more diversified, did not have a chance to eliminate this irrelevant members.

Even more interesting observation can be made based on entropy analysis. The winning ensembles did not feature the highest entropy. Apparently, weak ensemble members increased the entropy value. As the result, we have to conclude that there is no straight relations between the entropy and accuracy. Nonetheless, in our opinion, this should not discourage from further researches on application of diversity measures. Other well-known diversity measures shall be also tested.

Table 4. Accuracy and entropy of best ensemble vs committee size

Committee size	Best ensemble	Accuracy	Entropy
1	LADTree	0,721	0,000
2	LADTree|NB|	0,751	0,236
3	LADTree|NB|NBTree	0,757	0,210
4	LADTree|NB|NBTree|RBFNetwork	**0,759**	0,275
5	LADTree|LWL|NB|NBTree|RBFNetwork	**0,759**	0,314
6	LADTree|LWL|MLP|NB|NBTree| RBFNetwork	0,754	**0,336**
7	BFTree|IBk(3)|LADTree|NB|NBTree| RBFNetwork|SVM	0,752	0,310
8	BFTree|LWL|MLP|NB|NBTree| RBFNetwork|RandomTree|SVM	0,750	0,335
9	BFTree|IBk(3)|LADTree|MLP|NB|NBTree|R BFNetwork| RandomTree|SVM	0,750	0,333
10	BFTree|IBk(3)|J48|LADTree|MLP|NB| NBTree|RBFNetwork|RandomTree|SVM	0,747	0,326
11	BFTree|IBk(3)|J48|LADTree|LWL|MLP| NB|NBTree|RBFNetwork|RandomTree|SVM	0,741	0,332

5 Conclusions

In the paper we presented results of our research on designing classification system for computer aided diagnosis of migraine types. Authors proposed application of ensemble system which allowed to elevate classification accuracy to acceptable level comparing to classical classification algorithms. We showed that diversity of classifiers in the ensemble are key factor which determines improvement of classification accuracy. Nonetheless, it is hard to find simple methods of measuring the diversity as there are no strict relation between the diversity measure and ensemble accuracy. Based on experimental evaluation we found that collecting heterogeneous ensemble allows to significantly elevate the accuracy for more than ten percent points. We also showed that there are not straight relation between size of the ensemble and its diversity. The best results can be obtained by subset of selected classifiers.

Obtained results opens ways for further investigation on optimisation and automation ensemble designing process by application of heuristic algorithms for committee selection which utilize diversity measures.

Acknowledgement. This paper has been elaborated in the framework of the project Opportunity for young researchers, reg. no. cz.1.07/2.3.00/30.0016, supported by Operational Programme Education for Competitiveness and co-financed by the European Social Fund and the state budget of the Czech Republic.

The work was supported in part by the statutory funds of the Department of Systems and Computer Networks as well by the Polish National Science Center under the grant no. DEC-2013/09/B/ST6/02264

References

1. Liebowitz, J. (ed.): The Handbook of Applied Expert Systems. CRC Press (1998)
2. Simi, S., Simi, D., Cvijanovi, M.: Clinical and socio-demographic characteristics of tension type headache in working population. HealthMED 6(4), 1341–1347 (2012)
3. Olesen, J., Goadsby, P.J., Ramadan, N.M., Tfelt-Hansen, P., Welch, K., Michael, A.: The Headaches, 3rd edn. Lippincott Williams & Wilkins (2005)
4. Brown, M.R.: The classification and treatment of headache. Medical Clinics of North America 35(5), 1485–1493 (1951); PMID 14862569
5. Ad Hoc Committee on Classification of Headache. Classification of Headache. JAMA 179 (1962), doi:10.1001/jama.1962.03050090045008
6. Vos, T.: Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 380 (2012), doi:10.1016/S0140-6736(12)61729-2; PMID 23245607
7. MacGregor, E.A., Hackshaw, A.: Prevalence of migraine on each day of the natural menstrual cycle. Neurolog 63(2), 351–353 (2004)
8. Jain, A.K., Duin, P.W., Mao, J.: Statistical Pattern Recognition: A Review. IEEE Trans. on PAMI 22(1), 4–37 (2000)

9. Ruta, D., Gabrys, B.: Classifier Selection for Majority Voting. Information Fusion 6, 63–81 (2005)
10. Wozniak, M., Jackowski, K.: Some Remarks on Chosen Methods of Classifier Fusion Based on Weighted Voting. In: Corchado, E., Wu, X., Oja, E., Herrero, Á., Baruque, B. (eds.) HAIS 2009. LNCS (LNAI), vol. 5572, pp. 541–548. Springer, Heidelberg (2009)
11. Kuncheva, L.I.: Combining pattern classifiers: Methods and algorithms. Wiley-Interscience, New Jersey (2004)
12. Kuncheva, L.I., Whitaker, C.J.: Ten measures of diversity in classifier ensembles: Limits for two classifiers. In: IEE Workshop on Intelligent Sensor Processing, Birmingham, pp. 10/1–10/6 (2001)
13. Krzanowski, W., Partrige, D.: Software Diversity: Practical Statistics for its Measurement and Exploatation. Raport University of Exeter, Department of Computer Science (1996)

Hypertension Type Classification Using Hierarchical Ensemble of One-Class Classifiers for Imbalanced Data

Bartosz Krawczyk and Michał Woźniak

Department of Systems and Computer Networks,
Wroclaw University of Technology
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
{bartosz.krawczyk,michal.wozniak}@pwr.wroc.pl

Abstract. The paper presents the research on the computer support system which is able to recognize the type of hypertension. This diagnostic problem is highly imbalanced, because only ca. 5% of patient suffering from hypertension are diagnosed as secondary hypertension. Additionally the secondary hypertension could be caused by several disorders (in our work we recognize the five most popular reasons) which require strikingly different therapies. Thus, appropriate classification methods, which take into consideration the nature of the decision task should be applied to this problem. We decided to employ the original classification methods developed by our team which have their origin in one-class classification and the ensemble learning. They quality was confirmed in our previous works. The accuracy of the chosen classifiers was evaluated on the basis of the computer experiments which were carried out on the real data set obtained from the hypertension clinic. The results of the experimental investigations confirmed usefulness of the proposed, hierarchical one-class classifier ensemble and could be applied in the real medical decision support systems.

Keywords: classifier ensemble, pattern classification, one-class classifier, imbalanced data, hypertension.

1 Introduction

Medical decision support systems have been focus of intense research for years. According to [1] ca. 11% of expert systems are related to the medical problems, while more than 21% of scientific papers describe applications of machine learning algorithms to medical decision tasks. This work follows-on to our previous works on hypertension type classification [2]. The hypertension is sometimes called *a silent killer*, because many persons do not realize themselves that they suffer from this disorder, which can lead to the serious health problems as coronary heart disease, heart or kidney failure, and stroke to enumerate only a few. Therefore an accurate diagnostic method which could help physician to propose an appropriate therapy is still very desirable tool. Basically, we could

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_34

distinguished two main types of hypertension: the essential hypertension and the secondary one which could be caused by the several disorders. In this work we will use the hypertension taxonomy presented in Fig. 1.

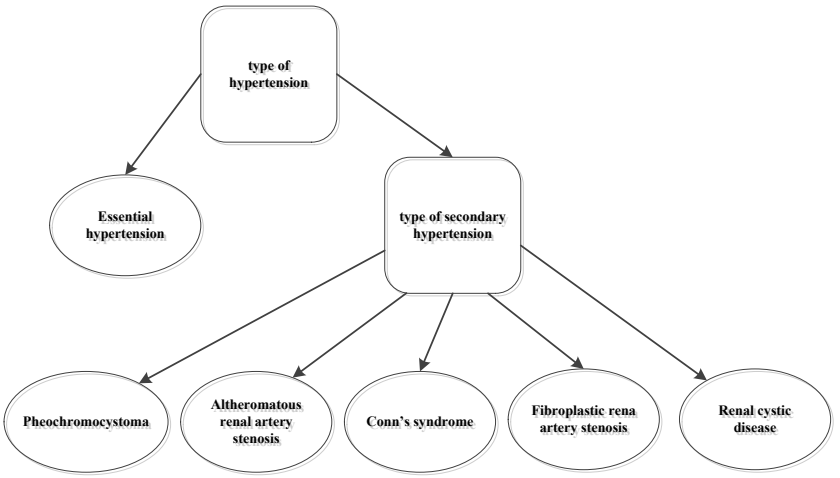

Fig. 1. The taxonomy of the hypertension

According to the medical reports ca. 95% of patient with high blood pressure suffer from the essential hypertension, while rest of them are diagnosed with the one of the five secondary types of hypertension. The problem of hypertension type diagnosis is the crucial stage for the appropriate therapy planning, but as we see it is highly imbalanced what causes that especially patients suffer from the secondary hypertension could be treated incorrectly. Therefore to solve this classification task we should use the classification methods which will take imbalanced data into consideration.

The work is organized as follows. Firstly, the hypertension problem will be described, then classification methods dedicated to imbalance pattern classification task and one-class classification approach are presented shortly. Afterwards, we present the experimental evaluation of chosen classification methods for the problem under consideration. The last section concludes the paper.

2 Hypertension

Blood pressure is determined by the amount of blood the heart pumps and the amount of resistance to blood flow in the arteries. Hypertension is a common condition in which the force of the blood against the artery walls is high enough. The hypertension is sometimes called *a silent killer*, because persons can have it for years without symptoms, but the damage to blood vessels and heart continues. Uncontrolled hypertension strongly increases the risk of health problems,

including kidney disorders, coronary heart problems, heart attack, and stroke to enumerate only a few. Normal blood pressure at rest is within the range of 100–140 mmHg systolic and 60–90 mmHg diastolic. Fortunately, the blood pressure measurement is easy and available for the most of the patients, therefore this disorder can be easily detected. Nowadays, the hypertension is recognized as the one of the main health problem, because e.g., [3] reports that in 2013 30-45% of Europeans suffer from it. It has also a huge impact on the global economy, e.g., the American Heart Association estimated the direct and indirect costs of high blood pressure in 2010 as $76.6 billion [4].

The hypertension's therapy planning is usually long and continuous process. The crucial role plays the recognition of its type. Basically, there can distinguish:

- Primary hypertension, known as essential one, that has no known cause and it is diagnosed in the majority of people, i.e., in ca. 95% of hypertension cases.
- Secondary hypertension, which is often caused by comorbid conditions, and is sometimes curable.

The physician is responsible for deciding if the hypertension is of an essential or a secondary type (so called the first level diagnosis). Because only ca. 5% of patients suffering from secondary hypertension, we face with the very hard, highly imbalanced classification problem. Additionally, there are several types of the secondary hypertension. The senior physicians from the *Broussais Hospital of Hypertension Clinic* and *Wroclaw Medical Academy* suggested its following classification:

1. fibroplastic renal artery stenosis,
2. atheromatous renal artery stenosis,
3. Conn's syndrome,
4. renal cystic disease,
5. pheochromocystoma.

3 Imbalanced Classification

Typical classification algorithms work under an assumption that the distribution of objects among the classes in the training set is roughly equal. However, many real-life applications are characterized by the fact, that it is impossible to gather equal number of examples from all classes, as some may appear less frequently or be more costly to gather. In case where one of the classes is represented by a significantly greater number of examples than other, we deal with a problem known as the imbalanced classification. Such an uneven distribution tends to result in a bias of the decision boundary produced by classifiers towards the majority class. This significantly damages the classifier performance on the minority class. With such a situation arises a need for applying carefully designed algorithms that can cope with such a difficult data distribution.

Recent works report that the unequal number of examples among classes is not the major source of problem [5]. In case, where the problem is imbalanced, but the minority class is well-represented by a significant number of objects, even standard algorithms can return good recognition rate [6]. The underlying difficulty is connected to specific data properties, that often are embedded in imbalanced data sets, such as class overlapping [7], small sample size or small disjuncts [8].

These data properties are the major reason behind a poor performance of standard classifiers on imbalanced data sets. Therefore, in recent years there has been a significant development of dedicated methods, that are able to overcome mentioned difficulties. They can be divided into four groups [9]:

- Data-level methods that work at the pre-processing phase of the data. They usually aim at re-balancing the distribution between the classes and they are not dependent on used classifier model. The best known technique here is SMOTE [10], which adds synthetic objects to the minority class.
- Classifier-level methods try to modify the existing classifiers in order to make them robust to unequal object distributions. This is mainly done by reducing or eliminating the bias towards the majority class or shifting the emphasis of the learning step towards the minority class.
- Cost-sensitive approaches introduce a high penalty factor for misclassifying the minority class objects. Instead of standard 0-1 loss function, they use a pre-defined cost matrix, that allows to define the penalty for the learning algorithm for misclassifying minority samples. The most popular are cost-sensitive decision trees, however cost-sensitive neural networks or support vector machines have been also introduced.
- Hybrid approach that uses a classifier ensemble together with one of the mentioned above techniques [11]. They take a full advantage of committee approaches combined with effective method for handling imbalance. Most popular methods include SMOTEBoost [12], EasyEnsemble [13] and AdaCost [14].

4 One-Class Classification

Because in our research we use one-class classification (OCC) approach [15], then let's present this concept shortly. During the training step of OCC only objects from a single class, known as the target concept ω_T, are at disposal. The purpose of OCC is the estimation of a decision surface that encloses all available data samples and thus describes the concept [16]. During the one-class classifier exploitation step, objects from different distributions, unknown during the training phase may appear. They represent data that do not belong to the target concept, and are labeled as outliers ω_O. OCC can be considered as learning in the absence of counterexamples. The target class should be separated from all possible outliers, and hence the decision boundary should be estimated in all directions in the feature space around the target class. This allows us to create

a pattern recognition system that is robust to appearance of new classes or lack of representative counterexamples.

OCC is an attractive solution to many real-life problems where data coming from a single class is abundant but other objects are hard or even impossible to obtain such as spam filtering/intrusion detection [17].

Despite the original aim of OCC to work on cases with no access to counterexamples, there is a number of reports that discuss the usefulness of OCC approach in cases, where objects from all of classes are at disposal. This is explained by several attractive properties of OCC, resulting from their different learning procedure. They do not minimize the classification error, but adapt to the features of the target class. They do not use all of the available knowledge from the training set, which results in worse recognition accuracy than multi-class methods for standard problems. However, in case of difficult data sets, OCC can outperform multi-class algorithms, as single-class classifiers are robust to many difficulties embedded in the nature of data. This seems as a very attractive proposal for dealing with imbalanced data sets, and our previous works confirm that using ensembles of OCC can return highly effective recognition systems for uneven class distributions [18, 19].

In this work, we propose to embed the background medical knowledge about the hypertension problem (see Fig. 1) into the process of designing the medical decision support system. The introduced system is realized as a two-step hierarchical architecture, with each step handling a decomposed part of the recognition task.

Step 1. On the upper level of the introduced architecture, we implement a single one-class classifier to distinguish between *essential* and *secondary* hypertension. In our data set, the *essential* hypertension class has significantly larger number of examples than all of the remaining classes, thus leading to an imbalanced problem. We counter this by using a OCC model that is trained on *essential* hypertension class as the target class. All of the remaining five secondary hypertension classes are fused together to create an outlier class. The system outputs a binary value - it can decide that a new object belongs to the target class (*essential* hypertension) or to the outlier class (*secondary* hypertension). In case of the latter decision, the recognition system moves to the second level of architecture, which is able to distinguish between secondary classes. By this, we are able to efficiently handle imbalanced problem at the first step of our architecture.

Step 2. This step aims at distinguishing between one of five types of *secondary* hypertension. This is implemented by decomposing the original multi-class problem with an ensemble of OCC algorithms. Each of the classes is handled by a dedicated one-class classifier, that adjusts to its properties. Then, we use an Error-Correcting Output Codes [20] to reconstruct an original multi-class problem from single-class responses. This allows us to handle difficult multi-class data with efficient decomposition strategy.

5 Experimental Investigations

The aims of the experiment were:

- propose an efficient medical decision support system for automatic diagnosis of hypertension types;
- examine the usefulness of the proposed hierarchical one-class classifier ensemble and compare it to several state-of-the-art methods.

5.1 Set-Up

The initial works on the hypertension type classification system was developed together with *Service d'Informatique Médicale* from the *University Paris VI* [2]. All data was getting from the medical database *ARTEMIS*, which contains the data of the patients with hypertension, whose have been treated in *Hôpital Broussais* in Paris. Although the set of symptoms necessary to correctly assess the existing hypertension is pretty wide, in practice for the diagnosis, results of 18 examinations (which came from general information about patient, blood pressure measurements and basis biochemical data) are used, whose are presented in Tab. 1.

Table 1. Description of features

#	name	#	name
1	sex	10	efffusion
2	body weight	11	artery stenosis
3	high	12	heart failure
4	cigarette smoker	13	palpitation
5	limb ache	14	carotid or lumbar murmur
6	alcohol	15	serum creatinine
7	systolic blood pressure	16	serum potassium
8	diastolic blood pressure	17	serum sodium
9	maximal systolic blood pressure	18	uric acid

The set-up of used classifiers was as follows:

- As a base one-class classifier for the proposed classifier ensemble, we decided to use Support Vector Data Description (SVDD) with RBF kernel and kernel parameters $\sigma = 0.3$, $C = 8$.
- As reference methods, we use:
  - C4.5 decision tree with post-pruning,
  - a multi-class Support Vector Machine (SVM) with RBF kernel and kernel parameters $\sigma = 0.1$, $C = 10$,
  - Random Forest (RandF) ensemble with 120 decision trees .
- The parameter values were established with a grid-search procedure.

– For comparison purposes we use the mentioned classifiers combined with SMOTE preprocessing, in order to counter the imbalance ratio between *essential* class and remaining ones. For SMOTE algorithm, we use 5 neighbors.

All experiments were done with the usage of combined 5x2 cv F test [21] with $\alpha = 0.05$, that allowed to assess the statistical significance of the obtained results.

5.2 Results

The results of the experiments, with the respect to geometric mean (G-mean) values and statistical analysis are given in Tab. 2.

Table 2. Results of experiments on hypertension type classification

Classifier	G-mean	Statistically better than
C4.5	50.92	-
SVM	55.12	C4.5
RandF	57.98	C4.5,SVM
C4.5+SMOTE	67.82	C4.5,SVM,RandF
SVM+SMOTE	68.28	C4.5,SVM,RandF
RandF+SMOTE	70.07	C4.5,SVM,RandF,C4.5+SMOTE,SVM+SMOTE
Hierarchical OCC	75.86	ALL OTHER METHODS

5.3 Discussion of the Results

On the basis of the presented results we may formulate a few interesting observations. The experiments showed, that our hypertension dataset poses a challenge for standard machine learning algorithms, and that it is not a trivial task to achieve a good performance for this problem. However, the output of the experiment proved the quality of the proposed hierarchical ensemble of one-class classifiers. Let us take a closer look into the performance of each methods.

Canonical classifiers deliver highly unsatisfactory results. Neither C4.5, SVM or Random Forest were able to efficiently discriminate between hypertension types. This comes from the fact, that we deal with a multi-class and highly imbalanced problem. If we had used standard accuracy as measure, these classifiers would perform satisfactory. However, G-mean metric allows us to examine their performance with the respect to uneven distribution between classes. And from it we can see, that they fail to properly recognize minority classes (*secondary* hypertension types).

When embedding a dedicated pre-processing method (namely SMOTE algorithm) into these classifiers, we can see a significant rise of the G-mean value. This is because SMOTE inputs artificial instances into minority classes and is able to reduce the classifier's bias towards the majority class. However, one should have in mind a strong limitation of such techniques. With the usage of SMOTE comes the main problem how many of the artificial samples we should generate. Intuition points that best results should be achieved when classes have equal number of objects. Yet with so big disproportion (approximately 9:1) after

some repetitions of this algorithm the new objects will be created only on the basis of previously artificially created ones. Therefore it is hard to conclude if so many artificial objects will be representative for the problem.

Our proposed hierarchical one-class ensemble does not suffer from the mentioned limitations. Additionally, it further significantly boosts the recognition rate, which is reflected by the highest G-mean score and backed-up with statistical testing. This shows that OCC can be an efficient tool for handling highly imbalanced data set, despite the fact that it does not use any knowledge about counterexamples. Extending this with a second-level architecture for decomposing multi-class problem with single-class methods allowed us to achieve a very good discrimination between five *secondary* hypertension classes. Combining this approaches into two-level architecture resulted in a robust and effective medical decision support system for hypertension type classification.

6 Conclusions

The paper presents the experimental evaluation of the set of compound classifiers for highly imbalanced multi-class classification task. The considered task was related to crucial problem of hypertension type diagnosis which is recognized as one of the main serious social disease. The proposed method based on one-class classifier ensemble significantly outperforms other considered methods as hybrid approaches used preprocessing as SMOTE. As we mentioned in the previous section, the problem of imbalance data classification is visible not only in disparity among number of training examples represented considered classes, but what maybe more important in the specific data properties. Therefore our future works on decision support systems for the hypertension type classification will look for an appropriate hybrid classifier which is able to take such properties into consideration.

Acknowledgment. This work was supported by The Polish National Science Centre under the grant PRELUDIUM number DEC-2013/09/N/ST6/03504 and by the statutory funds of the Department of Systems and Computer Networks, Wroclaw University of Technology.

References

1. Liebowitz, J.: The Handbook of Applied Expert Systems. Taylor & Francis (1997), `http://books.google.pl/books?id=6DNgIzFNSZsC`
2. Wozniak, M.: Two-stage classifier for diagnosis of hypertension type. In: Maglaveras, N., Chouvarda, I., Koutkias, V., Brause, R. (eds.) ISBMDA 2006. LNCS (LNBI), vol. 4345, pp. 433–440. Springer, Heidelberg (2006)
3. Mancia, G., et al.: esh/esc guidelines for the management of arterial hypertension: the task force for the management of arterial hypertension of the european society of hypertension (esh) and of the european society of cardiology (esc). European Heart Journal 34(28), 2159–2219 (2013)
4. Lloyd-Jones, D., Adams, R.J., Brown, T.M., Carnethon, M., Dai, S., De Simone, G., Ferguson, T.B., Ford, E., Furie, K., Gillespie, C., et al.: Heart disease and stroke statistics 2010 update a report from the american heart association. Circulation 121(7), e46–e215 (2010)

5. Sun, Y., Wong, A.K.C., Kamel, M.S.: Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence 23(4), 687–719 (2009)
6. Chen, X., Wasikowski, M.: Fast: A roc-based feature selection metric for small samples and imbalanced data classification problems. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 124–132 (2008)
7. Garcia, V., Mollineda, R.A., Sánchez, J.S.: On the k-nn performance in a challenging scenario of imbalance and overlapping. Pattern Analysis and Applications 11(3-4), 269–280 (2008)
8. Napierala, K., Stefanowski, J.: Identification of different types of minority class examples in imbalanced data. In: Corchado, E., Snášel, V., Abraham, A., Woźniak, M., Graña, M., Cho, S.-B. (eds.) HAIS 2012, Part II. LNCS (LNAI), vol. 7209, pp. 139–150. Springer, Heidelberg (2012)
9. Lopez, V., Fernandez, A., Moreno-Torres, J.G., Herrera, F.: Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. Expert Systems with Applications 39(7), 6585–6608 (2012)
10. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
11. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews 42(4), 463–484 (2012)
12. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: Smoteboost: Improving prediction of the minority class in boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 107–119. Springer, Heidelberg (2003)
13. Liu, X., Wu, J., Zhou, Z.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39(2), 539–550 (2009)
14. Sun, Y., Kamel, M.S., Wong, A.K.C., Wang, Y.: Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition 40(12), 3358–3378 (2007)
15. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Machine Learning 54(1), 45–66 (2004)
16. Tax, D., Duin, R.P.W.: Characterizing one-class datasets. In: Proceedings of the Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa, pp. 21–26 (2005)
17. Giacinto, G., Perdisci, R., Del Rio, M., Roli, F.: Intrusion detection in computer networks by a modular ensemble of one-class classifiers. Inf. Fusion 9, 69–82 (2008)
18. Krawczyk, B., Woźniak, M.: Diversity measures for one-class classifier ensembles. Neurocomputing 126, 36–44 (2014)
19. Krawczyk, B., Woźniak, M., Cyganek, B.: Clustering-based ensembles for one-class classification. Information Sciences 264, 182–195 (2014)
20. Wilk, T., Wozniak, M.: Soft computing methods applied to combination of one-class classifiers. Neurocomput. 75, 185–193 (2012)
21. Alpaydin, E.: Combined 5 x 2 cv f test for comparing supervised classification learning algorithms. Neural Computation 11(8), 1885–1892 (1999)

Handling Label Noise in Microarray Classification with One-Class Classifier Ensemble

Bartosz Krawczyk and Michał Woźniak

Department of Systems and Computer Networks,
Wroclaw University of Technology,
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
{bartosz.krawczyk,michal.wozniak}@pwr.wroc.pl

Abstract. The advance of high-throughput techniques, such as gene microarrays and protein chips have a major impact on contemporary biology and medicine. Due to the high-dimensionality and complexity of the data, it is impossible to analyze it manually. Therefore machine learning techniques play an important role in dealing with such data. In this paper, we investigate the influence of label noise on the effectiveness of classification system applied to microarray analysis. Popular methods do not have any mechanism for handling such difficulties embedded in the nature of data. To cope with that, we propose to use a one-class classifiers, which distinct from canonical methods, rely on objects coming from single class distributions only. They distinguish observations coming from the given class from any other possible decision about the examples, that were unseen during the classification step. While having less information to dichotomize between classes, one-class models can easily learn the specific properties of a given data set and are robust to difficulties embedded in the nature of the data. We show, that using ensembles of one-class classifiers can give as good results as canonical multi-class classifiers, while allowing to deal with unexpected label noise in the data. Experimental investigations, carried out on public data sets, prove the usefulness of the proposed approach.

Keywords: classifier ensemble, pattern classification, one-class classifier, bioinformatics, microarray, label noise.

1 Introduction

Contemporary high-throughput technologies produce massive volumes of biomedical data. Transcriptional research and profiling, with the usage of microarray technologies are powerful tools to gain a deep insight into the pathogenesis of complex diseases that are a plague to modern society, such as various forms of cancer. Contemporary works on cancer profiling proved, that gene expression patters can be used for highly accurate cancer subtype identification [1] - leukemia [2], melanoma [3], breast cancer [4] or prostate cancer [5] to name a few.

A. Madevska Bogdanova and D. Gjorgjevikj, *ICT Innovations 2014*,
Advances in Intelligent Systems and Computing 311, DOI: 10.1007/978-3-319-09879-1_35

Recognizing cancer characteristics, based on their individual expression profiles is a promising direction for finding necessary information for future patient-profiled therapy. Currently, there are no universal rules on how individuals respond to chemotherapy. Additionally, existing chemotherapies have in most cases severe side-effects and varied treatment qualities.

Microarray experiments generate massive amounts of data, characterized by a high complexity and dimensionality. This causes a need for an efficient decision support system to extract the meaningful information. Methods which have their origin in machine learning are widely used in this area [6], with two distinct approaches - unsupervised [7] and supervised learning [8]. In this paper we will focus on the latter one, because the supervised machine learning is a promising approach for analyzing microarray results in context of predicting patients outcome. Among the methods applied to this task one has to mention at least Support Vector Machines [9], Multiple Classifier Systems [10], which have gained an significant attention of the bioinformatics community in recent years. Additionally, Random Forest [11] and Rotation Forest [12] ensembles have displayed an excellent classification accuracy for small-sample, high dimensionality microarray data sets, outperforming single-model approaches.

Another important issue is handling the problem known as the *curse of dimensionality*, because microarray data usually is characterized by a relatively small number of objects, in comparison to dimensionality of their feature space often reaching several thousands. This is the source of difficulties for machine learning algorithms, resulting in a reduced accuracy and increased computational complexity. In such highly dimensional space, a significant number of features possesses small discriminative power and does not contribute anything valuable to the classification process. This makes feature selection a crucial step in microarray classification [13].

Although there are many applications of machine learning-based decision support systems in bioinformatics, there are still many unresolved problems, such as:

- How to integrate heterogeneous data sources to achieve better insight into the mechanism behind complex diseases?
- How to organize, store, analyze and visualize high-dimensionality data obtained from the biomedical data flood?
- How to deal with the problem of high-dimensionality, small sample size, which strongly affects the classification performance and may lead to overfitting, poor generalization and unstable predictors?
- How to deal with difficulties embedded in the nature of microarray data, such as noise or class imbalance, as canonical machine learning classifiers cannot cope with them easily?

In this paper we concentrate on the last of the mentioned issues.

We analyze the problem of label noise in the microarray classification, i.e., Label noise is a situation, in which object belonging to one class has assigned an incorrect label in the training set. This can be a result of domain expert or

labeling system error, and reduces the quality of the training data. This in turn leads to inputting an incorrect information to classifiers and significantly damage the generalization ability of estimated decision support. Canonical machine learning algorithms used in bioinformatics cannot tackle label noise and suffer from a reduction of the predictive accuracy. We propose to analyze microarray data with the usage of one-class classifiers, instead of commonly applied binary ones [14]. We use their weighted versions, that assign a weight to each object on the basis of its degree of relevance. Thus, noisy samples that are located far from the target class distribution receive low weight and do not damage the classifier performance.

To cope with the high dimensionality problem we apply an ensemble approach, based on Random Subspaces [15]. By decomposing the feature space we at the same time reduce the overall computational complexity of the classification model and assure initial diversity among the pool of individual classifiers in the committee. A diversity-based pruning method is applied to discard redundant classifiers and to choose mutually complementary one-class predictors.

Experiments carried on a set of public microarray data sets, show that the proposed approach maintains a good classification accuracy, while displaying an improved robustness to label noise.

2 One-Class Classification

The aim of one-class classification (OCC) is to detect one specific class from all of the remaining ones (e.g., selecting cats from all animals). The given class is denoted as target class ω_t, while the remaining objects are considered as outliers ω_O. During the learning only examples target class (known also as positive examples) are being presented to learner, while it is assumed that during the exploitation phase new, unseen objects from other classes may appear. This can be seen as learning in the absence of counterexamples.

OCC problems are common in the real world where positive examples are widely available but negative ones are hard, expensive or even impossible to gather. Such approach is very useful as well for many practical cases especially when the target class is "stable" and outlier one is "unstable"as in cases of spam filtering or intrusion detection (IDS/IPS) [16].

Significant popularity was achieved by methods that concentrate on estimation of a closed boundary for given data [17]. Boundary methods can effectively work with a small number of objects, which makes them a perfect tool for applications suffering from a small sample size, such as microarrays classification. The most popular methods from this group are one-class support vector machine (OCSVM) [18] and support vector data description (SVDD) [19]. In this work we will use an extension of the OCSVM, known as Weighted One-Class Support Vector Machine [20].

3 Proposed Approach

In this paper, we propose to employ a weighted one-class classification approach to microarray classification tasks, that suffer from the label noise problem. Let us list the main features and advantages of the proposed approach:

1. We decompose the microarray classification task (which is usually a binary one) into separate single-class problems. This allows us to use every available information (as there will be a dedicated ensemble for each class), while capturing the unique properties of each class (as OCC methods adjust themselves to find the optimal description of a given context).
2. The high dimensionality of the feature space is difficult to handle for one-class boundary classifiers. It significantly increases their complexity, the training and execution times and lead to a much more difficult task of estimating the volume of the boundary. To deal with this difficulty we use a Random Subspace ensemble to decompose the feature space into smaller competence areas and build an ensemble of simpler one-class models.
3. As Random Subspace may lead to creation of similar or weak classifiers, one should apply a pruning procedure in order to remove the irrelevant models from the pool of individual classifiers. In our approach we use a diversity-based method tuned for the specific nature of the OCC task.
4. By using weighted one-class models on significantly reduced competence spaces, we may detect objects that are far from the standard distribution of the target class. Such objects most probably are affected by label noise. We assign them a lower weight value, which significantly reduces their influence on the shape of the decision boundary and allows to alleviate the label noise problem.

3.1 Dealing with the High Dimensionality Problem

The boundary methods for OCC base their decision on computing a distance between the object x and the estimated boundary, which encloses the target class ω_T (or support vectors, that describe this boundary). For such models one may directly apply fusion methods, that are based on the discrete output (returned class label) of the individual classifiers e.g., - voting combiners. However, to apply more efficient combination methods, which assume the continuous outputs of each of the individuals, the support of an object x for a given class is required.

We propose to use the following heuristic support function produced on the basis of a distance:

$$F(x, \omega_T) = \frac{1}{c_1} exp(-d(x|\omega_T)/c_2), \tag{1}$$

which models a Gaussian distribution around the classifier, where $d(x|\omega_T)$ is a distance (Euclidean distance is used) from the evaluated object to the support vectors describing the target concept, c_1 is the normalization constant and c_2 is

the scale parameter. Parameters c_1 and c_2 should be fitted to the target class distribution.

To handle high dimensional data we use the Random Subspace method to partition the data set into many subspaces of smaller dimensionality. Each base classifier is trained on a new subset, which is highly smaller than the original feature space size. This boosts the training time, while applying ensemble principles makes sure that despite using weaker predictors, we still get the satisfying accuracy [21].

3.2 Pruning the Ensemble

As Random Subspace may produce classifiers of different level of individual quality and diversity, a classifier section step is most beneficial to forming an one-class ensemble. We propose to use a diversity measure, designed specifically for OCC. Our previous works have shown, that such methods allow to discard irrelevant predictors from the pool.

Let's assume that the highest ensemble diversity for a given object $x_j \in X$ is displayed by $[R/2]$ of the ensemble votes with the same value (ω_T or ω_O) and remaining R - $[R/2]$ with the other value. If all votes returned identical response the ensemble cannot be considered as a diverse one. Let us denote by $r(x_j)$ the number of one-class classifiers that correctly recognize the object x_j. Assuming there are N objects in the training set, one may use entropy to measure the diversity using the presented concept:

$$E_{oc}(\Pi^r) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{(R - [R/2])} \min\{r(x_j), R - r(x_j)\}. \tag{2}$$

where Π^r is the considered pool of classifiers.

This is a non-pairwise (global) diversity measure, which take values from [0,1]. 0 corresponds to identical ensemble and 1 corresponds to the highest possible diversity.

3.3 Combination Rule

As a fusion method we use a one-class mean vote, which combines binary output labels of one-class classifiers. It can be written as:

$$y_{mv}(x) = \frac{1}{L} \sum_{k} [(F_k(x, \omega_T) \geq \theta_k)], \tag{3}$$

where $[(\cdot)]$ is the *Iverson brackets* and θ_k is threshold for the target class. When a threshold equal to 0.5 is applied this rule transforms into a majority vote for binary problems.

4 Experimental Investigations

The aims of the experiment were:

- to examine the robustness of the commonly used in microarray analysis classifiers to different ratio of label noise;
- to establish the usefulness of the proposed one-class ensemble in handling noisy microarrays.

4.1 Set-Up

In this section we evaluate the proposed one-class ensemble on the basis of data sets available at [1], whose details are given in Table 1. Four different data sets were used.

Table 1. Statistics of the data sets used in the experiments

data set	samples (class 1 / class 2)	features
Breast Cancer	78 (34 / 44)	24481
Central Nervous System	60 (21 / 39)	7129
Colon Tumor	62 (22 / 40)	6500
Lung Cancer	181 (31 / 150)	12533

We examine the performance of classifiers in three different scenarios: with 0%, 15% and 30% of objects subject to label noise. We simulate the label noise by switching the class label into the opposite one.

To put the obtained results into context we have tested the performance of multi-class classifiers used for this task:

- SVM - single SVM (trained with RBF kernel and SMO procedure),
- RandF - Random Forest (consisting of 100 decision trees),
- RotF - Rotation Forest (consisting of 100 decision trees),
- WOCSVM - Weighted One-Class SVM.
- Proposed - ensemble of weighted Support Vector Data Description (SVDD) with RBF kernel and kernel parameters $\sigma = 0.3$, $C = 8$. Each Random Subspace ensemble consisted of 5% of original feature space. We have created 100 classifiers for the ensembles dedicated to each of data sets.

Results are based on leave-one-out cross-validation (LOOCV). The Friedman ranking test [22] was done for comparison over multiple benchmark data sets.

4.2 Results

Results with the respect to G-mean and statistical rankings are given in Tab. 2 - 4.

[1] `http://datam.i2r.a-star.edu.sg/datasets/krbd/`

Table 2. G-mean and statistical rankings of examined methods for case with 0% of label noise

dataset	SVM	RandF	RotF	WOCSV	Proposed
Breast Cancer	90.73	91.89	92.47	87.21	91.92
Central Nervous System	91.25	94.11	92.79	90.22	93.55
Colon Tumor	82.39	86.42	86.15	77.92	85.22
Lung Cancer	79.18	82.28	83.96	77.12	82.31
Avg. rank	4.00	1.84	1.84	5.00	2.32

Table 3. G-mean and statistical rankings of examined methods for case with 15% of label noise

dataset	SVM	RandF	RotF	WOCSV	Proposed
Breast Cancer	86.35	88.29	88.82	84.94	89.46
Central Nervous System	85.03	87.32	88.06	85.82	88.06
Colon Tumor	80.15	83.89	84.57	79.82	84.57
Lung Cancer	77.54	80.82	81.59	76.03	82.48
Avg. rank	4.50	3.00	1.50	4.50	1.50

Table 4. G-mean and statistical rankings of examined methods for case with 30% of label noise

dataset	SVM	RandF	RotF	WOCSV	Proposed
Breast Cancer	70.59	73.75	76.29	76.12	81.34
Central Nervous System	75.26	78.12	78.82	77.57	83.44
Colon Tumor	69.56	70.49	74.02	71.95	76.18
Lung Cancer	69.12	74.53	76.04	74.38	77.73
Avg. rank	5.00	3.50	2.50	3.00	1.00

4.3 Discussion of the Results

The results allows us to draw several interesting conclusions about the performance of examined classifiers under the influence of label noise.

In case of the lack of label noise (0% of objects) we deal with the standard problem of microarray classification. Here SVM, RandF and RotF deliver very good performance, as reported in numerous research papers from last decade. Using single-model WOCSVM approach (one classifier per each class), we notice that it deliver the worst performance from all of the mentioned methods. This is due to the fact, that standard boundary methods cannot handle well such high-dimensional data and work with limited knowledge (without access to counterexamples). However, the proposed weighted one-class ensemble is able to achieve similar performance as one outputted by multi-class methods. This shows, that with the use of feature space partitioning and pruning, the proposed method can efficiently handle standard data sets, no worse than canonical methods.

In case of a small label noise (15% of objects) we can see a drop in G-mean value for all of the examined methods. SVM and Random Forest suffer the most from the examined classifiers. Rotation Forest is able to cope with the noise. The proposed ensemble delivers identical performance to Rotation Forest, and

in two cases outperform all other methods. This shows that our method can adapt itself to noisy scenarios and efficiently filter small degree of label noise with the weighting procedure.

In case of a large label noise (30% of objects) we can see that all of the canonical multi-class methods deliver impaired performance, as they are not able to build efficient decision surface on the basis of noisy training set. What is interesting, in such case even the single-model WOCSVM is able to perform very well and deliver better performance than SVM and Random Forest. This shows that weighed OCC can reject uncertain objects from taking a significant part in the process of shaping the decision boundary. The proposed method is able to outperform all other methods for all data sets. This shows the usefulness of using one-class ensemble decomposition for handling microarrays with a large noise ratio.

In summary, the proposed method delivers similar performance to canonical multi-class methods on standard microarray datasets, while offering a robust behavior in case of label noise presence.

5 Conclusions

In this paper we presented the preliminary study on the classification with unreliable teacher problem. We mainly focused on the one-class classification paradigm and prove that the ensemble of OCCs is rendered useful, especially when the label noise is pretty high. In the future we would like to carry out the computer experiments on the wider range of data sets, especially coming from microarray analysis problems. We believe that such approach, based on OCC ensemble have a huge potential and still awaits for proper attention from the machine learning community.

Acknowledgment. This work was supported by the Polish National Science Center under the grant no. DEC-2013/09/B/ST6/02264 and by the statutory funds of the Department of Systems and Computer Networks, Wroclaw University of Technology.

References

1. Tinker, A.V., Boussioutas, A., Bowtell, D.D.L.: The challenges of gene expression microarrays for the study of human cancer. Cancer Cell 9(5), 333–339 (2006)
2. Silveira, V.S., Scrideli, C.A., Moreno, D.A., Yunes, J.A., Queiroz, R.G.P., Toledo, S.C., Lee, M.L.M., Petrilli, A.S., Brandalise, S.R., Tone, L.G.: Gene expression pattern contributing to prognostic factors in childhood acute lymphoblastic leukemia. Leukemia and Lymphoma 54(2), 310–314 (2013)
3. Schatton, T., Murphy, G.F., Frank, N.Y., Yamaura, K., Waaga-Gasser, A.M., Gasser, M., Zhan, Q., Jordan, S., Duncan, L.M., Weishaupt, C., Fuhlbrigge, R.C., Kupper, T.S., Sayegh, M.H., Frank, M.H.: Identification of cells initiating human melanomas. Nature 451(7176), 345–349 (2008)
4. Finak, G., Bertos, N., Pepin, F., Sadekova, S., Souleimanova, M., Zhao, H., Chen, H., Omeroglu, G., Meterissian, S., Omeroglu, A., Hallett, M., Park, M.: Stromal gene expression predicts clinical outcome in breast cancer. Nature Medicine 14(5), 518–527 (2008)

5. Lynch, C.C., Hikosaka, A., Acuff, H.B., Martin, M.D., Kawai, N., Singh, R.K., Vargo-Gogola, T.C., Begtrup, J.L., Peterson, T.E., Fingleton, B., Shirai, T., Matrisian, L.M., Futakuchi, M.: Mmp-7 promotes prostate cancer-induced osteolysis via the solubilization of rankl. Cancer Cell 7(5), 485–496 (2005)
6. Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armananzas, R., Santaf, G., Perez, A., Robles, V.: Machine learning in bioinformatics. Briefings in Bioinformatics 7(1), 86–112 (2006)
7. Wang, Y., Yu, Z., Anh, V.: Fuzzy c-means method with empirical mode decomposition for clustering microarray data. International Journal of Data Mining and Bioinformatics 7(2), 103–117 (2013)
8. Ringner, M., Peterson, C., Khan, J.: Analyzing array data using supervised methods. Pharmacogenomics 3(3), 403–415 (2002), www.scopus.com; cited By 43 (since 1996)
9. Bariamis, D., Maroulis, D., Iakovidis, D.K.: Unsupervised svm-based gridding for dna microarray images. Computerized Medical Imaging and Graphics 34(6), 418–425 (2010)
10. Woźniak, M., Grana, M., Corchado, E.: A survey of multiple classifier systems as hybrid systems. Information Fusion 16, 3–17 (2014)
11. Moorthy, K., Mohamad, M.S.: Random forest for gene selection and microarray data classification. In: Lukose, D., Ahmad, A.R., Suliman, A. (eds.) KTW 2011. CCIS, vol. 295, pp. 174–183. Springer, Heidelberg (2012)
12. Liu, K., Huang, D.: Cancer classification using rotation forest. Computers in Biology and Medicine 38(5), 601–610 (2008)
13. Inza, I., Larraaga, P., Blanco, R., Cerrolaza, A.J.: Filter versus wrapper gene selection approaches in dna microarray domains. Artificial Intelligence in Medicine 31(2), 91–103 (2004)
14. Krawczyk, B.: Combining one-class support vector machines for microarray classification. In: FedCSIS, pp. 83–89 (2013)
15. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844 (1998)
16. Noto, K., Brodley, C., Slonim, D.: Frac: A feature-modeling approach for semi-supervised and unsupervised anomaly detection. Data Mining and Knowledge Discovery 25(1), 109–133 (2012)
17. Tax, D.M.J., Juszczak, P., Pękalska, E.z., Duin, R.P.W.: Outlier detection using ball descriptions with adjustable metric. In: Yeung, D.-Y., Kwok, J.T., Fred, A., Roli, F., de Ridder, D. (eds.) SSPR&SPR 2006. LNCS, vol. 4109, pp. 587–595. Springer, Heidelberg (2006)
18. Schölkopf, B., Smola, A.: Learning with kernels: support vector machines, regularization, optimization, and beyond. In: Adaptive Computation and Machine Learning. MIT Press (2002)
19. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Machine Learning 54(1), 45–66 (2004)
20. Bicego, M., Figueiredo, M.A.T.: Soft clustering using weighted one-class support vector machines. Pattern Recognition 42(1), 27–32 (2009)
21. Wilk, T., Woźniak, M.: Soft computing methods applied to combination of one-class classifiers. Neurocomput. 75, 185–193 (2012)
22. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (2006)

Author Index