M. Elena Renda · Miroslav Bursa
Andreas Holzinger · Sami Khuri (Eds.)

LNCS 9267

Information Technology in Bio- and Medical Informatics

6th International Conference, ITBAM 2015
Valencia, Spain, September 3–4, 2015
Proceedings

Lecture Notes in Computer Science 9267

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

More information about this series at http://www.springer.com/series/7409

M. Elena Renda · Miroslav Bursa
Andreas Holzinger · Sami Khuri (Eds.)

Information Technology in Bio- and Medical Informatics

6th International Conference, ITBAM 2015
Valencia, Spain, September 3–4, 2015
Proceedings

Springer

Editors
M. Elena Renda
Institute of Informatics and Telematics
Pisa
Italy

Andreas Holzinger
Medical University Graz
Graz
Austria

Miroslav Bursa
Czech Technical University in Prague
Prague
Czech Republic

Sami Khuri
San Jose State University
San Jose, CA
USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-22740-5 ISBN 978-3-319-22741-2 (eBook)
DOI 10.1007/978-3-319-22741-2

Library of Congress Control Number: 2015945607

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Biomedical engineering and medical informatics represent challenging and rapidly growing areas. Applications of information technology in these areas are of paramount importance. Building on the success of ITBAM 2010, ITBAM 2011, ITBAM 2012, ITBAM 2013, and ITBAM 2014, the aim of the sixth ITBAM conference was to continue bringing together scientists, researchers, and practitioners from different disciplines, namely, from mathematics, computer science, bioinformatics, biomedical engineering, medicine, biology, and different fields of life sciences, so they can present and discuss their research results in bioinformatics and medical informatics. We hope that ITBAM will continue serving as a platform for fruitful discussions between all attendees, where participants can exchange their recent results, identify future directions and challenges, initiate possible collaborative research, and develop common languages for solving problems in the realm of biomedical engineering, bioinformatics, and medical informatics. The importance of computer-aided diagnosis and therapy continues to draw attention worldwide and has laid the foundations for modern medicine with excellent potential for promising applications in a variety of fields, such as telemedicine, Web-based healthcare, analysis of genetic information, and personalized medicine.

Following a thorough peer-review process, we selected nine long papers for oral presentation and two short papers for poster session for the sixth annual ITBAM conference. The organizing committee would like to thank the reviewers for their excellent job. The articles can be found in the proceedings and are divided into the following sections: Medical Terminology and Clinical Processes and Machine Learning in Biomedicine. The papers show how broad the spectrum of topics in applications of information technology to biomedical engineering and medical informatics is.

The editors would like to thank all the participants for their high-quality contributions and Springer for publishing the proceedings of this conference. Once again, our special thanks go to Gabriela Wagner for her hard work on various aspects of this event.

June 2015

M. Elena Renda
Miroslav Bursa
Andreas Holzinger
Sami Khuri

Organization

General Chair

Christian Böhm	University of Munich, Germany

Program Committee Co-chairs

Miroslav Bursa	Czech Technical University Prague, Czech Republic
Andreas Holzinger	Medical University Graz, Austria
Sami Khuri	San José State University, USA
M. Elena Renda	IIT - CNR, Pisa, Italy, Honorary Chair

Program Committee

Werner Aigner	FAW
Tatsuya Akutsu	Kyoto University, Japan
Andreas Albrecht	Queen's University Belfast, UK
Peter Baumann	Jacobs University Bremen, Germany
Veselka Boeva	Technical University of Plovdiv, Bulgaria
Roberta Bosotti	Nerviano Medical Science s.r.l., Italy
Miroslav Bursa	Czech Technical University, Czech Republic
Christian Böhm	University of Munich, Germany
Rita Casadio	University of Bologna, Italy
Sònia Casillas	Universitat Autònoma de Barcelona, Spain
Kun-Mao Chao	National Taiwan University
Vaclav Chudacek	Czech Technical University in Prague, Czech Republic
Hans-Dieter Ehrich	Technical University of Braunschweig, Germany
Christoph M. Friedrich	University of Applied Sciences Dortmund, Germany
Alejandro Giorgetti	University of Verona, Italy
Jan Havlik	Czech Technical University in Prague, Czech Republic
Volker Heun	Ludwig-Maximilians-Universität München, Germany
Larisa Ismailova	NRNU MEPhI, Moscow, Russia
Alastair Kerr	University of Edinburgh, UK
Sami Khuri	San Jose State University, USA
Michal Krátký	Technical University of Ostrava, Czech Republic
Vaclav Kremen	Czech Technical University in Prague, Czech Republic
Jakub Kuzilek	Czech Technical University, Czech Republic
Gorka Lasso	CICbioGUNE
Lenka Lhotska	Czech Technical University, Czech Republic
Roger Marshall	Plymouth State University, USA
Elio Masciari	ICAR-CNR, Università della Calabria, Italy

Erika Melissari	University of Pisa, Italy
Henning Mersch	RWTH Aachen University, Germany
Jean-Christophe Nebel	Kingston University, UK
Vit Novacek	National University of Ireland, Galway
Cinzia Pizzi	Università degli Studi di Padova, Italy
Clara Pizzuti	Institute for High Performance Computing and Networking (ICAR)-National Research Council (CNR), Italy
Nicole Radde	Universität Stuttgart, Germany
Maria Elena Renda	CNR-IIT, Italy
Stefano Rovetta	University of Genoa, Italy
Huseyin Seker	De Montfort University, Uk
Jiri Spilka	Czech Technical University in Prague, Czech Republic
Kathleen Steinhofel	King's College London, UK
Karla Stepanova	Czech Technical University, Czech Republic
Roland R. Wagner	University of Linz, Austria
Viacheslav Wolfengagen	Institute JurInfoR-MSU
Borys Wrobel	Polish Academy of Sciences, Poland
Filip Zavoral	Charles University in Prague, Czech Republic
Songmao Zhang	Chinese Academy of Sciences, China
Qiang Zhu	The University of Michigan, UK

Contents

Medical Terminology and Clinical Processes

From Literature to Knowledge: Exploiting PubMed to Answer Biomedical Questions in Natural Language 3
Pinaki Bhaskar, Marina Buzzi, Filippo Geraci, and Marco Pellegrini

Using Twitter Data and Sentiment Analysis to Study Diseases Dynamics 16
Vincenza Carchiolo, Alessandro Longheu, and Michele Malgeri

An Open Data Approach for Clinical Appropriateness 25
Mario A. Bochicchio, Lucia Vaira, Marco Zappatore, Giambattista Lobreglio, and Marilena Greco

Machine Learning in Biomedicine

A Logistic Regression Approach for Identifying Hot Spots in Protein Interfaces 37
Peipei Li and Keun Ho Ryu

The Discovery of Prognosis Factors Using Association Rule Mining in Acute Myocardial Infarction with ST-Segment Elevation 49
Kwang Sun Ryu, Hyun Woo Park, Soo Ho Park, Ibrahim M. Ishag, Jang Hwang Bae, and Keun Ho Ryu

Data Mining Techniques in Health Informatics: A Case Study from Breast Cancer Research 56
Jing Lu, Alan Hales, David Rew, Malcolm Keech, Christian Fröhlingsdorf, Alex Mills-Mullett, and Christian Wette

Artificial Neural Networks in Diagnosis of Liver Diseases 71
José Neves, Adriana Cunha, Ana Almeida, André Carvalho, João Neves, António Abelha, José Machado, and Henrique Vicente

How to Increase the Effectiveness of the Hepatitis Diagnostics by Means of Appropriate Machine Learning Methods. 81
Alexandra Lukáčová, František Babič, Zuzana Paraličová, and Ján Paralič

Ant-Inspired Algorithms for Decision Tree Induction: An Evaluation on Biomedical Signals 95
Miroslav Bursa and Lenka Lhotská

Poster Session

Microsleep Classifier Using EOG Channel Recording: A Feasibility Study . . . 109
Martin Holub, Martina Šrutová, and Lenka Lhotská

Author Index 115

Medical Terminology and Clinical Processes

From Literature to Knowledge: Exploiting PubMed to Answer Biomedical Questions in Natural Language

Pinaki Bhaskar, Marina Buzzi, Filippo Geraci(✉), and Marco Pellegrini

CNR, Institute for Informatics and Telematics, Via G. Moruzzi 1, Pisa, Italy
{pinaki.bhaskar,marina.buzzi,filippo.geraci,
marco.pellegrini}@iit.cnr.it

Abstract. Researchers, practitioners and the general public strive to be constantly up to date with the latest developments in the subjects of biomedical research of their interest. Meanwhile the collection of high quality research papers freely available on the Web has increase dramatically in the last few years and this trend is likely to continue. This state of facts brings about opportunities as well as challenges for the construction of effective web-based searching tools. Question/Answering systems based on user interactions in Natural Language have emerged as a promising alternative to traditional keyword based search engines. However this technology still needs to mature in order to fulfill its promises. In this paper we present and test a new graph-based proof-of-concept paradigm for processing the knowledge base and the user queries expressed in natural Language. The user query is mapped as a subgraph matching problem onto the internal graph representation, and thus can handle efficiently also partial matches. Preliminary user-based output quality measurements confirm the viability of our method.

1 Introduction

The steady increase of the amount of biomedical data available in internet accessible repositories in the last two decades has come together with a similar increase in internet accessible biomedical literature (full papers or abstracts). Moreover, this trend has been reinforced by the fact that a few entry points (like the MEDLINE/PubMed repository) are sufficient to obtain a fairly representative view of the worldwide high quality biomedical scientific production. This state of facts brings about opportunities as well challenges.

The opportunity is for researchers, practitioners and the general public to be constantly up to date with the latest developments and findings in their medical (sub)field of interest from authoritative sources. For example, the goal of "evidence-based medicine" (EBM) [21], that is the idea that the diagnosis of individual patients can be based on, or at least supported by, finding relevant facts via "ad-hoc" searches in the biomedical literature, is indeed made more realistic by the availability at a finger-tip of the whole corpus of recent literature.

M.E. Renda et al. (Eds.): ITBAM 2015, LNCS 9267, pp. 3–15, 2015.
DOI: 10.1007/978-3-319-22741-2_1

The challenge is mainly in bridging the gap between rather vaguely defined and informally expressed user needs and the rigidity of standard traditional key-word based search engine interfaces currently in use.

To cope with this challenge a new generation of IR systems has emerged, which are based on formulating questions in unstructured natural language, and expecting as an answer a specific concise description in natural language of a relevant fact extracted from the appropriate knowledge base (QA systems).

A recent development, which witness the pressing need for technological improvements in Biomedical QA systems, has been the organization of dedicated QA tool evaluation challenges (e.g. http://www.bioasq.org/ now in its third year, and the medical track of CLEF http://www.clef-initiative.eu/).

In this paper we describe a proof-of-concept system with the intent of exploring the potentiality of novel graph representation of the knowledge-base and the use of efficient on-line sub-graph pattern matching algorithms in order to extract at query-time a pool of ranked relevant text snippets. This approach can be contrasted with that of a recent work by D. Hristovski et al. [13] where the the semantic relationship among named entities extracted from the off-line literature digestion are just fed to a database (thus missing the key graph abstraction). The user query in [13] is directly translated into a direct single query to the database, expressed with the sql-like Lucene DB query interface, oriented to exact matches. Our approach to query processing based on sub-graph pattern matching is more general and flexible, allowing for an easy expression also of partial matches.

2 Related Work

The first and most obvious way to solve a user's health information need is to ask a query to one of the generalist internet search engines. Wang et al. [26] compared usability and effectiveness of four general purpose search engines (namely: Google, Yahoo!, Bing, and Ask.com) for medical information retrieval, showing that current ranking methods need to be improved to support users in the task of retrieving accurate and useful information. Allam et al. [1] note that for controversial health issues the standard document ranking strategies adopted by generalist SE are at risk of introducing a strong cognitive bias. These two phenomena highlight the importance of developing simple user-friendly tools for direct access to authoritative medical information sources.

Specialized search engines based on medical literature such as MedSearch [17] and MedicoPort [5] have been developed within the general SE framework of a keyword search generating a list of snippets. In particular, Can et al. [5] focus on usability for untrained people, and show how increasing result precision is attained with various techniques including query conversions, synonyms mapping, focussed crawling, etc. Moreover the final ranking is based on a Pagerank-like method. Luo et al. [17] also addresses similar challenges using various techniques to improve its usability and the quality of search results. Lengthy queries are converted into shorter representative queries by extracting

a subset of important terms to make the query processing faster while improving the results' quality. Moreover, their system provides diversified search results by aggregating documents with similar informative content.

A third approach that has similarities, but also key differences with the two cases mentioned above is that of Clinical Decision Support Systems (DSS) that use medical literature SE as one of their information sources [7,23]. The role of DSS is in supporting clinicians in medical tasks (diagnoses, therapies, research, case studies, etc.) and must make extensive use of many different information sources, besides the literature, and in particular of the phenotype of the individual patients for which a decision has to be made.

Extended surveys on the biomedical QA systems have been compiled by S.J. Athenikos and H. Han [2] and O. Kolomiyets and MF. Moens [16]. They recognize this research area as one of the most dynamic within the more general field biomedical text mining [8,22,27].

Specific attempts in the direction of user-friendly accurate medical and/or biological QA systems are embodied in the systems askHERMES[1] [6], EAGLi[2] [10] and HONQA[3] [9], which are well described in the comparative usability study [3].

3 Method

In this section we describe a novel framework aimed at answering a user question posed in natural language. As a knowledge base we used the set of open access articles (where we extracted title and abstract) provided by PubMed in the NCBI ftp site.

In short, our system works in four main steps (see Fig. 1 for a graphical representation). Firstly, we extract from the question a set of relevant concepts and the relationships among them, then we retrieve all the sentences in our knowledge base containing these concepts with the appropriate relationships, later we assign a relevance score to each sentence according to the predicted meaning of the question, finally we merge sentences belonging to the same article into a single snippet, rank them and present the results to the user.

We modeled each sentence of a document as a graph where each node corresponds to a concept and an edge connects two concepts if they are related within the document. As dictionary of concepts we used the MeSH ontology. We determined the relationships among concepts using a customized set of rules. Then, we merged all the generated graphs into a single master graph collapsing all the nodes corresponding to the same concept into a single node. We used the graphDB model to store the graph into a MySQL database.

Once the user poses a new question, we convert it into a graph by extracting the relevant concepts and their relationships. The retrieval of all the possibly relevant sentences (and the corresponding articles) reduces to a sub-graph matching

[1] http://www.askhermes.org/.
[2] http://eagl.unige.ch/EAGLi/.
[3] http://services.hon.ch/cgi-bin/QA10.

problem. In our system we used the method described in [14] since it is based on the graphDB model. Than, we assign a score to each retrieved sentence according to its relevance to the question.

Our score is based on the relative position of terms in the phrase, on the classical TF-IDF, and on the type of the terms (biomedical/not biomedical). In particular, we give high score to terms according to their presence inside: the question, the MeSH dictionary or both.

In order to revert to articles instead of sentences as base result unit, we merge together in a single snippet the sentences belonging to the same article, averaging their score. Finally we rank the snippets by sorting them in decreasing score order and return them to the user.

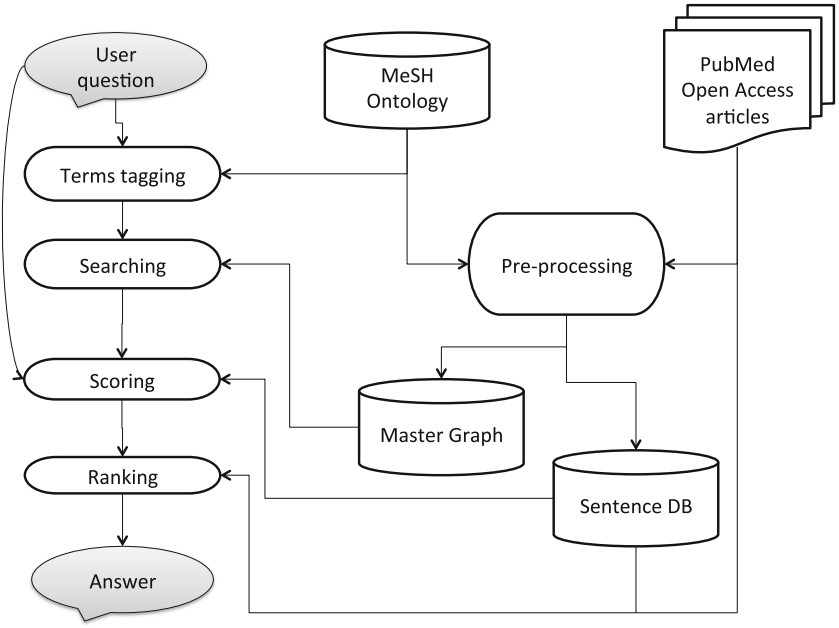


Fig. 1. Architecture of our QA framework

3.1 Knowledge Base Creation

The acceptance in the large majority of biomedical communities of the open access publication policy had the effect of making available a large collection of research articles that can be freely downloaded and used as a knowledge base. In particular, we focused on the *Open Access Subset* dataset provided by PubMed in the NCBI ftp site. This dataset is frequently updated and, at the moment we downloaded it, consisted in more than 790 thousand documents belonging to about 4 thousand journals for a total size of 52.92 GB.

Our knowledge base consists in a large undirected graph where each node corresponds to a well-defined biomedical concept or named entity, and the edges

model the relationship among pairs of concepts. For our retrieval purposes we did not consider the direction of the relationship and we did not assign a category to the nodes. Although this could appear as a simplistic approach, it guarantees a higher matching probability for small questions.

As a dictionary of biomedical terms we exploited the MeSH ontology. MeSH is a vast thesaurus organized as a hierarchy with 16 main categories including: anatomy, organisms, diseases, chemicals, drugs, processes therapeutic techniques and equipment. MeSH associates to each term a list of *entry terms* which consist in synonyms and alternate forms that can be used interchangeably. We collapsed all the entry terms into a single node of our master graph. Including articles in our knowledge base, we restricted to the title and abstract only. As observed in [4] this is not a limiting choice since most of the concepts exposed in a paper can be inferred from its title and abstract, while the rest of the paper contains only technical details. As a result, this restriction did not affect the effectiveness of our system.

As first step to include articles (from now on we refer to them indicating only title and abstract) in the knowledge base we split them into sentences. We treated titles as single phrases. Each sentence is scanned moving a sliding window and searching MeSH with the corresponding text. Since terms can have variable length ranging from a single word to five or six words, we used sliding windows of different sizes so that to capture all the possible medical terms inside the sentence. Notice that scanning with different window length can be done in parallel to avoid blowing the computational cost up.

More often multi-word terms are specific concepts where each component is itself a medical term. For example, the term *adult stem cell* includes the term *cell* which is more general and the term *adult*. In our case, however, using the most general term can determinate an increase of the number of retrieved irrelevant documents. In fact, in the above example, the focus of the term is not the cell in general, but a specific category of cells. Complicating the things, the term *adult* defines a concept belonging to a different MeSH category. However, considering it as an independent concept can be misunderstanding because in the above example it is used as an adjective to better specify the category of stem cells. In order to remove these general or misunderstanding terms, once we parsed a sentence we checked for nested terms (i.e. terms completely included into another term) and removed those included in a longer term.

A different situation arises when two concepts overlap but they are not nested. For example, the snippet sentence *tandem repeats sequence analysis* contains two overlapping terms *tandem repeats sequence* and *sequence analysis*. In this case tandem repeats are the object of the sequence analysis and thus both terms are relevant and related. The overlap is only the result of the fact that the two concepts are close in the text and have a word in common. In addition, the concept *tandem repeat sequence* has the alternative form *tandem repeat* in its entry terms list. Using this alternative form the two concepts of the example stop to overlap.

Inferring relationships among relevant terms in a text is a key point in the natural language processing field and general rules are far apart. In absence

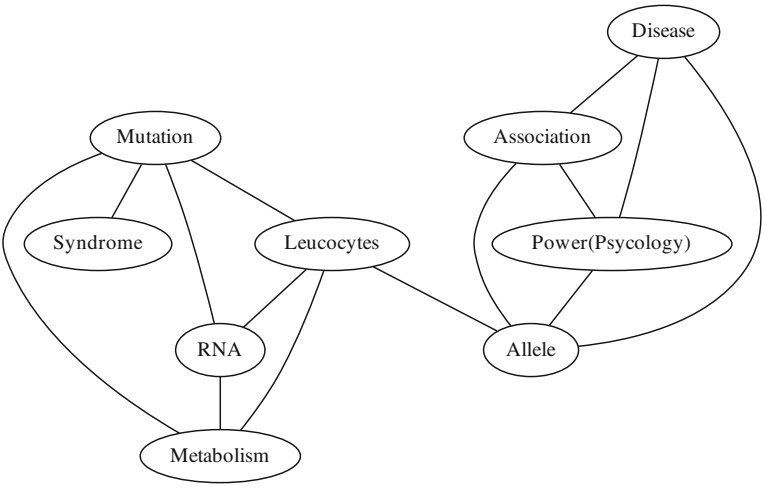


Fig. 2. Generated graph for two abstracts

of manually curated training data that enable the use of statistical methods, only rule-based algorithms are available [20]. Following this latter approach we observed that a minimal set of rules is able to correctly capture most of the relationships among medical concepts contained in an abstract. We observed that the distance between two related concepts is quite small and never exceeds the sentence. If a certain term is semantically connected to two concepts belonging to different sentences it is repeated in both phrases. According to this observation we constrained our system to check relationships only among terms within the same sentence.

Another agreed statement in natural language processing is that relationships are most often expressed with verbs. Motivated from this assumption, some effort has been done to create manually curated resources that are freely available. Among them, VerbNet [15] is one of the most complete for English. We observed that, in the abstracts, the verb is likely to be between the two related terms. The most frequent exception is when, instead of a single term, a sentence contains: two terms connected through a conjunction, or a list of terms. We treated lists as if they were a single concept. For example, in the sentence: *indels* and *SNPs* are common genomic *mutations*, we connect independently both the terms *indels* and *SNPs* with *mutations*, but we do not connect them among each other.

In Fig. 2 we show an example outcome of the processing of two abstracts.

3.2 Data Structure Organization

Data organization is a particularly important part of the structural design of those applications required to manage large datasets. Different data structure designs, can lead to a drastic speeded up of certain operators at the cost of a lower performance of other operations. In our QA system the three biggest data structure store: the sentences, the MeSH ontology and the master graph. Except at pre-processing time, all these structures do not require any on-line update.

Since the retrieval of sentences does not require full text searching, but it is only based on a pattern matching over the master graph, we can store them into a SQL table. We used a unique external key to associate the nodes of the master graph (namely the biomedical terms) with the sentences where they have been identified.

MeSH terms are accessed at query time to identify biomedical terms in the question. This operation can be efficiently done using an inverted index. Standard DBMSs (i.e. MySQL, postgreSQL) allow indexing string fields using the standard *select* command for querying. As a result, we were able to map the MeSH ontology into an SQL table.

Efficiently managing a large graph with thousands of nodes and millions of edges can be a thorny task. Trivial solutions like the adjacency matrix or the adjacency list are impractical for memory or performance reasons. We used the graphDB model [12] that exploits the DBMS strength to store the graph into a SQL database. According to this model, each node and its annotations are stored into a table and the edges are stored into another table. One of the practical advantages of the graphDB model is that it provides a simple and scalable way to store large graphs without requiring explicit handling of low-level details such as the memory management and the possible concurrent access. As shown in [14], another important characteristic of this model is that all the instances of a pattern can be efficiently retrieved using a limited number of *select* operations and a single *join*.

3.3 Question Answering Process

The first step of the question answering process consists in translating the user question into a graph. This process is done using the same procedure described in Sect. 3.1 for abstract preprocessing.

Searching. We use the question graph is as a model to perform a structural pattern-matching search using the algorithm proposed in [14]. In short, the algorithm works as follows: given an arbitrary pivot node, the question graph is divided by mean of a depth-first visit into a limited number of paths originating from the pivot. Each of these paths is retrieved with a single SQL *select* operation. Returned paths are pruned and merged using a SQL *join*. Since in our master graph nodes corresponding to the same concept have been collapsed into a single node, the output of the searching procedure consists in a single resulting graph.

Reverting from the subset of retrieved nodes on the master graph to sentences is done maintaining a mapping between concept nodes and IDs of the sentences where the concepts have been identified. According to the approximate pattern-matching model [11] it is suffice that all the retrieved nodes map a certain sentence ID to add it to the list of results.

Scoring. In order to evaluate the relatedness of a result sentence (from now on referred as answer) with the question, our system tokenizes the answer and identifies three categories of terms:

- **question's medical term**: that are biomedical terms belonging to both the question and the answer;
- **general medical term**: that are biomedical terms belonging to the answer but not present in the question;
- **general question's term**: that consist in the words of the question not tagged as medical terms.

Only answer's terms belonging to one of these three categories will contribute to the scoring and thus to the final ranking. Our score leverage on: the distance from the boundaries of the sentence, the TF-IDF, and the type (medical/non medical) of the terms. Let $A = \{t_1, t_2, \ldots, t_n\}$ be an answer with n terms and t_i be the i-esim term of A. We define the relatedness of the answer A with the question Q as:

$$R(A,Q) = \sum_{i=1}^{n} \left(\lambda(t_i) + i \left(1 - \frac{i-1}{n} \right) + \frac{tf(t_i)}{df(t_i)} \right) \times b(t_i) \tag{1}$$

where $\lambda()$ is:

$$\lambda(t_i) = \begin{cases} \lambda \text{ if } \neg\exists j < i \quad : \quad t_j = t_i \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

the boosting factor $b()$ is:

$$b(t_i) = \begin{cases} 5 \text{ if } t_i \text{ is a question's medical term} \\ 3 \text{ if } t_i \text{ is a general medical term} \\ 1 \text{ if } t_i \text{ is a general question's term} \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

and $\frac{tf(t_i)}{df(t_i)}$ is the standard TF-IDF for the term t_i in the knowledge base.

The rationale of the function $\lambda()$ is that of assigning an extra score to the first occurrence of a new term. As a result, an answer matching all the question's terms will have a higher score than an answer matching multiple times a subset of the question's terms. We empirically set $\lambda = 50$ so that to ensure that the score of a term occurring multiple times has a low probability to become greater than the sum of the scores of two terms occurring only once.

In order to define a priority order among the three categories of terms, the boosting factor $b()$ is introduced in Eq. (1). In absence of a formal criterion, the values of the boosting factor $b()$ have been empirically set.

Except stop words, general question terms are very relevant words that play an important role to clarify the meaning of a question and thus their contribution to the relevance score is essential. However, general words are more common than medical concepts and, thus, the term $\lambda()$ in Eq. (1) can dominate the overall score. In order to avoid this problem, we set $\lambda = 0$ for this category of terms.

Ranking. After scoring we merge together all the sentences belonging to the same article into a single snippet averaging their score. Then we rank the snippets by sorting them in decreasing score order. Finally, the ranked snippets are returned to the user interface for displaying.

4 User Interface for Biomedical Q&A

An intuitive user interaction is a key aspect for the success of a searching interface. A visual and aesthetic experience, emotion, identification, stimulation, meaning and value are main drivers of the user experience [18]. The variety of goals and skills of medical literature searching audience highlights the need to provide an interface available on different devices and browsers.

We designed our system user interface applying participatory design techniques, fast prototyping and early test with a restricted set of users (two people with biomedical background one computer scientist). This working organization allowed us a fast refinement loop of the system interface and functionalities.

We used HTML5, CSS3, and Javascript to develop a device independent and responsive UI. Our user interface has been designed as to enable a simple interaction and assuring a high level of accessibility and usability also for people with different abilities. In particular, we matched the constraints to assure an easy interaction also using assistive technologies (i.e. screen readers and voice synthesizers) for navigating the interface.

The interface design is minimalist and includes only a box for the search and the underlying results area. The result interface is split in 2 logical sections: the search box and the results, that are marked using WAI-ARIA role search (within the form) and the main with associated the aria-labelledby "Results". Furthermore heading level ($< h1 >$) can be associated to each result to enable easy jump from one to another using a screen reader command.

Focus is another important element to make interactions easier. Once the search page is loaded, we set the focus on the search box enabling the user to immediately formulate the question. Once the results are presented to the user, the focus moves to the first result to facilitate exploration.

5 Experiments and Validation

In this section we report details about our tests on the ability of our system to retrieve a relevant answer to user questions. As for most information retrieval tasks, in absence of a manually curated golden standard, the evaluation of a question answering system is a complicated task. The subjective nature of the concept of relevance makes the creation of a golden standard a hard task [24] and the formal assessment of a QA platform impractical. Among the few experimental assessment strategies, user evaluations are often preferred.

As mentioned in Sect. 3.1 we used the PubMed's *Open Access Subset* dataset to build the knowledge base. After preprocessing we obtained a graph with $27,149$ nodes and $4,732,885$ edges. We carried out a user study where we evaluated the system effectiveness for 96 well-formulated questions submitted from human experts. We requested the user to inspect the first ten results of each question and judge whether the answers are appropriate. Judgment is a score in the range $(1, 5)$ where 3 is the minimum score to consider the answer relevant. We left to the user the option of not evaluating some answers. Our evaluation

strategy allowed us to reduce the effort of the experts at the cost of the impossibility to measure the system's recall.

5.1 Evaluation Metrics

Let $C_k(Q) = \{c_1, \ldots, c_k\}$ be outcome of the judgment of the top k answers to a certain question Q, we define

$$N_k(Q) = \frac{\max_{i=1}^{k} c_i}{5}$$

as the normalised correctness rate of the question Q. The score $N_k(Q)$ is bounded in the range $[0, 1]$ and it is maximum only if there exists an answer with highest judgment score among the first k answers. Let n be the overall number of evaluated questions, we define the overall correctness rate N_k as

$$N_k = \sum_{i=1}^{n} N_k(Q_i).$$

We exploited different metrics to evaluate different aspects our system. In particular, we used: an extended version of the $c@1$ introduced in 2009 in the ResPubliQA exercise of the CLEF[4] initiative [19], the standard accuracy measure, and the mean reciprocal rank (MRR) [25].

Given a set of n questions, $c@1$ is a measure of the global correctness of the answers returned by a QA system. We extended the original formulae to the case of systems that return k answers as follows:

$$c@k = \frac{1}{n}\left(N_k + |U|\frac{N_k}{n}\right) \tag{4}$$

where U is the subset of unanswered questions. Notice that, since in the CLIF exercise the judgment of answers was binary (correct/not correct), the term N_k in the original formulae reduces to the number of correctly answered questions.

As a second performance measure, we used the *accuracy*@k which is a traditional measure applied to QA evaluations. We introduced a slight modification to the formulae in order to remove from the denominator unanswered questions.

$$accuracy@k = \frac{N_k}{n - |U|}$$

Ranking per se is one of the most important components of a QA system. The above measures give limited information about the ranking quality that can be derived only comparing the measures for different assignments of k. We included in our evaluation also the mean reciprocal rank (MRR) that provides a quantification of the ranking quality. Let $Rank_i$ be the position of the first correct answer for question Q_i, in our case we classify an answer as correct if the evaluator assigned a score higher or equal to 3. The mean reciprocal rank is defined as:

[4] http://www.clef-initiative.eu/.

$$MRR = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{Rank_i}$$

5.2 Results

Table 1 reports the result of the evaluation with the three measures for different values of k.

Table 1. Evaluation results

Measure	$k = 1$	$k = 5$	$k = 10$	$k > 10$
$c@k$	0.72	0.76	0.83	0.94
$accuracy@k$	0.56	0.67	0.78	0.91
MRR	0.66			

The MRR value 0.66 indicates that, on average, a good scoring answer is found either in the first or the second position in the returned list of results. The $c@k$ value for $k = 1$ and $k = 5$ indicates that about 85 % of the times an high quality answer is found within the top 5 answers. These measures attest the good quality of the ranking function. The $c@k$ value for $k > 10$ indicates that a good scoring answer is found for about 95 % of the queries. This indicates the quality of the subgraph matching strategy.

A manual inspection of the cases of failure (only 11 of the 96 questions remained unanswered) indicates that these are mostly due to mistakes in the first steps of the query processing procedure relative to named entity recognition with the MeSH ontology, due, for example, to the use of acronyms or commercial names of drugs. We reckon that we can improve our performance by integrating specialized knowledge data bases within our framework for specific medical sub-domains.

6 Conclusions

This work presents a proof-of-concept system for a QA search engine in the medical domain based on a Natural Language user interface and on an efficient partial sub-graph pattern matching methodology. Though the first evaluations are encouraging, thus establishing the valitdity of the approach, there is scope for improvements by incorporating additional domain knowledge and by refining the named-entity recognition step. We plan this as future work.

Acknowledgments. We acknowledge the support of the Italian Registry of ccTLD ".it" and the ERCIM 'Alain Bensoussan' Fellowship Programme.

References

1. Allam, A., Schulz, P., Nakamoto, K.: The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating google output. J. Med. Internet Res. **16**(4), e100 (2014)
2. Athenikos, S.J., Han, H.: Biomedical question answering: a survey. Comput. Meth. Prog. biomed. **99**(1), 1–24 (2010)
3. Bauer, M., Berleant, D.: Usability survey of biomedical question answering systems. Hum. Genomics **6**(1), 17 (2012)
4. Bleik, S., Mishra, M., Huan, J., Song, M.: Text categorization of biomedical data sets using graph kernels and a controlled vocabulary. IEEE/ACM Trans. Comput. Biol. Bioinf. **10**(5), 1211–1217 (2013)
5. Can, A.B., Baykal, N.: Medicoport: a medical search engine for all. Comput. Meth. Programs Biomed. **86**(1), 73–86 (2007)
6. Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J.J., Ely, J., Yu, H.: Askhermes: an online question answering system for complex clinical questions. J. Biomed. Inf. **44**(2), 277–288 (2011)
7. Celi, L.A., Zimolzak, A.J., Stone, D.J.: Dynamic clinical data mining: search engine-based decision support. JMIR Med. Inf. **2**(1), e13 (2014)
8. Cohen, A.M., Hersh, W.R.: A survey of current work in biomedical text mining. Briefings Bioinf. **6**(1), 57–71 (2005)
9. Cruchet, S., Gaudinat, A., Boyer, C.: Supervised approach to recognize question type in a QA system for health. Stud. Health Technol. Inf. **136**, 407–412 (2008)
10. Gobeill, J., Patsche, E., Theodoro, D., Veuthey, A.L., Lovis, C., Ruch, P.: Question answering for biology and medicine. In: 9th International Conference on Information Technology and Applications in Biomedicine, 2009. ITAB 2009, pp. 1–5, November 2009
11. Gori, M., Maggini, M., Sarti, L.: Exact and approximate graph matching using random walks. IEEE Trans. Pattern Anal. Mach. Intell. **27**(7), 1100–1111 (2005)
12. Güting, R.H.: GraphDB: modeling and querying graphs in databases. In: VLDB, vol. 94, pp. 12–15. Citeseer (1994)
13. Hristovski, D., Dinevski, D., Kastrin, A., Rindflesch, T.C.: Biomedical question answering using semantic relations. BMC Bioinf. **16**(1), 16 (2015)
14. Kaplan, I.L., Abdulla, G.M., Brugger, S.T., Kohn, S.R.: Implementing graph pattern queries on a relational database. Technical report, Lammerce Livermore National Laboratory (2008)
15. Kipper, K., Korhonen, A., Ryant, N., Palmer, M.: Extending verbnet with novel verb classes. In: Proceedings of LREC, vol. 2006, p. 1. Citeseer (2006)
16. Kolomiyets, O., Moens, M.F.: A survey on question answering technology from an information retrieval perspective. Inf. Sci. **181**(24), 5412–5434 (2011)
17. Luo, G., Tang, C., Yang, H., Wei, X.: Medsearch: a specialized search engine for medical information retrieval. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 143–152. ACM (2008)
18. Nielsen, J.: Designing Web Usability: The Practice of Simplicity. New Riders Publishing, Thousand Oaks (1999)
19. Peñas, A., Forner, P., Sutcliffe, R., Rodrigo, Á., Forăscu, C., Alegria, I., Giampiccolo, D., Moreau, N., Osenova, P.: Overview of respubliQA 2009: question answering evaluation over european legislation. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 174–196. Springer, Heidelberg (2010)

20. Regev, Y., Finkelstein-Landau, M., Feldman, R., Gorodetsky, M., Zheng, X., Levy, S., Charlab, R., Lawrence, C., Lippert, R.A., Zhang, Q., Shatkay, H.: Rule-based extraction of experimental evidence in the biomedical domain: the KDD cup 2002 (task 1). SIGKDD Explor. Newsl. **4**(2), 90–92 (2002)
21. Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B., Richardson, W.S.: Evidence based medicine: what it is and what it isn't. BMJ **312**(7023), 71–72 (1996)
22. Simpson, M.S., Demner-Fushman, D.: Biomedical text mining: a survey of recent progress. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 465–517. Springer, New York (2012)
23. Soldaini, L., Cohan, A., Yates, A., Goharian, N., Frieder, O.: Retrieving medical literature for clinical decision support. In: Hanbury, A., Kazai, G., Rauber, A., Fuhr, N. (eds.) ECIR 2015. LNCS, vol. 9022, pp. 538–549. Springer, Heidelberg (2015)
24. Voorhees, E.M., Tice, D.M.: Building a question answering test collection. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2000, pp. 200–207. ACM, New York (2000)
25. Voorhees, E.M., et al.: The TREC-8 question answering track report. In: TREC. vol. 99, pp. 77–82 (1999)
26. Wang, L., Wang, J., Wang, M., Li, Y., Liang, Y., Xu, D.: Using internet search engines to obtain medical information: a comparative study. J. Med. Internet Res. **14**(3), e74 (2012)
27. Zweigenbaum, P., Demner-Fushman, D., Yu, H., Cohen, K.B.: Frontiers of biomedical text mining: current progress. Briefings in Bioinf. **8**(5), 358–375 (2007)

Using Twitter Data and Sentiment Analysis to Study Diseases Dynamics

Vincenza Carchiolo, Alessandro Longheu(✉), and Michele Malgeri

Dip. Ingegneria Elettrica, Elettronica e Informatica,
Università Degli Studi di Catania, Catania, Italy
{vincenza.carchiolo,alessandro.longheu,
michele.malgeri}@dieei.unict.it

Abstract. Twitter has been recently used to predict and/or monitor real world outcomes, and this is also true for health related topic. In this work, we extract information about diseases from Twitter with spatio-temporal constraints, i.e. considering a specific geographic area during a given period. We exploit the SNOMED-CT terminology to correctly detect medical terms, using sentiment analysis to assess to what extent each disease is perceived by persons. We show our first results for a monitoring tool that allow to study the dynamic of diseases.

Keywords: Health Information Systems (HIS) · Twitter · Natural Language Processing (NLP) · SNOMED-CT · Sentiment analysis

1 Introduction

The amount of digital health related data [6] is becoming more and more huge, being generated both by healthcare industries [23] (e.g. medical records and exams) as well as by social media and virtual networks, where individuals share their experiences and opinions about different topics, including personal health (illnesses, symptoms, treatments, side effects).

While data owned by healthcare industries are often accessible only with restrictions, social media data are generally publicly available, therefore they represent an enormous resource for mining interesting healthcare insights. Among various social networks, the one on-the-edge is Twitter [29], the micro-blogging service whose restriction of 140 characters for post encouraged the development of a kind of shorthand and speed in composing messages.

Twitter has been recently used as an information source to predict and/or monitor real world outcomes [3], from extreme event analysis as the 2013 Syria sarin gas attack [31] or the earthquakes in Japan [24], to more playful scenarios as the inferring of U.S. citizens' mood during the day [22] or the forecast box-office revenues for movies [2].

This work was supported by the "Programma Operativo Nazionale Ricerca & Competitivita" 20072013 within the project "PON04a2C - Smart Health 2.0" supported by MIUR (Minister of Education, University and Research).

M.E. Renda et al. (Eds.): ITBAM 2015, LNCS 9267, pp. 16–24, 2015.
DOI: 10.1007/978-3-319-22741-2_2

Exploiting virtual social networks for healthcare purposes has been recently named with the neologisms *Infodemiology* and *Infoveillance* [8], and also Twitter has been exploited, as in [1] the micro-blog is used to detect flu trends, or in [25], where authors tracked and examined disease transmission in particular social contexts via Twitter data, or in [9], where social media improves healthcare delivery by encouraging patient engagement and communication.

In this paper, we monitor health related information using both Twitter data and medical terms present in the SNOMED-CT terminology [15], currently the most comprehensive medical terminology worldwide adopted. Tweets are considered within a specific geographic area, and we extract a (possibily continuous) stream of messages within a given time window, retaining just all those concerning diseases. Then, using natural language processing [12] and sentiment analysis techniques [10,17], we assess to what extent each disease is present in all tweets over time in that region. Our proposal therefore results in a monitoring tool that allow to study the dynamic of diseases.

Exploiting tweets for health-related issues is not new; in [27] authors present a practical approach for content mining of tweets that is somehow similar to our proposal except for the initial selection of keywords. Indeed, we do not outline in advance a list of *significant* keywords for tweets extraction, rather we adopt the SNOMED-CT collection to extract any health related tweets. Similarly, in [1] and [16] a predefined list of flu related keywords (e.g. "H1N1") is considered to accomplish its task, whereas we do not focus on a specific disease. In [13], the temporal diversity of tweets is examined during the known periods of real-world outbreaks for a better understanding of specific events (e.g. diseases). As in our case, time is considered, whereas topic dynamics is inferred using an unsupervised clustering technique (instead of the official SNOMED-CT cited previously); the use of sentiment analysis however is not considered.

The paper is organized as follows. In Sect. 2 we describe the overall architecture of our proposal, and how the data collection and analysis are performed. In Sect. 3 we show an application to a real case, providing concluding remarks and future works in Sect. 4.

2 Architecture

The overall architecture of our proposal is depicted in Fig. 1. As introduced in the previous section, the first step is the extraction of geolocalized tweets; to this purpose, we developed a Python application that extracts a stream of tweets both during a desired time period and within a given region (a box with specified NE and SW coordinates). Note that for better results, only geolocalized tweets have been considered; a less precise solution is to use the user's provided location but this could lead to misinformation when specified location is not correct.

After having collected tweets, we want to extract only those with health-related content, i.e. where at least a medical term is present. At this step, Natural Language Processing (NLP in Fig. 1) techniques are required to properly filter each tweet by:

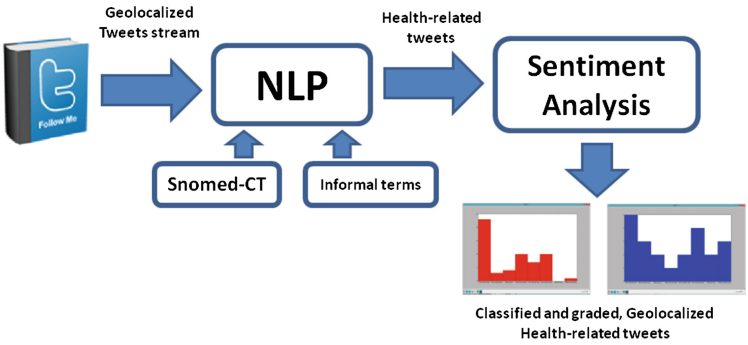


Fig. 1. Application architecture

- removing non-English tweets
- removing irrelevant information, as links, retweet details and usernames
- applying standard text processing operations as tokenization, stopwords removal, stemming and indexing [4].

2.1 Health-Related Tweets Extraction

In order to discard tweets that do not contain any medical term, we search for index terms in the SNOMED-CT terminology. To better clarify how this search is performed, we briefly cite the SNOMED-CT core components (details can be found in [26]) that are:

- *concepts*, that represent all entities that characterize health care processes; they are arranged into acyclic taxonomic hierarchies (according to a *is-a* semantics)
- *descriptions*, explaining concepts in terms of various clinical terms or phrases; these can be of three types, Fully Specified Names (FSNs) that is the main (formal) definition, Preferred Terms (PTs), i.e. the most common way of expressing the meaning of the concept, and Synonyms.
- *relationships* between concepts, e.g. the concept (disease) "Staphylococcal eye infection" has "Causative agent" relationship with "Staphylococcus" (different types of relationships exist depending on concepts type)
- *reference sets* used to group concepts e.g. for cross-maps to other standard purposes.

In this work, the first two items are considered. In particular, among all concepts hierarchies we focus in the "disorder/disease" since our goal is to detect tweets about diseases; therefore we do not consider other specific hierarchies (e.g. "surgical procedures"). Inside the disorder hierarchy, we search each index term extracted from tweets as a FSN, PT or synonym; if found, that tweet is further processed in order to establish to what extent the specified disorder is present using sentiment analysis (see below).

Note that to guarantee that all medical terms can be successfully detected, a list of additional informal terms is searched if nothing is found within SNOMED-CT. For instance, if the index term is the word "flu", this has positive match in the synonym list of "influenza" disease (the FSN), but the (also quite common) term "headache" is not explicitly present when browsing SNOMED-CT [11], where this disorder is instead referred as "migraine" both as FSN and its synonym. Including "headache" in an additional list (named "informal terms" in Fig. 1) is the simple solution we adopted; this list is considered just if nothing is found within SNOMED-CT.

Also note that several diseases are defined as a group of words (e.g. "Viral respiratory infection"), therefore during the indexing phase we also retainN-grams with N=2 and 3; diseases with more than three words can be easily disambiguated even with 3 words since not all words are generally significative (e.g. in "Disease due to Orthomyxoviridae" the first and the last words are enough for correct matching).

Finally, detected diseases may be hierarchically related, e.g. "influenza" and "pneumonia" are both children of "Viral respiratory infection" according to the "is-a" semantics. This information could be used for instance by replacing both children with their common parent, in order to build a more generalized, global view of diseases named in the given geographic area during the chosen time period. We choose however to preserve the best level of detail by not using a common ancestor as in the example, while on the other hand we will substitute all terms that represent the same disease with its FSN as indicated in SNOMED-CT. For instance, if different tweets refer to "flu", "grippe" and "influenza" they will be all considered as tweets about "influenza".

2.2 Tweets Classification

The next phase is the use of sentiment analysis in order to establish to what extent the disease detected in that tweet is present. Sentiment analysis or opinion mining [20] leverages NLP, text analysis and computational linguistics to extract subjective information, as the mood of the people regarding a particular product or topic; basically, the sentiment analysis can be viewed as a classification problem of labelling a given text (e.g. a statement within a tweet) as *positive*, *negative* or *neutral*.

Opinion mining has been applied to twitter data in several context, e.g. [2], where tweets are used to predict revenues for upcoming movies, or [7], where tweets allow to guess the political election results during U.S. presidential debate in 2008. Several approaches are adopted to perform sentiment analysis; typically, these are (1) machine learning algorithms with supervised models, where training examples labelled by human experts are exploited, or (2) unsupervised models, where classification is performed using proper syntactic patterns used to express opinions.

In the work here described we choose the latter approach. In particular, we first extract main statements from each tweet using the NLTK chunking package [19]; chunking, also called shallow parsing, allows to identify short phrases (clusters)

like noun phrases (NP) and verb phrases (VP), thus providing more information than just the parts of speech (POS) of words, but without building the full parse tree of the whole text (tweet). For instance, in the tweet "Last night was too rainy, this morning my headache is stabbing but fortunately my little syster has got over her terrible flu", the package produces the following chunks:

"Last night"(NP)
"was" (VP)
"too rainy" (NP)
"this morning" (NP)
"my headache" (NP)
"is stabbing" (VP)
"but fortunately" (NP)
"my little syster" (NP)
"has got over" (VP)
"her terrible flu" (NP).

Basically, the sentiment analysis we exploit to discover disease searches for them into NPs chunks (in the example, "headache" and "flu"), while the presence or absence of that diseases can be derived by analyzing VPs chunks. Therefore, in the tweet example the headache is present, while the flu is cited but no more present. We use a proper list of *positive* and *negative* verbs to this purpose, obviously taking into account negative verbal forms and propositions to guarantee a correct detection. In addition to the basic mechanism described here, we also estimate to what extent the given disease is present or not combining the linguistical distance (in terms of NP/VP chunks) between the disease and its associated verb and a proper rank we assigned to verbs and disease adjectives. In the example above, "terrible" and "is stabbing" both increase the relevance of their associated disease (details can be found in [5]). We exploit this estimation together with the number of tweets concerning a given disease in order to approximate its impact, e.g. assessing whether *few* people have *terrible* flu or *many* people are *few* cold in a given area during the monitoring time period.

Note that for each tweet, a set (generally small due to the limited lenght of tweets) of diseases could be detected. We do not associate however persons (twetter users) with diseases, rather we aim at achieving a "global" vision of the health status in the monitored area; an example of first results is provided in the following section.

3 Results

In this section we show how the approach illustrated in previous sections has been implemented to get first results.

The Python application we developed made use of the Tweepy libraries [28] and Twitter Stream APIs [30] to extracts the stream of tweets on March 2015 (1 month) within the area of New York City, delimited as a box with proper NE and SW coordinates (see Fig. 2); the OAuth APIs [14] has been used for authentication.

The total number of tweets collected was about 178,000 generated by about 60,000 unique users.

Fig. 2. The geographic area considered

Tweets have then been processed with the NLTK python based platform [18] to perform all text-processing operations described in the previous section; SNOMED-CT and the additional informal medical terms allow to isolate health related tweets, while the next phase (i.e. sentiment analysis) classify tweet statements (chunks) to assess whether and how diseases are present.

A list of all diseases extracted can be used to examine each one of them. In Fig. 3 the list of the most relevant diseases detected is shown, each with the number of tweets that contains at least a chunk referring to that disease.

Rank	Disease	# of tweets
1	Cold	36800
2	Headache	28200
3	Influenza	14700
4	Pneumonia	4500
5	Laryingis	3200

Fig. 3. The list of most detected diseases

As indicated in previous section, for each disease we also tried to estimate to what extent it is present at a given time. For instance in Fig. 4 we show how influenza is perceived by persons during March in the entire area examined. The two highest value detected from tweets concern the case where people healed from influenza (about 4500 tweets) and the opposite, where people tweet about their serious flu (6420 tweets). We believe that people tend to tweet *significant* information and probably having just a little bit of influenza is generally considered not so relevant.

Filtering data with space and/or time constraints makes it possible to assess the evolution of that disease, e.g. in Fig. 5 we represent the number of tweets detected across the three 10-day slots of March for “influenza”, showing that there has been an increment of influenza outbreaks during the second decade.

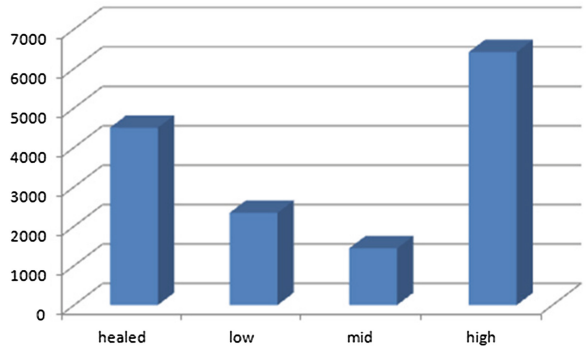


Fig. 4. # of tweets about "flu" in march 2015

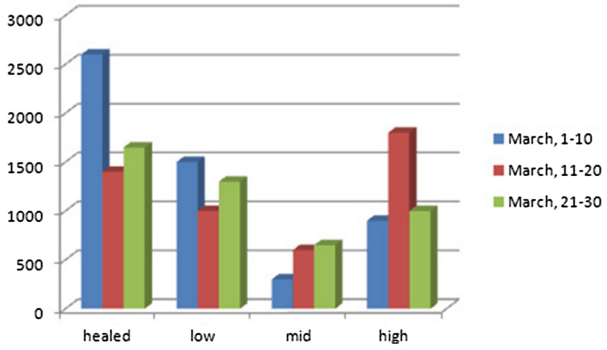


Fig. 5. The temporal evolution of influenza during march 2015

4 Conclusions

We introduced an approach to Tweeter data processing aiming at extracting health related information in a given area during an assigned period; this is achieved by also expoliting the SNOMED-CT medical terminology and sentiment analysis technique. The final goal is to get data for studying the spatio-temporal evolution of a selected disease in the area being considered, and first results are encouraging. We are considering other further questions as:

- the comparison with other existing proposal/tools, e.g. [21]
- the contribution that following and followers can provide to improve the accuracy and the meaning of collected data
- how profiling users (according to age, gender, residence area, device type...) leads to better (targeted) analysis; a related improvement is to address the biased demographic of users that could affect results (e.g. [32]).
- how to explore other sentiment anaysis methods, for instance combining lexical- and machine learning- based methods as suggested in [10], in order to improve the effectiveness of the proposed approach

– to gather a larger number of tweets (for instance, over a year or more) even in different geographical areas, to validate our proposal
– to more deeply explore SNOMED-CT, for instance by exploiting relationships between concepts for a more effective health-related tweets extraction.

References

1. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.H., Liu, B.: Predicting flu trends using twitter data. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 702–707, April 2011
2. Asur, S., Huberman, B.A.: Predicting the future with social media. CoRR abs/1003.5699 (2010). http://arxiv.org/abs/1003.5699
3. Atefeh, F., Khreich, W.: A survey of techniques for event detection in twitter. Comput. Intell. **31**(1), 132–164 (2015)
4. Baeza-yates, R., Ribeiro-Neto, B.: Modern Information Retrievial. ACM Press, Seattle (1999)
5. Carchiolo, V., Longheu, A., Cifalino, S.: Contestualizzazione spaziale di informazioni medico scientifiche tramite sensori sociali. DIEEI - Internal, Report (2015)
6. Cios, K.J., Moore, W.: Uniqueness of medical data mining. Artif. Intell. Med. **26**, 1–24 (2002)
7. Diakopoulos, N.A., Shamma, D.A.: Characterizing debate performance via aggregated twitter sentiment. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2010, pp. 1195–1198. ACM, New York (2010). http://doi.acm.org/10.1145/1753326.1753504
8. Eysenbach, G.: Infodemiology and Infoveillance. Am. J. Prev. Med. **40**(5), S154–S158 (2011). http://dx.doi.org/10.1016/j.amepre.2011.02.006
9. Fisher, J., Clayton, M.: Who gives a tweet: assessing patients interest in the use of social media for health care. Worldviews Evid.-Based Nurs. **9**(2), 100–108 (2012). http://dx.doi.org/10.1111/j.1741-6787.2012.00243.x
10. Gonçalves, P., Araújo, M., Benevenuto, F., Cha, M.: Comparing and combining sentiment analysis methods. In: Proceedings of the First ACM Conference on Online Social Networks, COSN 2013, pp. 27–38. ACM, New York (2013), http://doi.acm.org/10.1145/2512938.2512951
11. IHTSDO SNOMED CT Browser. http://browser.ihtsdotools.org/
12. Jackson, P., Moulinier, I.: Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization, 2nd edn. John Benjamins, Amsterdam (2007)
13. Kanhabua, N., Nejdl, W.: Understanding the diversity of tweets in the time of outbreaks. In: Proceedings of the 22nd International Conference on World Wide Web Companion, WWW 2013 Companion, pp. 1335–1342. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2013). http://dl.acm.org/citation.cfm?id=2487788.2488172
14. Kumar, S., Morstatter, F., Liu, H.: Twitter Data Analytics. Springer, New York (2013)
15. Lee, D., Cornet, R., Lau, F., de Keizer, N.: A survey of snomed-ct implementations. J. Biomed. Inform. **46**(1), 87–96 (2013). http://www.sciencedirect.com/science/article/pii/S1532046412001530

16. Lee, K., Agrawal, A., Choudhary, A.: Real-time disease surveillance using twitter data: demonstration on flu and cancer. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, pp. 1474–1477. ACM, New York (2013). http://doi.acm.org/10.1145/2487575.2487709
17. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014). http://www.sciencedirect.com/science/article/pii/S2090447914000550
18. Natural Language Toolkit. http://www.nltk.org/
19. Natural Language Toolkit chunk package. http://www.nltk.org/api/nltk.chunk.html
20. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008). http://dx.doi.org/10.1561/1500000011
21. Paul, M.: Discovering health topics in social media using topic models, April 2014. http://dx.doi.org/10.6084/m9.figshare.1007712
22. Pulse of the Nation. http://www.ccs.neu.edu/home/amislove/twittermood
23. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**(1), 3 (2014). http://dx.doi.org/10.1186/2047-2501-2-3
24. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 851–860. ACM, New York (2010). http://doi.acm.org/10.1145/1772690.1772777
25. Signorini, A., Segre, A.M., Polgreen, P.M.: The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. PLoS One **6**(5), e19467 (2011). doi:10.1371/journal.pone.0019467
26. Snomed, CT. http://www.ihtsdo.org/snomed-ct
27. Sunmoo Yoon, N.E., Bakken, S.: A practical approach for content mining of tweets. Am. J. Prev. Med. **45**(1), S122–S129 (2013)
28. Tweepy - A Python library for accessing Twitter API. http://www.tweepy.org/
29. Twitter. http://www.twitter.com/
30. Twitter Streaming APIs. https://dev.twitter.com/streaming/
31. Tyshchuk, Y., Wallace, W., Li, H., Ji, H., Kase, S.: The nature of communications and emerging communities on twitter following the 2013 syria sarin gas attacks. In: 2014 IEEE Joint on Intelligence and Security Informatics Conference (JISIC), pp. 41–47, September 2014
32. When Google got flu wrong. http://www.nature.com/news/when-google-got-flu-wrong-1.12413

An Open Data Approach for Clinical Appropriateness

Mario A. Bochicchio[1], Lucia Vaira[1(✉)], Marco Zappatore[1], Giambattista Lobreglio[2], and Marilena Greco[2]

[1] Department of Engineering for Innovation, University of Salento, Lecce, Italy
{mario.bochicchio, lucia.vaira, marcosalvatore.zappatore}@unisalento.it
[2] Department of Clinical Pathology, Vito Fazzi Hospital, Lecce, Italy
patologiaclinica.polecce@ausl.le.it, grecomarilena@gmail.com

Abstract. In recent years there have been partially unexpected qualitative and quantitative increase in clinical exams demand. Although on the one hand this is the positive result of better health awareness, mostly in terms of prevention, on the other hand it is the direct and logical consequence of the defensive behaviour, which arises from the potential occurrence of legal controversies and of the clinician's unawareness about the cost of examinations. To reduce the occurrence of unnecessary clinical tests we propose an approach based on Open Data and Open Software that can be adapted to existing medical information systems to enforce a suitable set of "appropriateness rules". The idea is to directly intervene at the moment of the request emission, in order to avoid unnecessary demands, which have no urgent and valid motivations and/or no value for patients.

Keywords: Open data · Clinical appropriateness · Open software · Rule engine

1 Introduction

The use of clinical laboratories has significantly increased over the last decades, while healthcare budgets worldwide are facing increasing pressure to reduce costs, improve efficiency and maintaining quality of care [1]. This is relevant because clinical laboratory practices contribute in a decisive fashion to the 70 % of the medical diagnoses and this means that the primary role of the laboratory in the diagnostic and clinical paths is by now certain, accepted and widely recognized. The largest sector in lab medicine in terms of test volume and revenue is clinical pathology [2] (66 %, $31.9 billion), followed by anatomic pathology (19 %, $9.0 billion) and molecular pathology/esoteric testing (8 %, $4.1 billion).

Many authors claim that too many laboratory tests are ordered in clinical practice. Daniels and Schroeder [3] found a 20-fold difference in laboratory utilization on patients with the same diagnosis, while others state that 30–50 % of tests conducted are required without valid motivations [4]. Several studies have suggested that

M.E. Renda et al. (Eds.): ITBAM 2015, LNCS 9267, pp. 25–33, 2015.
DOI: 10.1007/978-3-319-22741-2_3

inappropriate test requests are a primary reason for such an increase [5]. The rate of inappropriate test requests ranges from 4.5 % to 95 %, as shown in the systematic review of laboratory clinical audits by van Walraven and Naylor [6].

The appropriateness of clinical, or more generally, medical requests plays hence a key role in programs for quality improvement, a challenging task in the healthcare domain that can benefit from the use of a wide variety of tools and methods [7]. The increase in inappropriate requests stems from several factors including the routine clinical practice that leads to the adoption of strict protocols and guidelines; the defensive behaviour, which arises from the potential occurrence of legal controversies; the excessive frequency of repeat tests by apprentice medical staff because of uncertainty and clinician's unawareness about the cost of examinations [8]. Furthermore, by comparing hospitals from different countries, and those in the same country [9], great differences can be found in laboratory usage, which can be classified in at least four families of inappropriateness [10]:

- lack of knowledge about an already performed exam for a specific patient;
- unavailability or inaccessibility of previous test results;
- lack of knowledge about the response time for a given exam;
- doubts about the reliability of the obtained result.

In Italy, where the healthcare sector is essentially public, the Slow Medicine[1] movement, founded in 2001 [11], has launched the "*doing more does not mean doing better*" campaign similar to "Choosing Wisely"[2] in the United States, which aims to improve clinical appropriateness through the reduction of unnecessary tests and treatments. The campaign deems medicine as soaked with inappropriateness, wastes, conflicts of interest, and many cliché induce professionals and patients to consume more and more healthcare services in the illusion that this can improve health. The repeated request for tests is a component of the inappropriate usage of the laboratory that may be subject to evaluation and improvement initiatives. Several attempts to control inappropriate requests have been presented in literature so far, which included: rationing tests, redesigning of request forms, educating about appropriate tests for various conditions by discouraging futile repeat tests, educating about costs, issuing feedback information, and using protocols [12]. Unfortunately, the majority of these strategies has proven to be scarcely effective and those which have actually reduced requests were often been expensive in terms of time and/or manpower and have had no sustained effect once they were withdrawn.

In its broadest sense, an inappropriate request is one that should not be processed, generally because it is requested for the wrong patient, at the wrong time, in the wrong way, or is for the wrong test [13]. This last definition contains four basic concepts that can be summarized as it follows: do the right things, in the best way, at the right time to those who need it [1]. In other words:

- performing the right tests means choosing exams that are able to change the clinical/diagnostic/therapeutic practice;

[1] http://www.slowmedicine.it.

[2] http://www.choosingwisely.org/.

- performing tests in the best way implies the selection of the most suitable analytical methods and systems, by endorsing in the evaluation: sensitivity, specificity, accuracy, reliability, timing and productivity;
- performing tests at the right time means applying the appropriate diagnostic window in order to make the exam "clinically useful";
- performing tests to those who need (to the right patient) contains within itself the concept of efficiency: tests should be carried out taking into account two main attributes, that is the purpose and the optimal usage of resources.

Each clinical test has to respect some constraints (often of a temporal nature) in order to be appropriate but, at the same time, it has to be compatible with the patient status (drugs assumption, allergies, pathologies, nutrition …) as well. Such compatibility can be verified by adopting several kinds of mechanisms, but in general, the idea is to directly intervene at the moment of the request emission. To the best of our knowledge, this kind of "validation" is often provided as a supplementary feature by the commercial software adopted in the Operative Unit, under payment of additional fees. Therefore, due to the lack of resources and/or to political/institutional reasons, this service is often not taken into consideration. Another crucial issue in such a context is about the absence of open data and open rules on the clinical appropriateness.

For these reasons, in this paper we propose an approach, mainly based on Open Data and Open Software, which can be easily adaptable to existing clinical information systems in order to verify the appropriateness of laboratory test requests. Particular attention has been posed to sensitive information, which are mainly protected by applying proper anonymization techniques.

The paper is organised as follows: after the introduction in Sects. 1 and 2 presents background and related works in the field of clinical appropriateness looking for potential correlated studies. In Sect. 3 we delineate our proposal, analysing the potential solutions and presenting the software agent and the whole block architecture. In Sect. 4 we present a retrospective evaluation showing the potential savings reachable when using our proposed system. Finally the last Section is for conclusions.

2 Background and Related Works

Although the considerable relevance of the above-discussed "appropriateness" problem, only few contributions are available in literature on the topic. Efforts to remedy this problem have been tried for decades as well, but the problem still seems to exist [14]. In numbers, a search on Scopus,[3] IEEE Xplore,[4] ACM Digital Library[5] and PubMed[6] databases, for papers published from 1990 to 2015, returns 246 publications matching the "*clinical appropriateness*" pattern and 115 publications on "*medical appropriateness*". Most of them are "off-topic" or discuss about appropriateness

[3] Scopus Database, http://www.scopus.com/home.url.

[4] IEEEXplore Digital Library, http://ieeexplore.ieee.org/Xplore/home.jsp.

[5] ACM Digital Library, http://dl.acm.org/.

[6] PubMed Database, http://www.ncbi.nlm.nih.gov/pubmed.

considering a specific medical sector (not the "laboratory test" demands). Furthermore, most of the literature refers to the 90's, when the adoption of ICT on the theme was less developed. A synthetic report of the literature review process is depicted in Table 1 where each search pattern is associated to the number of contributions found, by differentiating them into "off-topic" publications and relevant ones. Among these works, the one by Charles et al. [15] is the most similar to our approach, although it dates back to 1998. In that paper, an Internet-based system for the construction and maintenance of ontologies for clinical appropriateness criteria is presented. The system allowed users to edit the indexing terms and the semantic network that form the ontology for a set of appropriateness criteria.

Table 1. Literature review report from 1990 to 2015

	Scopus			IEEE Xplore		
	Total	**off topic**	**relevant**	**Total**	**off topic**	**relevant**
clinical appropriateness	123	117	6	6	6	0
medical appropriateness	50	46	4	8	4	2
filters	Title, abstract, keywords			Full Text and metadata		

	ACM DL			PubMed		
	Total	**off topic**	**relevant**	**Total**	**off topic**	**relevant**
clinical appropriateness	7	7	0	110	103	7
medical appropriateness	17	14	3	40	36	4
filters	Any field			Any field		

3 Our Proposal

The aim of the proposed approach is to develop an Open Software Agent (OSA in the following) based on Open Data, easy to adapt to existing systems and able to verify the appropriateness of laboratory services requests, hence increasing the level of operators' awareness about the clinical tests and ensuring a better level of service.

3.1 Open Data Approach

Open Data holds a great potential in the health sector [16]. Their adoption for the definition of appropriateness criteria (or rules) allows to overcome the main limit of the existing commercial systems based on a pre-defined core of rules which are subject to obsolescence and not open to the scientific debate. On the other hand, the usage of open software and the resulting possibility to share the design, construction and maintenance costs among the interested users allows to overcome the problems related to the high costs of commercial systems. In this perspective, the collaboration of doctors, patients and pharmaceutical companies can help to continuously update and improve the appropriateness rules. This is in clear contrast with the existent systems, in which every single department or hospital needs to update its own private repository. The idea is to create an open and sharable "appropriateness rules" repository, which can improve the comprehension, facilitate the discussions on the topic and better support the validation

of rules. For example, assuming that doctors discover a new appropriateness rule (e.g. that screening a pregnant patient for hemoglobinopathy after the first pregnancy is redundant) then, after the approval of a scientific committee, this is added to the shared repository and immediately applied by all the OSA operating in the connected laboratories to block this kind of inappropriate requests.

3.2 Appropriateness Rules Management

The content of any clinical appropriateness criteria can be directly translated to IF-THEN rules. Considering the same example of hemoglobinopathy, the rule can be thought as the statement:

IF the patient is pregnant AND pregnancy IS NOT the first
THEN hemoglobinopathy is inappropriate

From a technical point of view, this can be managed in two different ways:

1. adopting a rule engine, which filters the exam requests and gives the adequate response (appropriate/inappropriate requests);
2. exploiting the existing conceptual model (a mapping between database and appropriateness rules). The rules are considered as SQL query statements, which return a Boolean value (yes/no) reflecting the appropriateness.

In the first case, the rule engine allows to separate business logic from application logic. The behaviour of the system can then be modified without code changes or recompilation/redeployment cycles. Rules are stored in a file so they can be changed with a rule editor and each rule consists of a conjunction of conditional elements corresponding to the IF part (left-hand-side or LHS) of the rule, and a set of actions corresponding to THEN part (right-hand-side or RHS) [17]. Data are stored in database connected with the main software for clinical appropriateness and the rule engine will pick the necessary data from the concerned tables. The data flow is straightforward: data representing each new request of clinical test is passed to the engine, rules are executed and later, if appropriate, each request is transmitted to the laboratory to be fulfilled. In our context, this can be done interactively, when requests are submitted by the personnel of each ward. In more details, each element of the request is evaluated by the rule engine, then the RHSs of those rules are evaluated and the request is approved/rejected. In the second case each appropriateness rule is described in terms of SQL queries. These queries are directly applied to request data. In case of inappropriateness the same queries can notify the violated constraint. The decision on which of the two possibilities is better is quite hard, depending on several factors (e.g. rules complexity, software and hardware constraints etc.). Rule engines are often considered easier to use and integrate than database tables. In fact, they can provide high flexibility since there are no queries, no tables, and no code. The rule engine controls all the logic, in addition rules are easier to understand than SQL code and they can be effectively used to bridge the gap between business analyst and developers. Finally, keeping rules in one place leads to a greater reusability. In summary, rule engines are considered appropriate for general setting. On the other hand, they also bring lots of extra costs, complexities and performance consumptions while performing checks on database with the help of

normal SQL queries can be less resource-demanding and more efficient. The conceptual model of the Database of Rules (DoR in the following) is shown in Fig. 1. A given "Rule Statement", which is expressed in natural language, can be composed of other, more basic "Rule Statements" and each rule is implemented in SQL by a "Rule Expression". SQL clauses (e.g. WHERE, HAVING, GROUP BY and ORDER BY) are represented in figure by the "Condition" class; logical operators (AND/OR/NOT) are used to combine conditions. In this way, a "Rule Expression" is composed by one or more "Condition". The appropriateness concept is here exposed by means of the "Error" class.

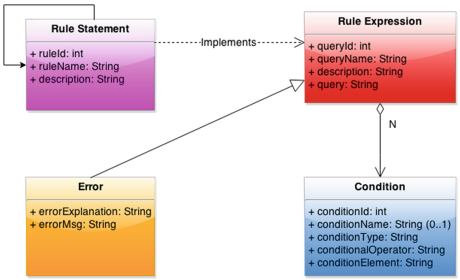


Fig. 1. Conceptual model of the database of rules

3.3 Architecture Overview

In this subsection we will discuss about the "as is" architectural model and the "to be" one in order to demonstrate the impact of usage for our proposed OSA. As represented in Fig. 2a, each single ward represents an applicant. Whenever a ward demands for an exam for a specific patient, it will use the management software shared by each ward in order to activate the exam request process. The request is taken over by the software installed in the Clinical Pathology Lab (CPL-Sw). Our software agent, named CLAP (CLinical APpropriateness) system, will be "placed before" the main software in the operative unit of clinical pathology as depicted in Fig. 2b. It will be in charge of

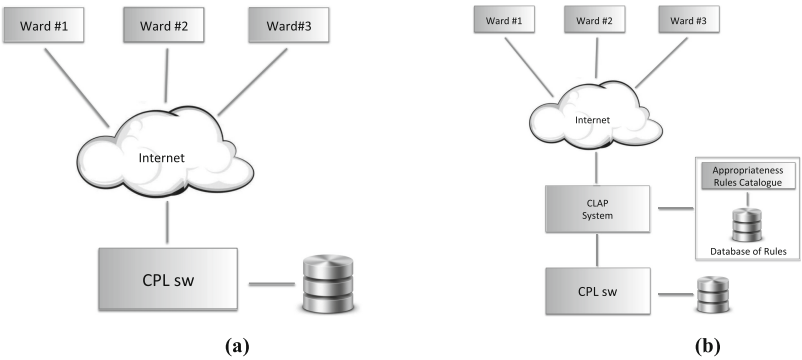


Fig. 2. (a) AS-IS and (b) TO-BE

checking the requests and, if appropriate, to send them to the CPL-Sw. The DoR will include the whole set of clinical appropriateness rules that are applied to each request. This is built as an open and sharable clinical appropriateness rules catalogue so that every operative units of each interested department could use it.

The behaviour of the CLAP System can be summarized as:

- check on the temporal distances between two test requests;
- proactive behaviour of the system (presenting the exam as already done and showing the result if the test validity period is still effective);
- check on the clinical profile of the patient (preventing the execution of inappropriate tests based on particular conditions).

4 Experimental Evaluation

The adoption of the CLAP system in clinical laboratories could allow significant savings for diagnostic tests. In order to prove this assertion we have performed a retrospective evaluation by analysing exam requests in 6 months (Sept 2014 – Feb 2015) in the unit of clinical pathology of the main hospital of Lecce, in Italy. Such operative unit generates costs of about 3 M€ per year (reagents, consumables, machinery maintenance, …) and serves nearly 1 million of patients. In Fig. 3 are reported the numbers and the percentage of test services in 2013 distinguished by diagnostic area.

	# tests	%
Clinical chemical and biomarker	1.780.376	72,92%
Coagulation	248.801	10,19%
Haematology, autoimmunity, allergology	190.005	7,78%
Hormonology	81.807	3,35%
Toxicology	55.694	2,28%
Electrophoforesis	32.254	1,32%
Virology	25.538	1,05%
Urine	27.082	1,11%
Total	2.441.557	

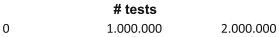


Fig. 3. Clinical services performed in 2013 in the OU of clinical pathology

A retrospective evaluation has shown direct potential savings estimated at about €600.000 per year. In particular, of 1.200.128 exam requests, approximately 218.000 (18 %) were considered inappropriate, about 110.000 (9 %) were uncertain, and near 880.000 (73 %) were deemed to be appropriate. By considering the Italian price list for the specialist outpatient care (of January 2013), and a cost of 1,2€ per clinical service, the overhead due to inappropriate requests in the investigated period was calculated at about €250.000. The details are represented in Fig. 4. Such savings, however, largely depend on the degree of computerization in the healthcare area and on the integration among the involved systems, but they represent a first signal of improvement.

Months	Appropriateness			Total
	Appropriate	Uncertain	Inappropriate	
Sept - Oct	168.983 (*72,80%*)	16.248 (*7,00%*)	46.888 (*20,20%*)	232.120
Nov - Dec	246.424 (*74,20%*)	13.284 (*4,00%*)	72.400 (*21,80%*)	332.108
Jan - Feb	457.848 (*72,00%*)	79.488 (*12,50%*)	98.565 (*15,50%*)	635.900
Total	873.255 (*72,76%*)	109.020 (*9,08%*)	217.852 (*18,15%*)	1.200.128
Estimated cost	€ 1.047.907	€ 130.824	**€ 261.422**	€ 1.440.153

Fig. 4. Economic burden of clinical services performed during September 2014 - February 2015

5 Conclusions

The reduction of unnecessary clinical tests and the enforcement of specific "appropriateness rules" can save considerable resources to public health. A retrospective evaluation performed on real data at the operative unit of clinical pathology of the main hospital of Lecce, in Italy, showed potential direct savings estimated at approximately € 600.000/year. The achievement of clinical appropriateness cannot, however, be reduced to a mere matter of saving money, but it should be inspired by the need to raise awareness among professionals and disseminate knowledge on the proper use of diagnostic prescriptions. The development of an automated system that can efficiently supervise the tests' requests making use of an open and sharable repository may be a first solution to the problem. For these reasons we feel that the development of a proposal based on open source technologies and open data may represent an opportunity for savings resources while enhancing the quality and efficiency of the laboratory analyses.

References

1. Fryer, A.A., Smellie, W.S.: Managing demand for laboratory tests: a laboratory toolkit. J. Clin. Pathol. **66**, 62–72 (2013)
2. Terry, M.: Lab Industry Strategic Outlook Market Trends and Analysis 2007. Washington G-2 Reports (2007)
3. Daniels, M., Schroeder, S.A.: Variation among physicians in the use of laboratory tests: relation to clinical productivity and outcomes. Med. Care **15**, 482–487 (1977)
4. Valenstein, P., Leiken, A., et al.: Tests ordering by multiple physicians increases unnecessary laboratory examinations. Arch. Pathol. Lab. Med. **112**, 238–241 (1988)
5. Bareford, D., Hayling, A.: Innappropriate use of laboratory services: long term combined approach to modify request patterns. BMJ **301**, 1305–1307 (1990)
6. Van Walraven, C., Naylor, C.D.: Do we know what inappropriate laboratory utilization is? A systematic review of laboratory clinical audits. JAMA **280**, 550–558 (1998)
7. Batalden, P.B., Davidoff, F.: What is "quality improvement" and how can it transform healthcare? Qual. Saf. Health Care **16**(1), 2–3 (2007)
8. Rodriguez-Espinosa, J.: Clinical laboratory: use and misuse, management models and health expenditure. Med. Clin. (Barc) **125**, 622–625 (2005)
9. Larsson, A., Palmer, M., Hulte`n, G., Tryding, N.: Large differences in laboratory utilisation between hospitals in Sweden. Clin. Chem. Lab. Med. **38**, 383–389 (2000)

10. Pleban, M., Mussap, M.: ICT, automazione e appropriatezza: le logiche organizzative e le logiche diagnostiche. Riv Med Lab-JLM **5**(2), 92–101 (2004)
11. Bonaldi, A., Vernero, S.: Italy's Slow Medicine: a new paradigm in medicine. Recenti Prog. Med. **106**(2), 85–91 (2015)
12. Prinsloo, E.A.M., Dimpe, M.W., et al.: Doctors' use of laboratory tests in the diagnosis and treatment of patients. S. Afr. J. Epidemiol. Infect. **25**(3), 16–20 (2010)
13. Fryer, A.A., Hanna, F.W.: Managing demand for pathology tests: financial imperative of duty of care? Ann. Clin. Biochem. **46**, 435–437 (2009)
14. Catrou, P.G.: Is that lab test necessary? Am. J. Clin. Pathol. **126**, 335–336 (2006)
15. Kahn, C.E.: An Internet-based ontology editor for medical appropriateness criteria. Comput. Methods Programs Biomed. **56**, 31–36 (1998)
16. Verhulst, S., Noveck, B.S. et al.: The Open Data Era in health and social care: a blueprint for the National Health Service (NHS England) to develop a research and learning programme for the open data in health and social care, May 2014
17. Gupta, A., Forgy, C., et al.: High-speed implementations of rule-based systems. ACM Trans. Comput. Syst. (TOCS) **7**(2), 119–146 (1989)

Machine Learning in Biomedicine

A Logistic Regression Approach for Identifying Hot Spots in Protein Interfaces

Peipei Li and Keun Ho Ryu(✉)

Database/Bioinformatics Laboratory, Chungbuk National University, Cheongju, South Korea
{lipeipei,khryu}@dblab.chungbuk.ac.kr

Abstract. Protein–protein interactions occur when two or more proteins bind together, often to carry out their biological function. A small fraction of interfaces on protein surface found providing major contributions to the binding free energy are referred as hot spots. Identifying hot spots is important for examining the actions and properties occurring around the binding sites. However experimental studies require significant effort; and computational methods still have limitations in prediction performance and feature interpretation.

In this paper we describe a hot spots residues prediction measure which provides a significant improvement over other existing methods. Combining 8 features derived from accessibility, sequence conservation, inter-residue potentials, computational alanine scanning, small-world structure characteristics, phi-psi interaction, and contact number, logistic regression is used to derive a prediction model. To demonstrate its effectiveness, the proposed method is applied to ASEdb. Our prediction model achieves an accuracy of 0.819, F1 score of 0.743. Experimental results show that the additional features can improve the prediction performance. Especially phi-psi has been found to give important effort. We then perform an exhaustive comparison of our method with various machine learning based methods and those previously published prediction models in the literature. Empirical studies show that our method can yield significantly better prediction performance.

Keywords: Protein–protein interactions · Binding sites · Protein hot spots prediction · Logistic regression

1 Introduction

Protein–protein interactions occur when two or more proteins bind together, often to carry out their biological function [1]. Many of the most important molecular processes in the cell such as signal transmission are carried out by large molecular machines that are built from a large number of protein components organized by their protein–protein interactions [2]. A small fraction of interfaces on the protein surface are found providing major contributions to the binding free energy [3]. To identify hot spots is important for examining the actions and properties occurring around the binding sites, and therefore provides important clues to the function of a protein.

Alanine scanning mutagenesis [4], which systematically mutates a target residue into alanine and measures the binding free energy change, is a popular technique used

M.E. Renda et al. (Eds.): ITBAM 2015, LNCS 9267, pp. 37–48, 2015.
DOI: 10.1007/978-3-319-22741-2_4

to examine the energetic importance of a residue in the binding of two proteins. Based on it, Alanine Scanning Engergetics database (ASEdb) [5] is created with experimental hot spots. Another database, binding interface database (BID) which contains binding free energy strengths mining from the primary scientific literature is also well known [6]. Experimental methods are absolutely with high accuracy, however they are time consuming and expensive. Therefore computational methods are required to solve the problem. However it is still a challenging task in bioinformatics area.

In recent years several studies focus on researching different characteristics between hot spots and non-hot spots residues. It is proved that hot spots are clustered at the core of the protein interface surrounding by O-ring residues at its rim [7]. Studies find that hot spots are statistically correlated with structurally conserved residues [8, 9]. Also the relevance of the amino acid physicochemical and biological properties values in the hot spots area on protein surface with the interface residues has been well investigated [10, 11].

Based on the research of the characteristics of hot spots, a number of computational methods have been developed to predict hot spots residues from interface residues. Especially feature based methods achieve relative good predictive results. In [12], an efficient approach namely APIS that uses support vector machine (SVM) to predict hot spot using a wide variety of 62 features from a combination of protein sequence and structure information is developed. F-score method is used as a feature selection method to remove redundant and irrelevant features and improve the prediction performance. Nine individual features based predictor is finally developed to identify hot spots with F1 score of 0.64. HotPoint [13] is a server providing the hot spot prediction results considering criteria: Hot spots are buried, more conserved, packed, and known to be mostly of specific residue types. Based on the benchmark dataset it achieves an accuracy of 0.70. Another automated decision-tree approach combines two knowledge-based models to improve the ability to predict hot spots: K-FADE uses shape specificity features calculated by the Fast Atomic Density Evaluation (FADE) program, and K-CON uses biochemical contact features [14]. Graph representation of homodimeric protein complexes is another approach to predict hot spots. In [15], combining small-world network characteristics and solvent accessibility the predicted highly central residues show a general tendency to be close to the annotated hot spots.

Although computational methods have been well developed and achieve a relative success with good performance, they are still under limitation. First the features used in predicting method are not comprehensive. Second the features previously identified as being correlated with hot spots are still insufficient.

In this paper, we present a statistical model for predicting hot spots residues in protein interaction interfaces. First 8 features extracted from Accessible, sequence conservation, Inter-residue potentials, Computational alanine scanning, Small-world structure characteristics, Phi-psi interaction, and Contact number are used as input. Then logistic regression is used to construct prediction model. To evaluate our model we first do large-scale experiments. Our method yields an accuracy of 0.819, F1 score of 0.743 on ASEdb set. We then perform an exhaustive comparison of our method with various machine learning based methods and than those previously published approaches in the literature. Empirical studies show that our method can yield significantly better prediction accuracy.

2 Material and Method

2.1 Benchmark Dataset

We present results from dataset of experimental hot spots from the Alanine Scanning Energetics database (ASEdb).

ASEdb, which contains interface residues experimentally mutated to alanine, is treated as an experimental data set. It is first published by [5] and now used by researchers widely. After removing residues with sequence identity not more than 35 % we final selected 149 amino acid residues from 14 protein complexes, in which 58 are considered as hot spots whose observed binding free energies are ≥2.0 kcal/mol, 91 are labeled as non-hot spots whose binding free energy is <0.4 kcal/mol.

2.2 Features for Hot Spots Prediction

The feature vector used in logistical regression model consists of a total of 8 features-derived from sequence, structure and we provide a brief description of these features as well as some of the options we considered in this section.

2.2.1 Accessibility

Accessible surface area (ASA) which refers to the surface area of the molecules that are accessible to solvent [16] can improve the accuracy of identifying computational hot spots in protein interface [17]. Especially in [18] authors prove that ASA in complex state discriminates better hot spots from non-hotspots than ASA in monomer state. We use ASA in complex in this work. The values of each residue in the data set are calculated by Naccess [19].

2.2.2 Propensity Scaled Sequence Conservation Score

The importance of sequence conservation features is well studied. Researchers have applied in catalytic radiuses identification in enzymes [20] and hot spots residues prediction [21] in literatures. It was shown that central residues are highly conserved in sequence alignments and non-exposed to the solvent in the protein complex and concluded that these residues either correspond to experimental hot spots or are in contact with experimentally annotated hot spots.

Sequence conservation scores are derived based on multiple sequence alignment (MSA) of homologs gathered from HSSP (Homology-Derived Secondary Structure of Proteins) [22]. A multiply sequence alignment is estimated using Rate4Site program [23] which detects conserved amino-acid sites by computing the relative evolutionary rate for each site. Conservation scores obtained by Rate4Site are scaled between 1 and 9. In addition, amino acids are proved to have varying propensity, so we use conserved residue propensity calculated in [21] to rescale conservation score by the following formulation:

$$pScore_i = score_i \times P_k \tag{1}$$

Where $pScore_i$ refers to Propensity scaled sequence conservation score, $score_i$ is the sequence conservation score calculated by Rate4Site program, and P_k is the propensity of residue type k.

2.2.3 Inter-residue Potentials

Inter-residues potentials play a curial role in many protein folding and binding problems [24]. Knowledge-based solvent mediated inter-residue potentials extracted from protein interfaces are used to calculated pair potentials. In this work, we use the same method as described in [18].

2.2.4 Small-World Structure Characteristics

Protein structures can be modeled as network systems where amino acid residues are nodes and their interactions with each other are the edges. In [25], they compare with random networks and regular networks, and find out that the network of protein structure are characterized by relatively high values of clustering coefficients and small values of characterized path length. In other words the protein structures exhibit characteristics that resemble a small-world network.

In small-world networks, betweenness is calculated as the number of times residue k is included in the shortest path between each pair of residues in the protein, normalized by the total number of pairs. It is proved that highly central residues with high betweenness value show a general tendency to be close to hot spot residues [15]. Betweenness is used as a feature in our prediction model. A neighboring residue is defined as the distance between two in the protein surface which is less than 6 Å.

2.2.5 Phi-Psi Interaction Features

Protein backbone torsion angles (Phi) and (Psi) involve two rotation angles rotating around the Cα-N bond (Phi) and the Cα-C bond (Psi). Due to the planarity of the linked rigid peptide bonds, these two angles can essentially determine the backbone geometry of proteins. We calculate these values according to BioJAVA [26].

2.2.6 Contact Number

Contact number is a simple solvent exposure measure that estimates residue burial in proteins [27, 28]. It has been used in enhancing protein fold recognition [29]. Contact number is generally defined as the number of Ca atoms of other residues within a user-defined sphere around the Ca atom of the residue at hand. The radius of the sphere is typically chosen to be between 8 and 14 Å.

2.3 Logistic Regression Model

Logistic regression is a venerable, but capable probabilistic binary classifier. It is well-understood, mature and comfortable. Logistic regression accuracy is comparable to new-fangled state-of-the-art SVMs. So we choose it to construct hot spots prediction model.

Given X be a protein amino acid residues dataset with binary output y (whether the residue is hot spots or not), let x_i be the d-dimensional vector of residue-specific features,

y_i be the hot spots label of residue i. We model the conditional distribution of the random variable y by a logistic regression, thus our regression model is

$$P(y = 1 \mid x; \theta) = \frac{1}{1 + \exp^{-\theta^T X}} \tag{2}$$

The parameter θ to be calculated is a vector.
Assume

$$P(y = 1 \mid x; \theta) = h_\theta(x) \tag{3}$$

$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x) \tag{4}$$

We can derive likelihood and log likelihood of the data X, y under the LR model with parameters θ as

$$L(\theta) = \prod_{i=1}^{m} (h_\theta(x_i))^{y_i} (1 - h_\theta(x_i))^{1-y_i} \tag{5}$$

$$l(\theta) = \sum_{i=1}^{m} y_i \log h_\theta(x_i) + (1 - y_i) \log(1 - h_\theta(x_i)) \tag{6}$$

The logistic regression model is executed with SPSS [30] on data set from ASEdb.

2.4 Evaluation Measure

Precision, recall and accuracy are three widely used metrics employed in classification. And in additional F1 measure as a weighted average of the precision and recall is also used for assessment of protein-protein interface hot spot prediction methods.

Let TP, FP, TN, and FN denote the numbers of true positive (a predicted residue included in the benchmark dataset), false positive (a predicted residue not listed in the benchmark dataset), true negative (a hot spot residue in the benchmark dataset which has been missed by prediction method) and false negative (a non-hot spot residue in the benchmark dataset which has been correctly predicted) respectively. A formal definition of these metrics is given below.

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$A = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{10}$$

3 Experiments and Results

3.1 Feature Analysis

As known redundant and irrelevant features can reduce classification accuracy, we first do simple analysis on data set ASEdb. As in previous studies, compASA, pScore, potential, and robetta features are widely used in hot spots prediction and they are well studied. In this paper, we mainly focus on the importance of newly introduced features including betweenness, phi, psi and CN and their improvement on prediction performance. We will then perform various features combination to determine whether the introduced features can improve prediction performance or not.

In order to give a view of all selected features, data distribution plot is shown in Fig. 1. Among all the features, compASA and robetta show more evident difference between hot spots residues and non-hot spots residues. Feature phi and psi show less difference. This indicates that phi and phi may not be a good discriminating factor by itself.

In additional correlation analysis for each feature is explored (Table 1). Except compASA and potentials, compASA and CN have correlation coefficients between 0.55 and 0.5; all other correlation coefficients are less than 0.5. It means that all features are less relevant.

Variables in the logistic regression equation are listed in Table 2. Features compASA, Robetta and Phi have signification scores less than 0.05, so the three features gives more contribution than other features.

Table 3 shows the prediction performance with basic features (compASA, pScore, potential, and robetta) and additional features (betweenness, phi, psi and CN). In the same data set, all the evaluation scores have been improved with 0.086, 0.074, 0.054 and 0.083 for precision, recall, accuracy and F1-score respectively. It proves that the additional features can improve the prediction performance.

3.2 Comparison with Other Machine Learning Classifiers

Using the same feature sets we do comparison with other classic machine learning methods, as decision tree, Naïve Bayes, and SVM. Prediction results using the same dataset are shown in Table 4. Except SVM classifier got a highest precision of 0.722, our method obtains a highest recall of 0.830, accuracy of 0.819 and F1 score of 0.743. Other classifiers do not show good prediction results.

3.3 Comparison with Existing Prediction Approaches

Performance is also compared with other existing hot spot prediction approaches. Robetta is designed residues with experimental $\Delta\Delta G$ larger than 1.0 kcal/mol are

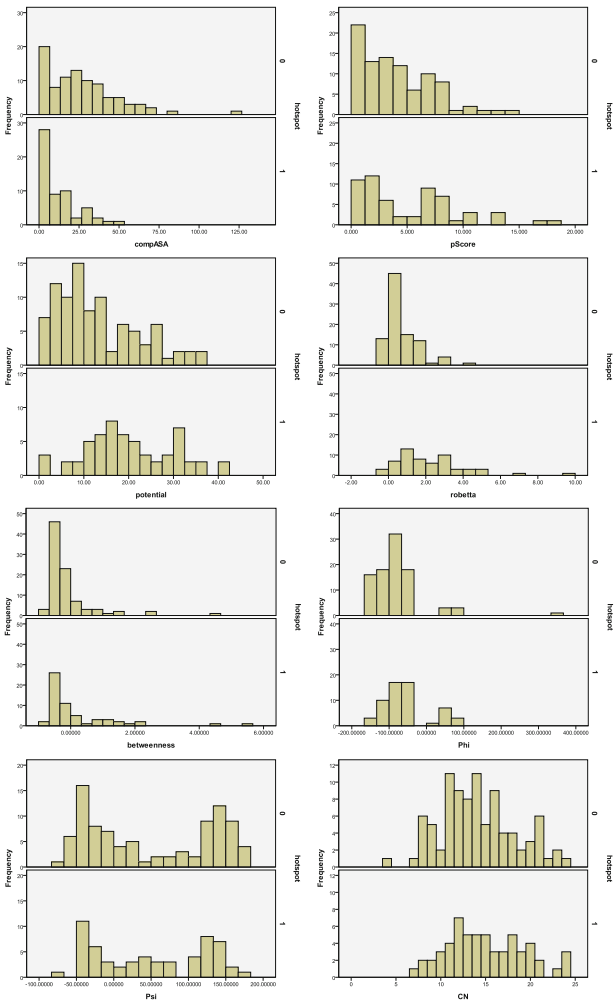


Fig. 1. Data distribution for compASA, pScore, potential, robetta, betweenness, phi, psi, CN and class hotspots. Up bars show hot spot residues, and down bars show non-hot spots residues.

labeled as computational hot spots. Results of KFC and Hotpoint are extracted from each method. As shown in Table 5, comparing with other methods, our method achieves a highest precision of 0.672, recall of 0.830, accuracy of 0.819 and F1 score of 0. 743 on ASEdb data set. Robetta, KFC and Hotpoint have no excellent performance.

3.4 Case Study

In additional to these large-scale experiments, we present one detailed case study in this section.

Table 1. Correlation analysis between features

	compASA	pScore	potentials	Robetta	betweenness	Phi	Psi	CN
compASA	1	−0.262	**−0.535**	−0.388	−0.242	0.267	−0.014	−0.516
pScore		1	0.234	0.359	0.367	0.092	−0.211	0.192
potentials			1	0.458	0.155	−0.005	0.104	0.318
Robetta				1	0.187	0.123	−0.027	0.047
betweenness					1	−0.083	−0.033	0.323
Phi						1	−0.215	−0.135
Psi							1	0.071
CN								1

Table 2. Variables in the equation

	B	Sig.	Exp(B)
compASA	−0.049	**0.007**	0.952
pScore	−0.071	0.288	0.931
potentials	0.015	0.562	1.015
Robetta	0.805	**0.000**	2.236
betweenness	0.317	0.232	1.373
Phi	0.012	**0.003**	1.012
Psi	0.002	0.396	1.002
CN	−0.030	0.622	0.970
Constant	0.715	0.584	2.044

Table 3. Comparison of prediction performance with basic and additional features

Methods	P	R	A	F1
Basic	0.586	0.756	0.765	0.660
Full	**0.672**	**0.830**	**0.819**	**0.743**

Table 4. Comparison of prediction performance of machine learning based methods

Classifiers	P	R	A	F1
Decision tree	0.559	0.655	0.664	0.603
Naïve Bayes	0.679	0.621	0.738	0.649
SVM	**0.722**	0.448	0.718	0.553
Our method	0.672	**0.830**	**0.819**	**0.743**

Table 5. Comparison of prediction performance of various existing methods

Methods	P	R	A	F1
Robetta	0.627	0.724	0.725	0.672
KFC	0.654	0.586	0.718	0.618
Hotpoint	0.638	0.517	0.698	0.571
Our method	**0.672**	**0.830**	**0.819**	**0.743**

Barstar (pdbID: 1BRS, chain D) is a small protein synthesized by the bacterium Bacillus amyloliquefaciens. Its function is to inhibit the ribonuclease activity of its binding partner barnase (pdbID: 1BRS, chain A), with which it forms an extraordinarily tightly bound complex within the cell until barnase is secreted [31]. As shown in Fig. 2, experimental defined hot spots in interaction surface 1BRSAD are Y29, D35, D39 in chain D, and K27, N58, R59, R83, R87, H102 in chain A, and E60 in chain A is found to be non-hot spots in ASEdb. Prediction results are listed in Table 6. Our methods correctly predicted 8 of the nine hot spots and all of the non-hot spots. Only one hot spot is incorrectly predicted as non-hot spots. For Robetta, 6 of the nine hot spots and all of the non-hot spots are correctly predicted. For KFC only 5 of the nine hot spots are corrected identified and all the other residues are wrongly classified.

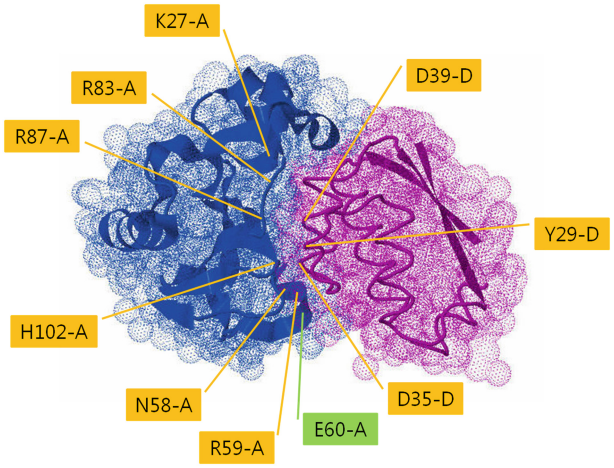


Fig. 2. The visualization of experimental hot spots and non-hot spots for chain A (left) and chain D (right) of protein complex IBRS. Hot spot residues and non-hot spots residues are marked. The figure is plotted using Jmol.

Table 6. Prediction results on protein complex 1BRS

Chain	Residue	hotspot	Our method	Robetta	KFC	Hotpoint
A	K27	Y	Y	N	N	N
A	N58	Y	N	N	N	N
A	R59	Y	Y	Y	N	N
A	E60	N	N	N	Y	N
A	R83	Y	Y	Y	Y	Y
A	R87	Y	Y	Y	Y	Y
A	H102	Y	Y	Y	N	Y
D	Y29	Y	Y	Y	Y	N
D	D35	Y	Y	N	Y	Y
D	D39	Y	Y	Y	Y	Y

HotPoint also correctly predicts 5 of the nine hot spots and the one non-hot spots. Especially K27 in chain A is only correctly predicted by our method. So our method has relative better performance in this case.

4 Conclusion and Future Research

In this paper, we presented a statistical model using logistic regression for predicting hot spots residues. Accessible surface area of protein complex, propensity scaled sequence conservation, inter-residue potentials, betweenness from small-world structure characteristics, phi, psi and contact number were collected from protein sequence, structure information to be used as input features. Computational alanine scanning value was also added to improve the prediction performance.

To demonstrate its effectiveness, the proposed method was applied to the Alanine Scanning Energetics database (ASEdb) benchmark datasets. In large-scale experiments, our prediction model achieved accuracy of 0.819, F1 score of 0.743 on ASEdb data set. Experimental results showed that the additional features can improve the prediction performance. Especially phi-psi has been found to give important effort. We then performed an exhaustive comparison of our method with various machine learning based methods (Decision tree, Naïve Bayes, SVM) and those previously published prediction models (Robetta, KFC, Hotpoint) in the literature. Empirical studies showed that our method can yield significantly better prediction accuracy in current benchmark datasets. Finally we provided one detailed case study of protein complex 1BRS and explained the hot spots prediction results.

Though we improved the prediction accuracy in this work, there is still room remained. In the future work, we will focus on exploring more efficient features for hot spots prediction. And the data set used for experiment widely are small and old, we expect researchers to give a new hot spots data sets for future research.

Acknowledgments. This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2013R1A2A2A01068923) and by the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1002)

References

1. Li, P., Heo, L., Li, M., Ryu, K.H.: Protein function prediction using frequent patterns in protein-protein interaction networks. FSDK **3**, 1664–1668 (2011)
2. Jones, S., Thornton, J.M.: Principles of protein-protein interactions. Proc. Natl. Acad. Sci. **93** (1), 13–20 (1996)
3. Clackson, T., Wells, J.A.: A hot spot of binding energy in a hormone-receptor interface. Science **267**(5196), 383–386 (1995)
4. Morrison, K.L., Weiss, G.A.: Combinatorial alanine-scanning. Curr. Opin. Chem. Biol. **5**(3), 302–307 (2001)

5. Thorn, K.S., Bogan, A.A.: ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics **17**(3), 284–285 (2001)
6. Fischer, T.B., Arunachalam, K.V., Bailey, D., Mangual, V., Bakhru, S., et al.: The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics **19**(11), 1453–1454 (2003)
7. Bogan, A.A., Thorn, K.S.: Anatomy of hot spots in protein interfaces. J. Mol. Biol. **280**(1), 1–9 (1998)
8. Ma, B., Elkayam, T., Wolfson, H., Nussinov, R.: Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc. Natl. Acad. Sci. **100**(10), 5772–5777 (2003)
9. Keskin, O., Ma, B., Nussinov, R.: Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. J. Mol. Biol. **345** (5), 1281–1294 (2005)
10. Chen, X., Jeong, J.: Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics **25**(5), 585–591 (2009)
11. Li, N., Sun, Z., Jiang, F.: Prediction of protein-protein binding site by using core interface residue and support vector machine. BMC Bioinform. **9**(1), 553 (2008)
12. Xia, J.F., Zhao, X.M., Song, J., Huang, D.S.: APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinform. **11**, 174 (2010)
13. Tuncbag, N., Keskin, O., Gursoy, A.: HotPoint: hot spot prediction server for protein interfaces. Nucleic Acids Res. **38**, W402–W406 (2010)
14. Darnell, S.J., Page, D., Mitchell, J.C.: An automated decision-tree approach to predicting protein interaction hot spots. Proteins **68**, 813–823 (2007)
15. Del, Sol: A. and O'Meara, P.: Small-world network approach to identify key residues in protein-protein interaction. Proteins **58**(3), 672–682 (2005)
16. Shrake, A., Rupley, J.A.: Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J. Mol. Biol. **79**, 351–371 (1973)
17. Rost, B., Sander, C.: Conservation and prediction of solvent accessibility in protein families. Proteins **20**, 216–226 (1994)
18. Tuncbag, N., Gursoy, A., Keskin, O.: Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. Bioinformatics **25**(12), 1513–1520 (2009)
19. Hubbard, S.J., Thornton, J.M.: NACCESS. Department of Biochemistry and Molecular Biology, University College, London (1993)
20. Sankararaman, S., Sha, F., Kirsch, J.F., Jordan, M.I., Sjölander, K.: Active site prediction using evolutionary and structural information. Bioinformatics **26**(5), 617–624 (2010)
21. Guney, E., Tuncbag, N., Keskin, O., Gursoy, A.: HotSprint: database of computational hot spots in protein interfaces. Nucleic Acids Res. **36**, D662–D666 (2008)
22. Dodge, C., Schneider, R., Sander, C.: The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res. **26**(1), 313–315 (1998)
23. Mayrose, I., Graur, D., Ben-Tal, N., Pupko, T.: Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol. Biol. Evol. **21**(9), 1781–1791 (2004)
24. Jernigan, R.L., Bahar, I.: Structure-derived potentials and protein simulations. Curr. Opin. Struct. Biol. **6**(2), 195–209 (1996)
25. Greene, L.H., Higman, V.A.: Uncovering network systems within protein structures. J. Mol. Biol. **334**(4), 781–791 (2003)
26. Holland, R.C., Down, T.A., Pocock, M., Prlić, A., Huen, D., et al.: BioJava: an open-source framework for bioinformatics. Bioinformatics **24**(18), 2096–2097 (2008)

27. Pollastri, G., Baldi, P., Fariselli, P., Casadio, R.: Prediction of coordination number and relative solvent accessibility in proteins. Proteins **47**, 142–153 (2002)
28. Li, P., Pok, G., Jung, K.S., Shon, H.S., Ryu, K.H.: QSE: A new solvent exposure measure for the analysis of protein structure. Proteomics **11**(19), 3793–3801 (2011)
29. Karchin, R., Cline, M., Karplus, K.: Evaluation of local structure alphabets based on residue burial. Proteins. **55**, 508–518 (2004)
30. Levesque, R.: SPSS Programming and Data Management: A Guide for SPSS and SAS Users, 4th edn. SPSS Inc., Chicago Ill (2007)
31. Hartley, R.W.: Barnase and barstar: two small proteins to fold and fit together. Trends Biochem. Sci. **14**(11), 450–454 (1989)

The Discovery of Prognosis Factors Using Association Rule Mining in Acute Myocardial Infarction with ST-Segment Elevation

Kwang Sun Ryu[1], Hyun Woo Park[1], Soo Ho Park[1], Ibrahim M. Ishag[1], Jang Hwang Bae[2], and Keun Ho Ryu[1(✉)]

[1] Database / Bioinformatics Laboratory, Chungbuk National University, Cheongju-si, South Korea
{ksryu, hwpark, soohopark, Ibrahim, khryu}@dblab.chungbuk.ac.kr
[2] Division of Cardiology, Department of Internal Medicine, Chungbuk National University, Cheongju city, Chungbuk, Korea
Jangwhan.bae69@gamil.com

Abstract. Association rule mining has been applied actively in order to discover the hidden factors in acute myocardial infarction. There has been minimal research regarding the prognosis factor of acute myocardial infarction, and several previous studies has some limitations which are generation of incorrect population and potential data bias. Thus, we suggest the generation of prognosis factor based on association rule mining for acute myocardial infarction with ST-segment elevation. In our experiments, we obtain high interestingness factor based on Korean acute myocardial infarction registry which is corrected by 51 participating hospitals since 2005. The interestingness of the factor is evaluated by confidence. It is expected to contribute to prognosis management by high reliability factor.

Keywords: Association rule mining · Acute myocardial infarction · Prognosis factor

1 Introduction

Acute Myocardial infarction (AMI) has increased due to various habitual causes such as wrong eating habit, diabetes, and as so on. It is also known as a heart attack, which is the irreversible necrosis of myocardium secondary to prolonged ischemic heart disease. This usually results from an imbalance in oxygen supply and demand, which is most often caused by plaque rupture with thrombus formation in a coronary artery, resulting in an acute reduction of blood supply to a portion of the myocardium [1].

The AMI is studied by association rule mining to analyze the hidden factor. M Karaolis et al. developed a data mining system using association analysis based on the apriori algorithm for the assessment of heart event related risk factors [2]. In their study, they have used experimental data sets without control group, which might

M.E. Renda et al. (Eds.): ITBAM 2015, LNCS 9267, pp. 49–55, 2015.
DOI: 10.1007/978-3-319-22741-2_5

generate incorrect rules risking the reliability of their mining system. DG Lee et al. discovered meaningful rules using association rule mining based on Korean acute myocardial infarction registry (KAMIR). The study focused on young adult patients with AMI [3]. Previous studies have applied the association rule mining without the verification of population. In particular, these studies have overlooked basic concepts such as STEMI, NSTEMI, PCI, symptom to balloon and door to balloon time, which might generate less reliable rules. Therefore, we suggest the generation approach of association rule mining for acute myocardial infarction in order to extract the reliable factors that affects patient prognosis.

The rest of the paper is organized as follows. Section 2 describes the material and method, Sect. 3 the experimental result, and Sect. 4 the conclusion.

2 Material and Method

This section presents association rule mining and basic concepts. The outcomes of the patient's prognosis differ according to AMI categorizations such as STEMI, NSTEMI, underwent PCI, without PCI, Door to balloon and Symptoms to balloon time.

2.1 Basic Concepts

STEMI and NSTEMI. Initial treatment of Acute Myocardial Infarction (AMI) is different by diagnose of ST segment elevation Acute Myocardial Infarction (STEMI) and non ST segment elevation Acute Myocardial Infarction (NSTEMI). The STEMI should administer thrombolytic agent within 30 min after hospital arrival or operate Percutaneous Coronary Intervention (PCI) within 90 min after hospital arrival [4]. On the other hand, NSTEMI should operate coronary angiography within 48 h after indicating symptoms on the basis of risk factor [5]. Consequently, in this work, patients were classified into two groups; Namely, STEMI and NSTEMI.

PCI. Percutaneous coronary intervention (PCI) is performed to open blocked coronary arteries caused by coronary artery disease and to restore arterial blood flow to the heart tissue without coronary artery bypass grafting (CABG) [6]. In our work, patients whose records do not indicate that they have underwent PCI in initial therapy were excluded.

Symptom to Balloon and Door to Balloon Time. American College of Cardiology (ACC)/American (AHA) and European Society of Cardiology guideline describe percutaneous coronary intervention (PCI) as the preferred reperfusion strategy in ST-segment elevation, if first medical contact to balloon time or door to balloon time is < 120 min and symptom to balloon time is < 4 h, patients mortality is decreased [7, 8]. We exclude patients with over time in terms of symptom to balloon time and door to balloon time which occur bias of outcomes.

2.2 Association Rule Mining

An association rule is an implication expression of the form $(X \rightarrow Y)$, where X and Y are disjoint item-sets, i.e.,$(X \cap Y = \varphi)$ Support determine how often a rule appears in transactions, while confidence determines how frequently itemset in Y appears in transactions that contain X. Association rule mining finds all rules that satisfy user-defined minimum support threshold and minimum confidence threshold. Rules which are generated by the approach of association rule mining include relationship of all attributes. We apply apriori algorithm in association rule mining. The frequent item-sets are extended one item at a time. It is main idea is to generate k-th candidate item-sets from the (k-1)-th frequent item-sets and to find the k-th frequent item-sets from the k-th candidate item-sets. The algorithms terminates when frequent item-set can't be extended any more [9]. The weka's implementation of this algorithm has been run in our study [10].

3 Experiments and Results

This section presents several experiments based on the approach of association rule mining. The first experiment shows the data coding and generation of reliable population in order to apply to association rule mining.

The other experiments show the extracted interestedness rules which affect the prognosis of patients (Fig. 1).

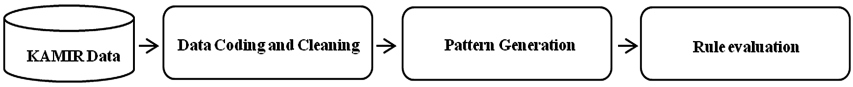


Fig. 1. The experimental process

3.1 Medical Database and Data Coding

Korean acute myocardial infarction registry (KAMIR) is a Korean, prospective, open, observation, multicenter, on-line registry of AMI with support from the Korean Society of Cardiology since November 2005. The 50 participating hospitals are capable of primary percutaneous coronary intervention (PCI). It has 14,855 AMI patients and including 141 risk factors. We select 17 factors which are age, gender, Killp class, history of ischemic heart disease, history of hypertension, history of diabetes, history of dylipidemia, history of smoking, family history of heart disease, comorbidities, angiographic findings, target lesion, lesion type, pre thrombosis in myocardial infarction (TIMI), left ventricle (LV) ejection fraction, Post TIMI, and stent type, which have already known or diagnosed in the past (Table 1).

Table 1. Data coding

Risk factor	Abbr	Code1	Code2	Code3
Age	A	45 > L	45 > M>65	H > 65
Gender	G	M:MALE	FM:FEMALE	
Killip class	K	Class 1 = I	Class 2 = II	Class 3 < III
Ischemic heart disease	IHD	YES = Y	NO = N	
Hypertension	H	YES = Y	NO = N	
Diabetes	D	YES = Y	NO = N	
Dylipidemia	DYL	YES = Y	NO = N	
Smoking	S	YES = Y	NO = N	
Family history	FH	YES = Y	NO = N	
Co-morbidities	C	YES = Y	NO = N	
Angiographic findings	AF	ONE = I	TWO = II	THREE = III
Target lesion	TL	RCA = R	LCA = L	LCX = LX
Lesion type	LT	TYPE A = A	TYPE B = B	THPE C = C
Pre TIMI	PRT	TIMI I = I	TIME II = II	TIMI III < III
Left ventricle ejection	LV	50 > L	50 < H	
POST TIMI	POT	TIMI I = I	TIME II = II	TIMI III < III
Stent type	ST	BARE = B	DRUG = D	

3.2 Data Cleaning

We eliminated considerable number of patients with following exclusion criteria to avoid data bias. Patients with more than 12 h delay from chest pain symptoms to revascularization (n = 2500), more than 2 h delay from emergency room appearance revascularization (n = 979), in-hospital death (n = 345), and underwent thrombosis without PCI (n = 39). Afterwards, random sampling was applied to overcome the class imbalance problem and which resulted in 525 patients that resemble the target population (Fig. 2).

3.3 Risk Factor for Prognosis

We defined minimum support threshold as 0.1 and minimum confidence threshold as 0.6. Major adverse cardiac event (MACE) consists of any cause death, all revisualization, repeated percutaneous coronary intervention, and coronary artert bypass graft surgery, which defined the key pattern. Therefore, the one thousand patterns associated with MACE are generated. The patterns were verified by subject matter experts (i.e. physicians) and confidence, as a result we have obtained interesting rules that include; 1: patient have high age without diabetes, dyslipidemia, and family history, 2: with high age without ischemic heart disease, diabetes, and family history, 3: with high age without ischemic heart disease, diabetes, dyslipidemia, and family history, 4: with male and killp class I without ischemic heart disease, post TIMI III, and dyslipidemia, 5: with age and post TIMI III without family history, 6: with high age without

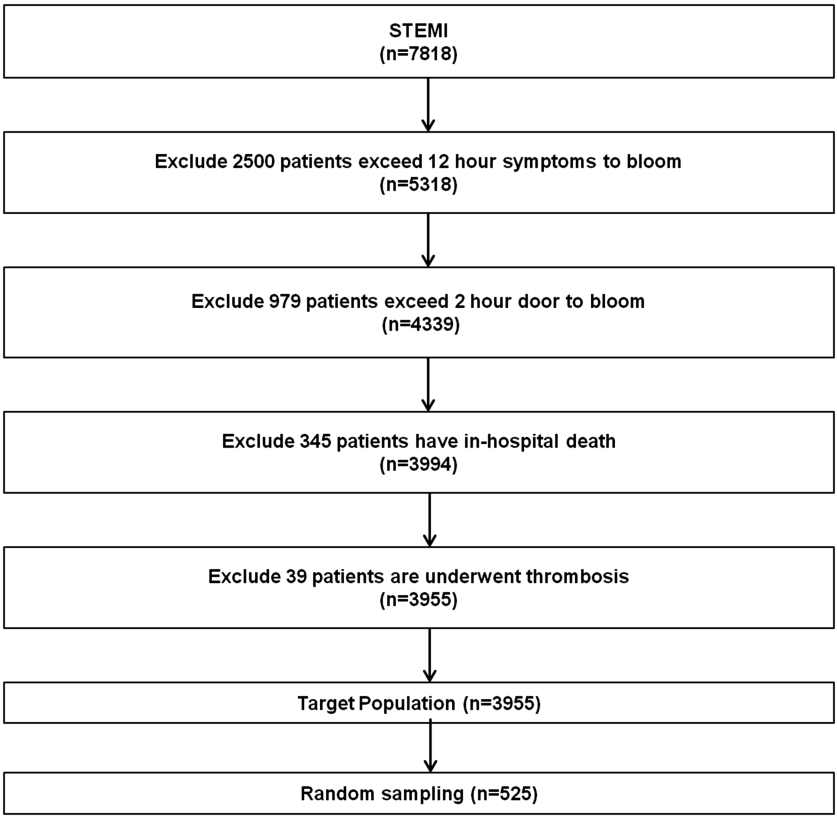


Fig. 2. Generation of target population

dyslipidemia, post TIMI III, and family history, 7: with killip class I, smoking, and post TIMI III, and male. These rule should affect adverse patients prognosis who underwent successful PCI with STEMI. The risk factor rules are given in Table 2.

Table 2. Risk factor rules

RULES	CONFIDENCE
RULE 1 : MACE = Y, A = H, D = N, DYL = N, FH = N	0.97
RULE 2 : MACE = Y, A = H, IHD = N, D = N, FH = N	0.97
RULE 3 : MACE = Y, A = H, IHD = N, D = N, DYL = N FH = N	0.97
RULE 4 : MACE = Y, G = M, K = I, IHD = N, POT = III, DYL = N	0.96
RULE 5 : MACE = Y, A = H, POT = III, FH = N	0.96
RULE 6 : MACE = Y, A = H, DYL = N, POT = III, FH = N	0.95
RULE 7 : MACE = Y, K = I, S = Y, POT = III, G = M	0.95

4 Conclusion and Future Research

In our study, we have introduced the right way to preprocess Acute myocardial infarction data in order to analyze prognosis factors. Furthermore, we have discovered medically relevant factors using association rule mining. The Rules 1, 2, and 3 showed that advanced age is higher major acute cardiac event (MACE) ratio without disease history. In other hand, diabetes, dylipidemia, ischemic heart disease, and family history are well known risk factors, which are not significant in AMI prognosis. Rule 4 shows Post TIMI with male and killip class one have higher major cardiac adverse events without disease history. In interpretation rules 5 and 6, Post TIMI with advanced age should affect to adverse prognosis without disease history. Rule 7 is killip class one, smoking, Post TIMI III, and male, which shows higher adverse prognosis.

Conclusively, Post TIMI III, advanced age, and smoking show significant risk factors in AMI prognosis. Especially, Post TIMI III with advanced age has shown highest relationship with adverse prognosis, whereas well known risk factors such as diabetes, dylipidemia, family history of heart disease, and ischemic heart disease have shown no significance MACE. This study is conducted without clear evaluation for the rules. Therefore, upcoming studies will consider other statistical measures to verify the discovered rules. Furthermore, our data have the limitation of generalization to define MACE by random sampling. To overcome those limitations, large scaled prospective randomized clinical studies are warranted.

Acknowledgments. This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2013R1A2A2A01068923) and by the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1002).

References

1. The Textbook of Cardiovascular medicine: The Korea Society Of Circulation (2004)
2. Karaolis, M., Moutris, J.A., Papaconstantinou, L., Pattichis, C.S.: Association rule analysis for the assessment of the risk of coronary heart events. In: 31th Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA (2009)
3. Lee, D.G., Ryu, K.S., Bashir, M., Bae, J.W., Ryu, K.H.: Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. J. Med. Syst. **37**(2), 9896 (2013)
4. Antman, E.M., Hand, M., Armstrong, P.W., Bates, E.R.: 2007 Focused up date of the ACC/AHA 2004 guidelines for the management of patients with ST-elevation myocardial infarction. In: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, Circulation, pp. 296–329 (2007)
5. Anderson, J.L., Antman, E.M., Adams, C.D., Bridges, C.R.: ACC/AHA 2007 guidelines for the management of patients with unstable angina/non ST-elevation myocardial infarction: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, Circulation, pp. 148–304 (2007)
6. Johns Hopkins Medicine. www.hopkinsmedicine.org/healthlibrary

7. Kalla, K., Christ, G., Karnik, R., Malzer, R., Norman, G., Prachar, H., Schreiber, W., Unger, G., Glogar, H.D., Kaff, A., Laggner, A.N., Maurer, G., Mlczoch, J., Slany, J., Weber H.S., Huber, K.: Implementation of Guidelines Improves the Standard of Care: The Viennese Registry on Reperfusion Strategies in ST-Elevation Myocardial Infarction (Vienna STEMI Registry), Circulation, vol. 113, pp. 2398–2405 (2006)
8. De Luca, G., Suryapranata, H., Ziklstra, F., van't Hof, A.W., Hoorntje, J.C., Gosselink, A.T., Dambrink, J.H.: Symtom-onset-to-balloon time and mortality in patients with acute myocardial infarction treated by primary angioplasty. J. Am. Coll. Cardiol. **42**, 991–997 (2003)
9. Tan, PN., Michael, S., Vioin, K.: Introduction to Data Mining (2006)
10. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Moragan Kaufmann Publisher, San Francisco (2005)

Data Mining Techniques in Health Informatics: A Case Study from Breast Cancer Research

Jing Lu[1(✉)], Alan Hales[1,2], David Rew[2], Malcolm Keech[3], Christian Fröhlingsdorf[1], Alex Mills-Mullett[1], and Christian Wette[1]

[1] Southampton Solent University, Southampton SO14 0YN, UK
Jing.Lu@solent.ac.uk
[2] University Hospital Southampton, Southampton SO16 6YD, UK
aahales@btinternet.com, D.Rew@soton.ac.uk
[3] University of Bedfordshire, Park Square, Luton LU1 3JU, UK
Malcolm.Keech@beds.ac.uk

Abstract. This paper presents a case study of using data mining techniques in the analysis of diagnosis and treatment events related to Breast Cancer disease. Data from over 16,000 patients has been pre-processed and several data mining techniques have been implemented by using Weka (Waikato Environment for Knowledge Analysis). In particular, Generalized Sequential Patterns mining has been used to discover frequent patterns from disease event sequence profiles based on groups of living and deceased patients. Furthermore, five models have been evaluated in Classification with the objective to classify the patients based on selected attributes. This research showcases the data mining process and techniques to transform large amounts of patient data into useful information and potentially valuable patterns to help understand cancer outcomes.

Keywords: Health informatics · Database technology · Clinical data environment · Electronic patient records · Breast cancer datasets · Data mining techniques · Knowledge discovery

1 Introduction

Within the healthcare domain there has been a vast amount of complex data generated and developed over the years through electronic patient records, disease diagnoses, hospital resources and medical devices. This data is itself a key resource and has the potential to play a vital role in enabling support for decision making by processing and analysing information for knowledge extraction. The growing amount of data exceeds the ability of traditional methods for data analysis and this has led to Knowledge Discovery in Databases (KDD): the process of automatically searching large volumes of data for interesting patterns, useful information and knowledge [4]. Data mining is the main step in KDD and it brings a set of techniques and methods that can be applied to this processed data to discover hidden patterns. For example, it can provide healthcare professionals with the ability to analyse patient records and disease treatment over time,

M.E. Renda et al. (Eds.): ITBAM 2015, LNCS 9267, pp. 56–70, 2015.
DOI: 10.1007/978-3-319-22741-2_6

which in turn can help to improve the quality of life for those facing terminal illnesses, such as breast cancer.

Within the University Hospital Southampton (UHS) clinical data environment, the Southampton Breast Cancer Data System (SBCDS) has been developed as a "proof of concept" system. In 2010, a consultant surgeon from UHS took the concept for a clinical data interface to the software specialist. There are already some valuable and complex analyses that have been developed within SBCDS and there is potential for further growth in functionality and capability of the system. This has motivated the initial stage of a collaborative research project between UHS and SSU (Southampton Solent University) with the following objectives: enhancement of the SBCDS user interface; expansion of its data mining capability; and exploitation of large-scale patient databases.

The work here will explore the application of different data mining techniques in breast cancer data analysis: from patient data collection and cleaning to the choice of data mining method; and from patterns discovery to modelling and evaluation. The remainder of this paper proceeds as follows: some related work is presented in Sect. 2 to highlight KDD and data mining in the context of health informatics. Section 3 describes the real-world case study comprising the UHS clinical data environment, electronic patient records and their breast cancer data system. Anonymised datasets have been extracted from SBCDS and pre-processing is considered in Sect. 4. Experimental results are compared and evaluated in Sect. 5, which showcases the potential of various mining approaches for generating new and interesting results. The paper draws to a close with concluding remarks indicating future directions.

2 Background and Related Work

This section will start with a description of data warehousing and KDD in health informatics then describe some data processing and mining methods used in breast cancer research.

2.1 KDD in Health Informatics

Health datasets come from various sources (e.g. clinical data, administrative data and financial data) and health information systems are generally optimised for high speed continuous updating of individual patient data and patient queries in small transactions. Using data warehousing can integrate data from multiple operational systems to provide population-based views of health information. Stolba et al. showed that a clinical data warehouse can facilitate strategic decision making, using a clinical evidence-based process for the generation of treatment rules [20]. Data originating from different medical sources was extracted, transformed and prepared for loading into the existing data warehouse structure. Li et al. introduced data mining technology into traditional clinical decision support systems to use the data in the hospital information system for knowledge mining activities [11]. The particular use of different data mining algorithms was also presented in the analysis and characterisation of symptoms and risk factors related with Chronic Obstructive Pulmonary Disease. Clustering was used to identify groups of

individuals presenting similar risk profiles and symptoms. Furthermore, association rules were used to identify correlations among the risk factors and the symptoms.

Data mining is the essential part of KDD – the overall process of converting raw data into useful information and derived knowledge – and could be particularly useful in healthcare and personalised medicine [8]. For instance, based on a patient's profile, history, physical examination and diagnosis, and utilising previous treatment patterns, new treatment plans can be effectively suggested [2]. Clinical data mining is an active interdisciplinary area of research that can be considered as arising from applying artificial intelligence concepts in medicine and healthcare. Several reviews and surveys have been reported to address the impact of data mining techniques in health applications [6]. Laxminarayan et al. propose a modified association rule mining technique to extract patterns from sequence-valued attributes such as sleep-related data by aggregating sets of events [9]. Razavi et al. discuss a decision tree model to predict recurrence of breast cancer, e.g. identifying high-risk patients in order to provide them with specialised treatment [17]. Sequential patterns mining has been applied to a General Practice database to find rules involving patients' age, gender and medical history [18].

2.2 Mining Approaches for Breast Cancer Data

Data mining methods have been applied in breast cancer studies with different objectives and data sources. At the early stage, Artificial Neural Networks (ANNs) were shown to be a powerful tool for analysing datasets where there are complex non-linear interactions between the input data and the patterns to be predicted. Burke et al. [1] compared the 5-year and 10-year predictive accuracy of various statistical models for breast cancer survival, including the pathological TNM staging model (T: size of the original/primary tumour; N: nearby/regional lymph nodes that are involved; M: distant metastasis), principal component analysis, classification and regression trees, and logistic regression.

Jerez-Aragones et al. presented a decision support tool for the prognosis of breast cancer relapse that combined a novel algorithm TDIDT (Top-Down Induction of Decision Trees) for selecting the most relevant factors for prognosis of breast cancer with a system composed of different neural network topologies [7]. They showed that the proposed system is a useful tool for clinicians to search through large datasets, seeking subtle patterns in prognostic factors, and that it may assist in the selection of appropriate adjuvant treatments for the individual patient. Delen et al. developed several prediction models for breast cancer survival – e.g. using ANN, decision trees and logistic regression – based on a large dataset drawn from the SEER (Surveillance, Epidemiology and End Results) cancer incidence and survival database in the USA [3]. They used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. Their approach to breast cancer survivability research provided insight into the relative predictive ability of different data mining methods.

The identification of breast cancer patients for whom chemotherapy could prolong survival has been treated as a data mining problem as well. Lee et al. has achieved this by clustering 253 breast cancer patients into three prognostic groups: Good, Poor and Intermediate [10]. Each of the three groups has a significantly distinct Kaplan-Meier

survival curve. Of particular significance is the Intermediate group, where patients with chemotherapy do better than those without chemotherapy in the same group. They also prescribe a procedure that utilises three non-linear smooth support vector machines for classifying breast cancer patients. These results suggested that the patients in the Good group should not receive chemotherapy while those in the Intermediate group should receive chemotherapy based on their survival curve analysis. Martin et al. examined factors related to the type of surgical treatment for breast cancer using a classification tree approach [15]. Data from the Western Australian Cancer Registry on women diagnosed with breast cancer from 1990 to 2000 was extracted and treatment preferences were predicted from covariates using both classification trees and logistic regression. They concluded that classification trees perform as well as logistic regression for predicting patient choice, but are much easier to interpret for clinical use.

The above related work shows that there have been extensive studies of using classification technologies in breast cancer and this paper will demonstrate the capability of classifiers for SBCDS. In addition, sequential patterns mining has been explored to show the applicability of an alternative data mining technique.

3 University Hospital Case Study

The real-world case study from UHS is introduced in this section in the context of its clinical data environment and electronic patient records, with the ultimate focus on the breast cancer data system.

3.1 Clinical Data Environment and EPR

The University Hospital Southampton has been at the forefront of developments in hospital-level computing within the NHS for around 20 years. It has acquired a portfolio of proprietary and commercially-sourced systems which provide a wide range of practical functions, including patient administration systems; document, report and results generation; and archiving. The UHS Clinical Data Environment (CDE) has been in progressive and incremental development with a mix of legacy, nationally-specified and locally-developed systems. This is a challenge which is common to all hospitals in the UK. Specifically, all documents and clinical results are accessible on a common digital platform. Figure 1 shows the overall architecture of the UHS-CDE and its relation with the Electronic Patient Record (EPR) and the Southampton Breast Cancer Data System. The recent Lifelines project has looked at displaying this integrated EPR graphically as an alternative to traditional tabular/paged format information.

The presentation of patient records using graphical techniques, colour and icons allows complex data to be rapidly visualised across the time history. An effective visualisation tool should conform to the mantra of “Overview > Zoom > Filter”, which has been promoted by Professor Schneiderman of the Human Computer Interaction Laboratory (HCIL) at University of Maryland [13]. It was a key concept in the design of visualisation tools for the study of complex and heterogeneous patient data, although

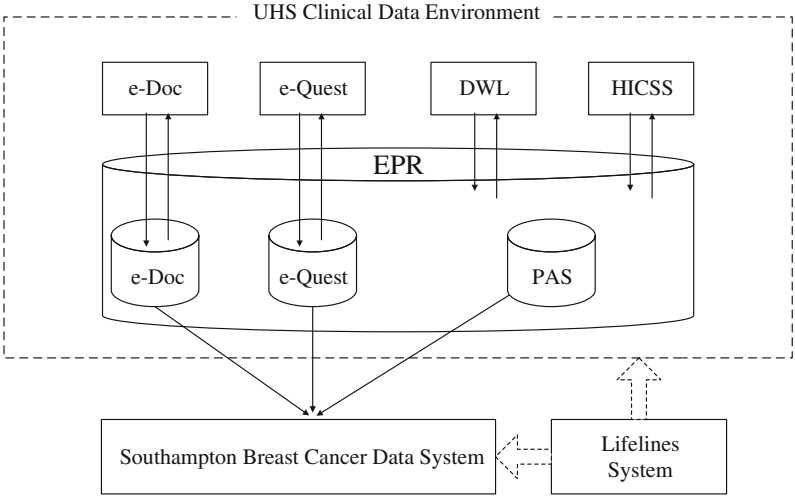


Fig. 1. UHS clinical data environment

their model was focused upon primary care and it was not developed further or implemented in practice.

UHS used an iterative design process to develop "in-house" an innovative, graphically rich, colour-coded, timeline-structured, interactive and real time EPR within the NHS Trust CDE. This coordinated and displayed icons relating to a wide range of clinical specialities onto a single screen of parallel workstream timelines [19]. Each icon indicates a document, test result or report: hovering over any icon reveals the metadata describing when and where it was generated, and by whom; clicking on the icon brings up the full document/record/report. The subject colour-coding reflects an established coding of paper records which has been in use at UHS for around 50 years and is therefore familiar to hospital staff.

The system which was developed is highly intuitive and has an easy-to-use interface, where the model may represent the preferred route of entry into UHS-CDE for the majority of data transactions.

3.2 UHS Breast Cancer Data System

It has been recognised that the timeline-based data visualisation model can be used as a generic tool with application in the study of all chronic diseases of childhood and adulthood, and as a template for other forms of health informatics research. It also has application as a clinical Decision Assistance Tool in specialist practice and for multidisciplinary teams. The concept of the Lifelines EPR has thus been extended to the development of an integrated data system within the UHS-CDE using breast cancer as an exemplar.

SBCDS has evolved on a "design, test and adjust" basis over three years of local development. It currently holds some 16,000 + records on all patients diagnosed and treated locally and a further 4,000 + cases referred for chemotherapy and radiotherapy

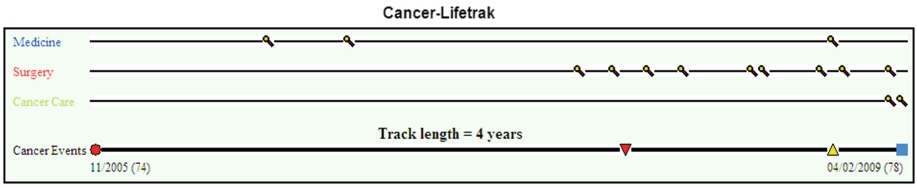


Fig. 2. Cancer Lifetrak record example

from other regional hospitals, since their records began in 1949. The SBCDS records are continuous since 1979 and supported by complete pathology records since 1990.

Figure 2 illustrates the key components of the "whole time course of the disease" structure and function of the SBCDS as an exemplar of applied clinical informatics. The Cancer-Lifetrak records the month of diagnosis: for example showing a red circle (left breast), an inverse triangle (local recurrence) and a yellow triangle (metastasis). A pink ribbon icon would represent death from breast cancer. Of significance for national statistical data is the fact that, although the progression would indicate a likely death from breast cancer, the patient in fact died from a late onset oesophageal cancer. This is evidenced by the supporting documents (clicking on a microscope icon brings up the underlying letter or report).

The successful "ground up" approach which has been developed to the design of a clinically-informative system highlights the opportunities for the next generation of programmers and technologists in health informatics, for which there is a pressing need. This has motivated the collaboration between UHS and the Solent Technology School, thus providing the basis for a number of mutual future benefits in respect of: practical refinements and extension of clinical data systems in the test bed of a forward-thinking NHS Hospital CDE; and education of a coming generation of informatics and technology students on the challenges of clinical systems programming. The rest of the paper will consider pre-processing and present some initial results from undergraduate projects using data mining techniques.

4 Pre-processing for Data Mining

Extracting clinical data from hospital information systems can be time-consuming and labour intensive. This section will demonstrate pre-processing practice with data extracted from the UHS breast cancer data system.

4.1 Research Framework

A research framework has been developed which integrates the data warehousing, OLAP and data mining technologies (see Fig. 3). The data to be included are anonymised patient records and diagnosis records from SBCDS, which also has direct

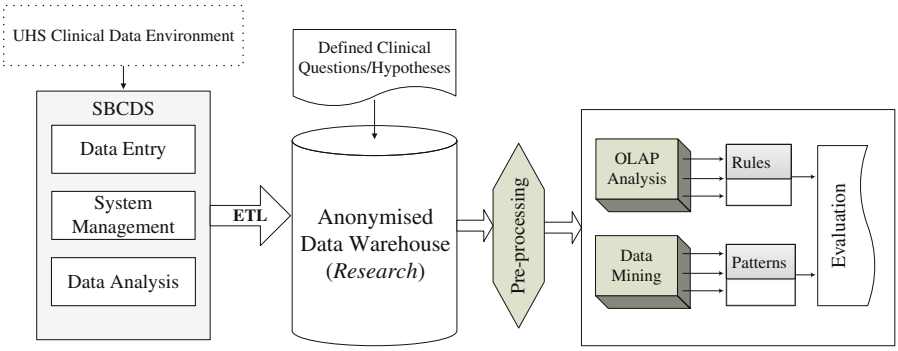


Fig. 3. Research framework for Solent-UHS case study

searches of further information from UHS-CDE, e.g. e-Doc, e-Quest and the Patient Administration System (PAS). The data has been extracted from the breast cancer data system and transformed into patient master and cancer detail tables, before loading to the university-hosted data warehouse.

Data pre-processing is used before mining in order to clean raw data and prepare the final dataset for use in later stages. Data cleaning and preparation stages include basic operations such as removing or reducing noise, handling outliers or missing values, and collecting the necessary information to model. The data from the data warehouse is then in a structured form and available for data mining, visualisation or analysis through (e.g.) OLAP.

4.2 Breast Cancer Datasets: Data Understanding and Preparation

The principle to extract data for this research project is to strictly avoid providing sufficient information that an individual might reasonably be identified by a correlation of seemingly anonymous individual items. Based on this principle, PATIENT_MASTER and CANCER_DETAIL tables have been exported that reflect SBCDS structure and loaded into a relational database. See Tables 1 and 2 for further details on the attributes and their explanation.

To make Table 2 more concise, some attributes are combined together in the display, e.g. CANCER_TYPE_CODE_L and CANCER_TYPE_CODE_R are combined by CANCER_TYPE_CODE_L/R. The data items are almost self-explanatory so long as the clinical aspects are understood. These datasets have been created in a "hybrid format" – containing both a coded and decoded column for each coded field, where the decode is not trivial – given that the main purpose of the research is analysing what exists and not trying to modify the data.

High data quality is one of the primary factors for successful knowledge discovery. There is a need to prepare suitable datasets for the corresponding data mining model before the results can be fully evaluated. Preparing the data for mining may consist of

Table 1. Description of attributes included in PATIENT_MASTER table

Attributes	Description
BCI_ID	Unique but anonymous patient ID
GENDER	(M)ale/(F)emale
YEAR_OF_BIRTH	Year of the Birth Date
YEAR_OF_DEATH	Year of Death or NULL if alive
POSTCODE_DISTRICT	2 Characters of the District section of post code
DECEASED_FLAG	Y if deceased, N or NULL if alive
BCANCER_COD_CODE	Probability code that the Cause of Death (COD) was due to Breast Cancer
OTHER_COD_CODE	Cause of Death code if not due to Breast Cancer
OTHER_CANCER_COD_CODE	Coded Cancer Type assigned as Cause of Death if not Breast Cancer
LOCAL_REFERRAL_FLAG	If (Y)es, originally referred to UHS

Table 2. Description of selected attributes included in CANCER_DETAIL table

Attributes	Description
BCI_ID	Unique but anonymous patient identifier.
BCI_DETAIL_ID	Unique patient/diagnosis event identifier
DIAGNOSIS_MONTH	Month the cancer diagnosis was confirmed
YEARS_OLD_AT_DIAGNOSIS	From DoB and Diagnosis month
LATERALITY	Breast side diagnosed with Cancer: (L)eft, (R)ight or (B)ilateral
CANCER_CATEGORY_CODE	Coded category of cancer type
CANCER_CATEGORY_DESC	Descriptive text of the previous coded field
CANCER_TYPE_CODE_L/R	Cancer type diagnosed on LHS/RHS (NULL if cancer is only on other side)
MAX_TUM_CM_CODE_L/R	Coded size band of tumour if left/right sided
MAX_TUMOR_GRADE_L/R	Tumour grade of the confirmed tumour in the left/right breast
TOTAL_LYMPH_NODES_ CODE_L/R	Coded count range of lymph nodes sampled from the left/right side
NUM_POS_LYMPH_NODES_CODE_L/R	Coded count range of cancer positive lymph nodes sampled from the left/right side
CONSULTANT	Coded value for the Consultant responsible for this diagnosis episode
TREATMENT_LOCATION	Hospital code for the location where treatment was given.
MULTI_FOCAL_L/R	(Y)es if the Breast Cancer presented as multiple tumour sites; (N)o if it is a single mass

several steps such as: selecting relevant attributes, generating the corresponding dataset, cleaning and replacing missing values. In this study 16,319 breast cancer patient records have been extracted from SBCDS and there are 22,380 records (instances) showing their cancer details. However, some of the data in the database is incomplete and incorrect, e.g. several patients were born before 1900 but are still considered to be alive and some other patients do not have an initial record of DoB. Considering the YEAR_OF_BIRTH after 1900 as a cut off, this removes 182 patients and their 197 cancer details from the datasets. The diagnosis of the remaining 16,137 patients is ranging between 1950 and 2014 – the data distribution of other attributes is displayed in Table 3.

Table 3. Attributes distribution among 16,137 breast cancer patients

GENDER	YEAR_OF_BIRTH	POSTCODE	DECEASED_FLAG	LOCAL_REFERRAL
F: 16077	From: 1900	Soton: 64 %	Alive: 10,133	Yes: 11,930
M: 60	To: 1998	Other: 36 %	Dead: 6,004	No/Missing: 3,865

Table 3 shows there are 60 men from SBCDS patient records. Indeed men can develop breast cancer as well, in the small amount of breast tissue behind their nipples. Breast cancer data represents a significant challenge to analyse and present in a way that is not confusing and potentially misleading. The data needs a degree of normalisation (i.e. to establish subsets of data that are logical to compare and contrast) before much sense can be made of it. For example, it is questionable trying to perform analysis on data for deceased and alive patients together, and significant thought must be given to what certain event profiles actually mean.

4.3 Disease Event Sequence Profiles

Querying the patient profiles is the starting point before pre-processing for data mining. Based on the data dictionary corresponding to Tables 1 and 2, the relevant attributes have been highlighted in grey and selected for this purpose. Table 4 shows some raw output from an SQL query against the data extracted for the case study – the data has been limited to local referrals. The query interleaves the date of each disease presentation event to give up to 6 event type/date pairs after the initial diagnosis.

The breast cancer event sequence profiles from Table 4 indicate further data incompleteness, e.g. for some of the records (especially for older patients) there may be no documentary evidence of the type of cancer. They are defined by analysis of the breast tissue under a microscope and in these cases the CANCER_TYPE will be NULL (empty). Then the challenge is deciding what to do with data which is incomplete or inconsistent. The expected survivability from different types of breast cancer varies quite widely so, if the cancer type is not known, it can be statistically questionable to use such data alongside data where the cancer type is known. Further filtering will be performed on the incomplete cases in Sect. 5 for sequential patterns mining.

Table 4. Sample results of patient profiles

INITIAL	PRES1	PRES2	PRES3	PRES4	PRES5	PRES6	STATUS
Loco-RR	Primary	Primary	Loco-RR	Metastatic			Dead
Metastatic	Primary	Metastatic					Dead
Other	Primary	Other	Loco-RR	Metastatic			Dead
Primary	Loco-RR	Loco-RR	Risk-Reduce				Alive
Primary	Loco-RR	Other	Loco-RR	Metastatic			Dead
Primary	Loco-RR	Primary	Loco-RR	Loco-RR	Loco-RR		Dead
Primary	Metastatic	Metastatic	Metastatic	Metastatic	Metastatic	Loco-RR	Dead
Primary	Other	Loco-RR	Loco-RR	Loco-RR			Alive
Primary	Primary	Metastatic	Other	Metastatic	Metastatic		Dead

Key: Primary Breast Cancer (Primary), Loco-Regional Recurrence (Loco-RR), Metastatic Disease (Metastatic), Risk Reducing Mastectomy (Risk-Reduce), Other Cancer Diagnoses (Other)

5 Experiments and Evaluation

This section starts with sequential patterns mining using disease event sequence profiles as input. Further data pre-processing and attribute selection was required for classification techniques to show the applicability of different classifiers in breast cancer research.

5.1 Sequential Patterns Mining

There are 185 distinct disease event sequence profiles which correspond to 12,048 instances. The following pre-processing approach has been pursued to ensure the data is represented as accurately as possible: removal of instances where (1) there is no presentation at all – 894 records have been deleted; (2) initial presentation is anything other than primary breast cancer – a further 170 records have been deleted; (3) two or more presentations of primary cancer exist (when cancer is unilateral). In summary about 8.8 % of the disease event sequences have been removed and this dataset is then divided into two sub-groups: alive (6,489) and dead (4,495). The GSP (Generalized Sequential Patterns) algorithm has been used through Weka [14] for sequential patterns mining.

Table 5. Sample results from sequential patterns mining

	minsup = 5 %	*minsup* = 0.5 %
ΣSPs	3	9
1-sequence	<{Loco-RR} > (365) <{Primary} > (6489)	<{Loco-RR} > (365), < {Metastatic} > (174) <{Primary} > (6489), < {Other} > (113)
2-sequences	<{Primary} {Loco-RR} > (365)	<{Loco-RR}{Loco-RR} > (38) <{Primary}{Loco-RR} > (365) <{Primary}{Metastatic} > (174) <{Primary}{Other} > (113)
3-sequences		<{Primary}{Loco-RR}{Loco-RR} > (38)

Table 5 shows the results for alive patients under minimum support *minsup* = 5 % and 0.5 % respectively, where the numbers of patients are in brackets.

It is noted that among the disease event sequences for live patients, about 90 % have a single item, i.e. only initial Primary Breast Cancer has been recorded. To discover the rules for the rest of the sequences, these singular instances have been removed and the new dataset transformed to contain 1,400 instances in ARFF as required by Weka. Similarly, the disease event sequences dataset for dead patients has been reduced and contains 5,490 instances. The same mining approach has been applied and new results are displayed in Table 6 for live patients. It is not necessary to list all 32 of the sequential patterns which have been discovered when *minsup* = 0.5 % – some representative ones have been selected and shown in Table 6. They are maximal patterns, i.e. they are not contained by other sequential patterns.

Table 6. Reduced results from sequential patterns mining

	minsup = 5 %	*minsup* = 0.5 %
ΣSPs	9	32
3-sequences	<{Primary}{Loco-RR}{Loco-RR}>	<{Primary}{Loco-RR}{Other}> <{Primary}{Metastatic}{Loco-RR}> <{Primary}{Other}{Loco-RR}> <{Primary}{Other}{Metastatic}> <{Primary}{Other}{Other}>
4-sequences		<{Primary}{Loco-RR}{Loco-RR}{Metastatic}> <{Primary}{Loco-RR}{Loco-RR}{Loco-RR}> <{Primary}{Metastatic}{Metastatic}{Metastatic}>

A directed acyclic Sequential Patterns Graph (SPG) has been used to represent the maximal sequence patterns [12]. Figure 4 shows the SPGs for both alive and dead patients when *minsup* = 5 %. The dotted areas indicate the differences between the two groups and this may provide further evidence of disease event sequence profiles between alive and dead patients.

With reference to Fig. 4, it is seen that nodes of SPG correspond to elements (or disease events) in a sequential pattern and directed edges are used to denote the sequence relation between two elements. Any path from a start node to a final node corresponds to one maximal sequence. Figure 5 further demonstrates the SPG model for alive patients only when *minsup* = 0.5 % based on the final column results in Table 6.

Taking the dotted line from Fig. 5 as a sample path for illustration, there are 365 patients with the diagnosis pattern of <{Primary}{Loco-RR}> and out of this group there are 38 cases presented as <{Primary}{Loco-RR}{Loco-RR}>. And finally only 3 instances match the pattern of <{Primary}{Loco-RR}{Loco-RR}{Metastatic}>.

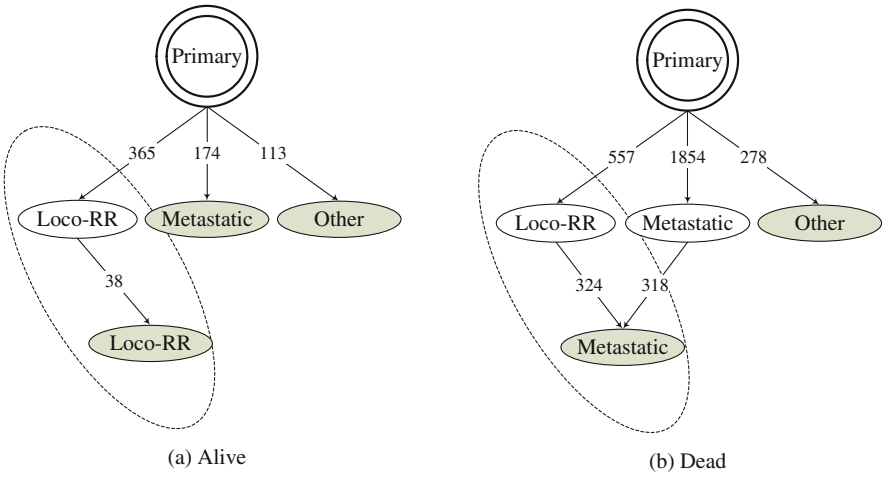


Fig. 4. SPG for maximal sequential patterns when *minsup* = 5 % (all patients)

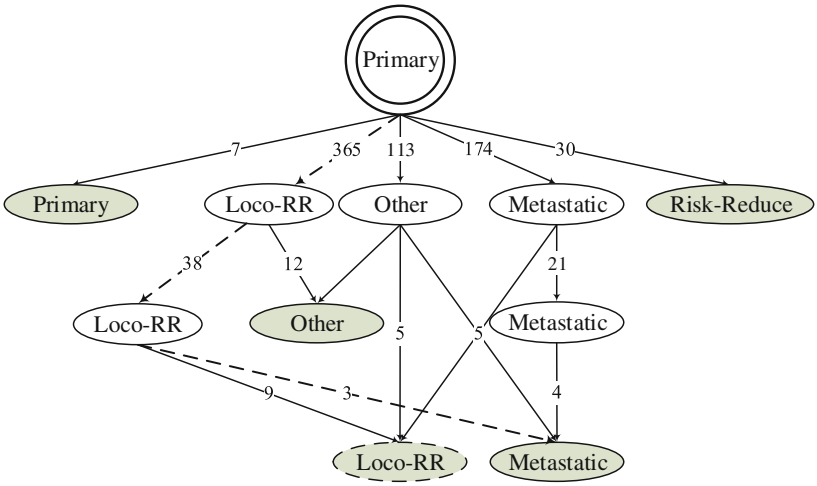


Fig. 5. SPG for maximal sequential patterns when *minsup* = 0.5 % (alive patients)

5.2 Classification Approaches

69 classification algorithms have been evaluated in Weka and compared according to a number of measures: (1) accuracy, sensitivity and specificity (ASS), (2) n-fold cross validation and (3) receiver operating characteristic (ROC). The study in this paper is based on 10-fold cross validation with the measures of correctly classified instances, incorrectly classified instances, Kappa statistic, absolute error, root mean squared error, etc. The percentage of correctly classified instances is often called accuracy and it is a metric widely used in machine learning [5]. However, for medical applications, classification

accuracy is not necessarily the best quality measure of a classifier and thus the Kappa statistic was used here: it measures the agreement of prediction with the true class where 1.0 signifies complete agreement.

The following classifiers were tested in Weka Explorer: Naïve Bayes, Multilayer Perceptron, LibSVM, Logistic and J48 [14]. These models were selected due to their past success and/or prevalence within the healthcare domain. The following attributes have been selected as the predictor sets: MNTHS_TO_1, MNTHS_TO_2, Age and specific treatments such as Hormone, Surgery, Primary Chemotherapy and Adjuvant Radiotherapy. The outcome STATUS is either "Alive" or "Dead". MNTHS_TO_1 is the number of months that have passed between initial presentation and the next episode, i.e. from INITIAL to PRES1 according to Table 4, and similarly for MNTHS_TO_2. Both of these two months/values have to exist, i.e. must not be NULL. With this restriction, about 8,500 instances have been removed because these patients only have "Primary Breast Cancer" as the initial presentation, with no further episodes as yet. Figure 6 presents the results from Weka Decision Tree J48 for illustration, which implements Quinlan's C4.5 algorithm [16].

The classification tree in Fig. 6 shows the treatment that breast cancer patients have received and the intervals from initial presentation to the first and subsequent episode. Terminal nodes in the tree are represented using rectangular boxes along with a numerical breakdown of how many patients are Alive/Dead. The structure of the tree can help to resolve questions on how specific treatment conditions relate to outcome; e.g. patients who have had a second diagnosis within 3.5 years and a third diagnosis after more than 34 months are most likely deceased. Additionally the outcome can be linked with the age of patients and their treatment choices. While the results here are

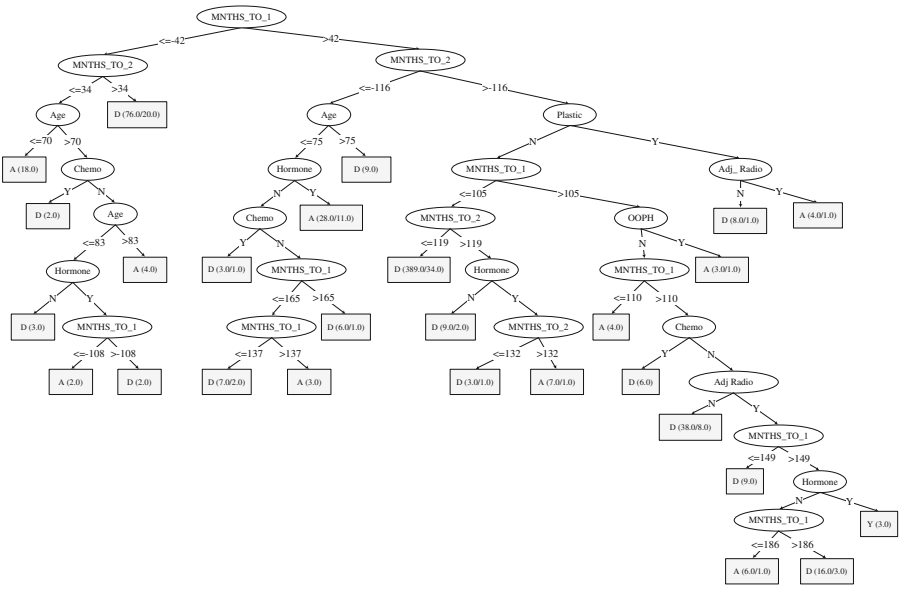


Fig. 6. Decision tree for disease event sequence profiles

under evaluation and open to further interpretation, the structure and visualisation of the above classification offers real potential for this health application.

6 Conclusion

Data mining techniques have been applied in this paper to a case study from breast cancer research, with particular focus on sequential patterns mining and classification. A large anonymised dataset drawn from the Southampton Breast Cancer Data System has been explored after data cleansing and transformation. The processes described here could be applicable to other types of disease or clinical decision. Moreover, the approaches could be extended to future research that employs other techniques such as clustering and temporal patterns mining, e.g. using fuzzy time intervals.

There is every prospect that data mining methods are capable of extracting patterns and relationships hidden behind these large clinical datasets. However, without the knowledge and insight of the domain experts, the results can be of limited value. The findings from this research have been discussed with the consultant surgeon and health systems specialist, and some suggestions made for the next stage. It would be interesting to pursue median time intervals rather than mean intervals and it may be possible to suggest a hypothesis based on the information revealed by this analysis: e.g. patients receiving a given treatment regime appear to survive longer than those on alternative regimes; or certain treatments do not seem to enhance patient survival.

Looking at the intervals between initial diagnosis and subsequent diagnoses across all patients would be very challenging to represent in a meaningful way. SBCDS does not record the order in which treatments occur between presentations. In some cases it will be obvious to the specialists, e.g. adjuvant oncology is always prior to surgery and hormone treatment is usually given at a certain stage of the treatment plan. Ultimately, the decisions about which information is sought will come from the clinical researchers, but for now it is valuable simply to generate some early outcomes as "food for thought" in terms of how the system and its data can be exploited.

References

1. Burke, H.B., Rosen, D., Goodman, P.: Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. In: Tesauro, G., Touretzky, D., Leen, T. (eds.) Advances in Neural Information Processing Systems, pp. 1063–1068. MIT Press, Cambridge (1995)
2. Campbell, K., Thygeson, N.N., Srivastava, J., Speedie, S.: Exploration of Classification Techniques as a Treatment Decision Support Tool for Patients with Uterine Fibroids. In: International Workshop on Data Mining for HealthCare Management, PAKDD (2010)
3. Delen, D., Walker, G., Kadam, A.: Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med. **34**(2), 113–127 (2005)
4. Fayyad, U., PiatetskyShapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine. **17**(3), 37–54 (1996)

5. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann. (2011)
6. Jacob, S.G., Ramani, R.G.: Data mining in clinical data sets: a review. Int. J. Appl. Inf. Syst. **4**(6), 15–16 (2012)
7. Jerez-Aragones, J.M., Gomez-Ruiz, J.A., Ramos-Jimenez, G., MunozPerez, J., Alba-Conejo, E.: A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif. Intell. Med. **27**(1), 45–63 (2003)
8. Holzinger, A.: Trends in interactive knowledge discovery for personalized medicine: cognitive science meets machine learning. IEEE Intell. Inform. Bull. **15**(1), 6–14 (2014)
9. Laxminarayan, P., Alvarez, S.A., Ruiz, C., Moonis, M.: Mining statistically significant associations for exploratory analysis of human sleep data. IEEE Trans. Inf Technol. Biomed. **10**(3), 440–450 (2006)
10. Lee, Y.J., Mangasarian, O.L., Wolberg, W.H.: Survival-time classification of breast cancer patients. Comput. Optim. Appl. **25**(1–3), 151–166 (2003)
11. Li, Q., Feng, J., Wang, L., Chu, H., Yu, H.: Method for knowledge acquisition and decision-making process analysis in clinical decision support system. In: Bursa, M., Khuri, S., Renda, M. (eds.) ITBAM 2014. LNCS, vol. 8649, pp. 79–82. Springer, Heidelberg (2014)
12. Lu, J., Chen, W.R., Adjei, O., Keech, M.: Sequential patterns post-processing for structural relation patterns mining. Int. J. Data Warehousing and Mining **4**(3), 71–89 (2008). IGI Global, Hershey, Pennsylvania
13. Mahajan, R., Shneiderman, B.: Visual and textual consistency checking tools for graphical user interfaces. IEEE Trans. Software Eng. **23**(11), 722–735 (1997)
14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1), 10–11 (2009)
15. Martin, M.A., Meyricke, R., O'Neill, T., Roberts, S.: Mastectomy or breast conserving surgery? factors affecting type of surgical treatment for breast cancer: a classification tree approach. BMC Cancer **6**, 98 (2006)
16. Quinlan, J. Ross. C4.5: Programs for Machine Learning. Elsevier (2014)
17. Razavi, A.R., Gill, H., Ahlfeldt, H., Shahsavar, N.: Predicting metastasis in breast cancer: comparing a decision tree with domain experts. J. Med. Syst. **31**, 263–273 (2007)
18. Reps, J., Garibaldi, J.M., Aickelin, U., Soria, D., Gibson, J.E., Hubbard, R.B.: Discovering Sequential Patterns in a UK General Practice Database. In: IEEE-EMBS International Conference on Biomedical and Health Informatics, pp. 960–963 (2012)
19. Rew, D.A.: Understanding outcomes in cancer surgery through time structured patient records. Indian J. Surg. Oncol. **2**(4), 265–270 (2011)
20. Stolba, N., Tjoa, A.: The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision making. Int. J. Comput. Syst. Sci. Eng. **3**(3), 143–148 (2006)

Artificial Neural Networks in Diagnosis of Liver Diseases

José Neves[1(✉)], Adriana Cunha[2], Ana Almeida[2], André Carvalho[2], João Neves[3], António Abelha[1], José Machado[1], and Henrique Vicente[4]

[1] Centro Algoritmi, Universidade Do Minho, Braga, Portugal
{jneves, abelha, jmac}@di.uminho.pt
[2] Departamento de Informática, Universidade Do Minho, Braga, Portugal
{adrianamadalenacunha, filipanovo, andrenekas}@gmail.com
[3] Drs. Nicolas and Asp, Dubai, United Arab Emirates
joaocpneves@gmail.com
[4] Departamento de Química, Centro de Química de Évora, Escola de Ciências e Tecnologia, Universidade de Évora, Évora, Portugal
hvicente@uevora.pt

Abstract. Liver diseases have severe patients' consequences, being one of the main causes of premature death. These facts reveal the centrality of one`s daily habits, and how important it is the early diagnosis of these kind of illnesses, not only to the patients themselves, but also to the society in general. Therefore, this work will focus on the development of a diagnosis support system to these kind of maladies, built under a formal framework based on Logic Programming, in terms of its knowledge representation and reasoning procedures, complemented with an approach to computing grounded on Artificial Neural Networks.

Keywords: Liver disease · Healthcare · Logic Programming · Knowledge representation and reasoning · Artificial neuronal networks

1 Introduction

Liver disease stands for one of the main causes of premature death, with high treatment costs, and the lost of working times [1]. Several factors may be associated to this illness, namely genetic factors, environmental and lifestyle issues (e.g. dietetic, exercise), viruses, obesity, and alcohol [2], or, in other words, alcoholism, autoimmune diseases, hereditary conditions, drugs and exposure to toxins through ingestion, inhalation, or skin absorption, long-term use of certain medications, diabetes, obesity, and even high levels of triglycerides in blood [3, 4]. Indeed, alcoholic disorder and non-alcoholic fatty syndromes are significant causes of chronic liver disease worldwide, i.e., histological lesions that can include steatosis, which may evolve to cirrhosis, and may lead to liver failure. Nonalcoholic steatohepatitis stands for the more severe end of this

M.E. Renda et al. (Eds.): ITBAM 2015, LNCS 9267, pp. 71–80, 2015.
DOI: 10.1007/978-3-319-22741-2_7

spectrum and is associated with infection progression and the development of liver fibrosis, cirrhosis and hepatocellular carcinoma [5–7].

Physicians usually start with the patient health history, ask about lifestyle habits and may recommend physical examinations, which may include blood, imaging, and/or tissue analysis. Regarding the former one, since liver contains thousands of enzymes, where a few of them are routinely used as indicators of its behaviour, namely ALkaline Phosphatase (ALP), ALanine aminoTransferase (ALT), ASpartate aminoTransferase (AST), Gamma-Glutamyl TransPeptidase (GGTP), 5'-NucleoTidase (5NT), Lactate DeHydrogenase (LDH), serum bilirubin test, albumin test, and even the prothrombin time test [8]. This test may help in measuring the liver's ability to synthesize cells, since most blood clotting factors are produced in the liver. Another commonly test used in this context is the Mean Corpuscular Volume (MCV). It is a measure of the average volume of red blood cells and their increase (macrocytosis) may point out to alcohol abuse and/or other problems [9].

Concerning imaging tests, several modalities are available, like computed tomography, magnetic resonance imaging and endoscopic ultrasonography [8]. Regarding tissue analysis, it consists in collecting a liver tissue sample in order to perform a laboratorial analysis. Liver biopsy remains as the definitive test to confirm the diagnosis of particular liver diseases like the Wilson one. However, this technique is absolutely contra-indicated in patients with inexplicable bleeding history, prothrombine time higher than 3 to 4 s over control, platelets less than 60000/mm^3, prolonged bleeding time, unavailability of blood transfusion support, suspected hemangioma and uncooperative patient behavior [8, 10].

Liver disease is typically asymptomatic until the development of clinical complications. Unfortunately, these snags appear at a relatively late stage of the progression of the disease. Furthermore, most risk factors for liver disease are also risk features for other ones (e.g., excess of alcohol consumption is associated with an increased risk of alcoholic liver disease and breast cancer [11]; obesity is associated with an increased risk of both non-alcoholic fatty liver disease and the coronary heart one [12]).

The stated above shows that it is difficult to make an early diagnosis of liver disease, since it needs to consider different conditions with intricate relations among them, where the available data may be incomplete, contradictory and even unknown. In order to overcome these drawbacks, the present work reports the founding of a computational framework that uses knowledge representation and reasoning techniques to set the structure of the information system and the associate inference mechanisms. We will centre on a Logic Programming based approach to knowledge representation and reasoning [13, 14], and look at a Soft Computing approach to data processing based on Artificial Neural Networks (ANNs) [15].

This paper is organized into five sections. In the former one an introduction to the problem presented is made. Then the proposed approach to knowledge representation and reasoning is introduced. In the third and fourth sections is introduced a case study and presented a solution to the problem. Finally, in the last section the most relevant conclusions are termed and the possible directions for future work are outlined.

2 Knowledge Representation and Reasoning

Many approaches to knowledge representation and reasoning have been proposed using the Logic Programming (LP) epitome, namely in the area of Model Theory [16, 17], and Proof Theory [13, 14]. In this work it is followed the proof theoretical approach in terms of an extension to the LP language. An Extended Logic Program is a finite set of clauses, given in the form:

$$\begin{aligned}\{\ & p \leftarrow p_1, \cdots, p_n, not\ q_1, \cdots, not\ q_m \\ & ?(p_1, \cdots, p_n, not\ q_1, \cdots, not\ q_m)(n, m \geq 0) \\ & exception_{p_1}, \cdots, exception_{p_j}(j \geq 0)\} :: scoring_{value}\end{aligned}$$

where "*?*" is a domain atom denoting falsity, the p_i, q_j, and p are classical ground literals, i.e., either positive atoms or atoms preceded by the classical negation sign ¬ [13]. Under this formalism, every program is associated with a set of abducibles [16, 17], given here in the form of exceptions to the extensions of the predicates that make the program. The term $scoring_{value}$ stands for the relative weight of the extension of a specific *predicate* with respect to the extensions of the peers ones that make the inclusive or global program.

In order to evaluate the knowledge that stems from a logic program, an assessment of the *Quality-of-Information* (*QoI*), given by a truth-value in the interval [0, 1], that stems from the extensions of the predicates that make a program, inclusive in dynamic environments, aiming at decision-making purposes, was set [18, 19]. Indeed, the objective is to build a quantification process of *QoI* and measure one's confidence (here represented as *DoC*, that stands for *Degree of Confidence*) that the argument values of a given predicate with relation to their domains fit into a given interval [20].

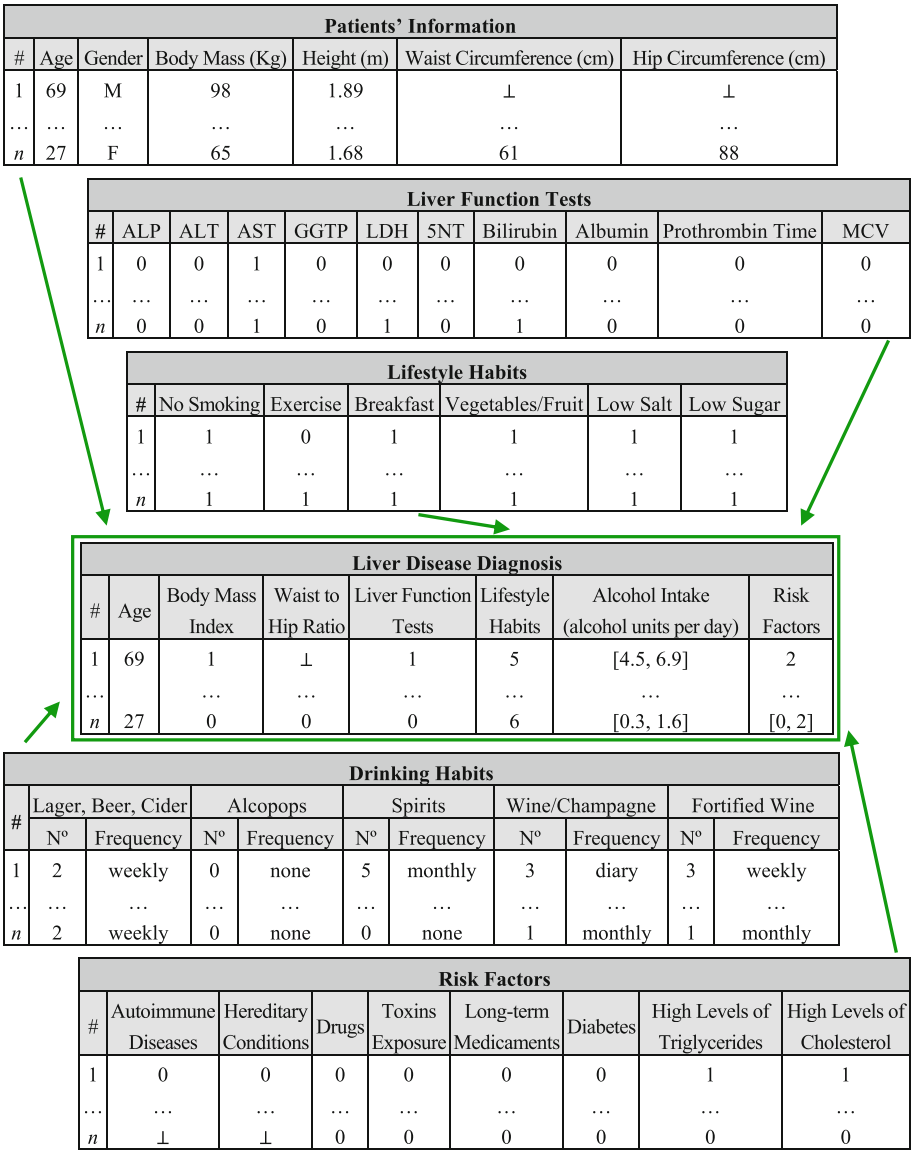
Therefore, the universe of discourse is engendered according to the information presented in the extensions of a given set of predicates, according to productions of the type:

$$predicate_i - \bigcup_{1 \leq j \leq m} clause_j(x_1, \cdots, x_n) :: QoI_i :: DoC_i \tag{1}$$

where ∪ and m stand, respectively, for *set union* and the *cardinality* of the extension of $predicate_i$.

3 A Case Study

As a case study, consider a database given in terms of the extensions of the relations (or tables) depicted in Fig. 1, which stands for a situation where one has to manage information about patients who may suffer from liver diseases. Under this scenario some incomplete and/or default data is also available. For instance, in the *Liver Disease*



Patients' Information

#	Age	Gender	Body Mass (Kg)	Height (m)	Waist Circumference (cm)	Hip Circumference (cm)
1	69	M	98	1.89	⊥	⊥
…	…	…	…	…	…	…
n	27	F	65	1.68	61	88

Liver Function Tests

#	ALP	ALT	AST	GGTP	LDH	5NT	Bilirubin	Albumin	Prothrombin Time	MCV
1	0	0	1	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…
n	0	0	1	0	1	0	1	0	0	0

Lifestyle Habits

#	No Smoking	Exercise	Breakfast	Vegetables/Fruit	Low Salt	Low Sugar
1	1	0	1	1	1	1
…	…	…	…	…	…	…
n	1	1	1	1	1	1

Liver Disease Diagnosis

#	Age	Body Mass Index	Waist to Hip Ratio	Liver Function Tests	Lifestyle Habits	Alcohol Intake (alcohol units per day)	Risk Factors
1	69	1	⊥	1	5	[4.5, 6.9]	2
…	…	…	…	…	…	…	…
n	27	0	0	0	6	[0.3, 1.6]	[0, 2]

Drinking Habits

#	Lager, Beer, Cider		Alcopops		Spirits		Wine/Champagne		Fortified Wine	
	Nº	Frequency	Nº	Frequency	Nº	Frequency	Nº	Frequency	Nº	Frequency
1	2	weekly	0	none	5	monthly	3	diary	3	weekly
…	…	…	…	…	…	…	…	…	…	…
n	2	weekly	0	none	0	none	1	monthly	1	monthly

Risk Factors

#	Autoimmune Diseases	Hereditary Conditions	Drugs	Toxins Exposure	Long-term Medicaments	Diabetes	High Levels of Triglycerides	High Levels of Cholesterol
1	0	0	0	0	0	0	1	1
…	…	…	…	…	…	…	…	…
n	⊥	⊥	0	0	0	0	0	0

Fig. 1. An extension of the relational model for liver diseases diagnosis.

Diagnosis table, the *Waist to Hip Ratio* in case 1 is unknown, while the *Alcohol Intake* ranges in the interval [4.5, 12.9].

In *Liver Function Tests* table, 0 (zero) and 1 (one) denote, respectively, *normal* and *abnormal* values. In *Lifestyle Habits* and *Risk Factors* tables, 0 (zero) and 1 (one) denote, respectively, *no* and *yes*. In *Drinking Habits* table *N* stands for the number of beverages consumed.

Table 1. Waist to hip ratio and disease risk related to obesity, stratified by age and gender.

Age (years)	Men				Women			
	Low	Moderate	High	Very high	Low	Moderate	High	Very high
[20,30[	< 0.83	[0.83, 0.89[	[0.89, 0.94[	≥ 0.94	< 0.71	[0.71, 0.78[	[0.78, 0.82[	≥ 0.82
[30,40[	< 0.84	[0.84, 0.92[	[0.92, 0.96[	≥ 0.96	< 0.72	[0.72, 0.79[	[0.79, 0.84[	≥ 0.84
[40,50[	< 0.88	[0.88, 0.96[	[0.96, 1.00[	≥ 1.00	< 0.73	[0.73, 0.80[	[0.80, 0.87[	≥ 0.87
[50,60[	< 0.90	[0.90, 0.97[	[0.97, 1.02[	≥ 1,02	< 0.74	[0.74, 0.82[	[0.82, 0.88[	≥ 0.88
≥60	< 0.91	[0.91, 0.99[	[0.99, 1.03[	≥ 1,03	< 0.76	[0.76, 0.84[	[0.84, 0.90[	≥ 0.90

In the *Liver Disease Diagnosis* table, the domain of *Body Mass Index* column is in the range [0, 2], wherein 0 (zero) denotes $BMI < 25$; 1 (one) stands for a *BMI* ranging in interval [25, 30[; and 2 (two) denotes a $BMI \geq 30$. The *Body Mass Index* (*BMI*) is evaluated using the equation $BMI = BodyMass/Height^2$ [21]. *Waist to Hip Ratio* column ranges in the interval [0, 3] according to Table 1, adapted from [22], wherein 0 (zero), 1 (one), 2 (two) and 3 (three) denote disease risk related to obesity, in terms of a qualification of *low*, *moderate*, *high* and *very high*.

The alcohol units for most common beverage were calculated in terms of Eq. 2, according to what is set in [23, 24], while the values of the *Alcohol Intake* column of *Liver Disease Diagnosis* table was calculated using Eq. 3.

$$Alcohol\ Units = Alcohol\ by\ Volume(\%) * Volume\left(cm^3\right)/1000 \tag{2}$$

$$Alcohol\ Intake = \sum_{\substack{beverage \\ types}} N * Frequency\ Factor * Alcohol\ Units \tag{3}$$

where *N* stands for the number of beverages consumed. The frequency factor is 1, 1/7, 1/30 and 0 depending on the intake frequency, i.e., daily, weekly, monthly or none.

The values presented in the remaining columns of *Liver Disease Diagnosis* table are the sum of the ones of the correspondent tables, ranging between [0, 10], [0, 6] and [0, 8], respectively for *Liver Function Tests*, *Lifestyle Habits* and *Risk Factors* columns.

Now, we may consider the relations given in Fig. 1, in terms of the extension of the *liver disease diagnosis* predicate, depicted in the form:

$$
\begin{aligned}
&\{\\
&\quad \neg liver_{disease\_diagnosis}(Age, BMI, W/H, LFT, LH, AIn, RF)\\
&\qquad\qquad \leftarrow not\ liver_{disease\_diagnosis}(Age, BMI, W/H, LFT, LH, AIn, RF)\\
&\quad liver_{disease\_diagnosis}\Big(\underbrace{1,\ 1,\ 0,\ 1,\ 1,\ 0.99,\ 1}_{attribute`s\ confidence\ values}\Big) :: 1 :: 0.86\\
&\qquad\qquad \underbrace{[0.7,0.7][0.5,0.5][0,1][0.1,0.1][0.8,0.8][0.2,0.3][0.25,0.25]}_{attribute`s\ values\ ranges\ once\ normalized}\\
&\qquad\qquad \underbrace{[0,1]\ \ [0,1]\ \ [0,1]\ \ [0,1]\ \ [0,1]\ \ [0,1]\ \ [0,1]}_{attribute`s\ domains\ once\ normalized}\\
&\ldots\\
&\}:: 1
\end{aligned}
$$

where its argument's values make the training and test sets of the Artificial Neural Network (ANN) given in Fig. 2. Now, let us consider a patient that presents the symptoms $Age = 58$, $BMI = \bot$, $W/H = 2$, $LFT = 2$, $LH = 3$, $AIn = [4.5, 13.1]$, $RF = 3$, to which it is applied the procedure presented in [20]. One may have:

$$
\begin{aligned}
&\{\\
&\quad \neg liver_{disease\_diagnosis}(Age, BMI, W/H, LFT, LH, AIn, RF)\\
&\qquad\qquad \leftarrow not\ liver_{disease\_diagnosis}(Age, BMI, W/H, LFT, LH, AIn, RF)\\
&\quad liver_{disease\_diagnosis}\Big(\underbrace{1,\ 0,\ 1,\ 1,\ 1,\ 0.91,\ 1}_{attribute`s\ confidence\ values}\Big) :: 1 :: 0.84\\
&\}:: 1
\end{aligned}
$$

4 Artificial Neural Networks

It was set a soft computing approach to model the universe of discourse of any patient suffering from liver disease, based on ANNs, which are used to structure data and capture complex relationships between inputs and outputs [25, 26]. ANNs simulate the structure of the human brain, being populated by multiple layers of neurons, with a valuable set of activation functions. As an example, let us consider the case listed above, where one may have a situation in which the diagnosis of liver disease is needed. In Fig. 2 it is shown how the normalized values of the interval boundaries and their DoC_s and QoI_s values work as inputs to the ANN. The output depicts a liver disease diagnostic, plus the confidence that one has on such a happening.

In this study 438 patients were considered with an age average of 65.4 years, ranging from 22 to 93 years old. Liver diseases was diagnosed in 73 cases, i.e., in

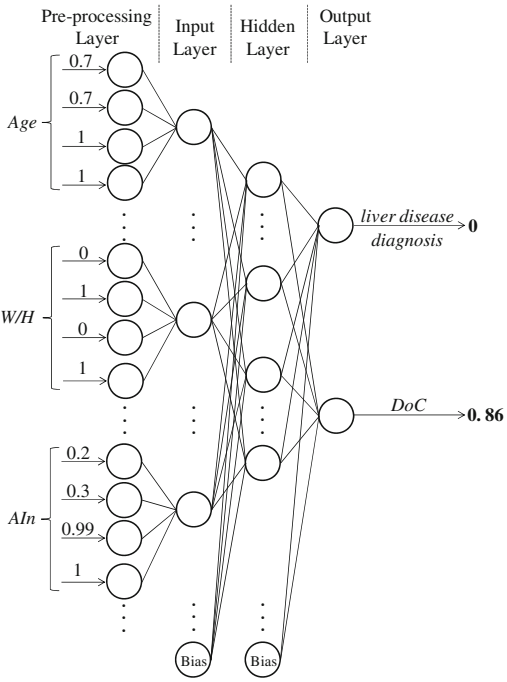


Fig. 2. The artificial neural network topology.

16.7 % of the analysed population. The gender distribution was 41.6 % and 58.4 % for female and male, respectively.

In each simulation, the available data was randomly divided into two mutually exclusive partitions, i.e., the training set with 70 % of the available data, used during the modeling phase, and the test set with the remaining 30 % of the cases, used after training in order to evaluate the model performance and to validate it. The back propagation algorithm was applied in the learning process of the ANN. The activation function used in the pre-processing layer was the identity one. In the other layers was used the sigmoid activation function.

A common tool to evaluate the results presented by the classification models is the coincidence matrix, a matrix of size $L \times L$, where L denotes the number of possible classes. This matrix is created by matching the predicted and target values. L was set to 2 (two) in the present case. Table 2 presents the coincidence matrix (the values denote the average of the 30 runs). A perusal of Table 2 shows that the model accuracy was 96.1 % for the training set (296 correctly classified in 308) and 95.4 % for test set (124 correctly classified in 130).

Based on coincidence matrix it is possible to compute sensitivity, specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) of the classifier. Briefly, sensitivity and specificity are measures of the performance of a binary classifier. Sensitivity evaluates the proportion of true positives that are correctly identified as such, while specificity translates the proportion of true negatives that are correctly

Table 2. The coincidence matrix for ANN model.

Target	Predictive			
	Training set		Test set	
	True (1)	False (0)	True (1)	False (0)
True (1)	49	3	20	1
False (0)	9	247	5	104

identified. Moreover, it is necessary to know the probability of the classifier that give the correct diagnosis. Thus, it is also calculated both PPV and NPV, while PPV stands for the proportion of cases with positive results which are correctly diagnosed, NPV is the proportion of cases with negative results which are successfully labeled. The sensitivity ranges from 94.2 % to 95.2 %, while the specificity ranges from 95.4 % to 96.5 %. PPV ranges from 80.0 % to 84.4 %, while NPV ranges from 98.8 % to 99.0 %. Thus, it is our claim that the proposed model is able to diagnosis liver diseases properly. The inclusion of other patient's characteristics, like lifestyle and drink habits may be responsible for the good performance exhibited by the presented model.

5 Conclusions and Future Work

Diagnosing *liver disease* has shown to be a hard task. On the one hand, the parameters that cause the disorder are not fully represented by objective data. On the other hand, liver disease is asymptomatic until the development of clinical complications that are manifested at a relatively late stage of the progression of the disease. Therefore, it is mandatory to consider many different conditions with intricate relations among them. These characteristics put this problem into the area of problems that may be tackled by Artificial Intelligence based methodologies and techniques to problem solving.

This work presents the founding of a computational framework that uses powerful knowledge representation and reasoning techniques to set the structure of the information and the associate inference mechanisms. This finding is built on a set of presuppositions, namely:

- Data is not equal to information;
- The translation of the raw measurements into interpretable and actionable read-outs is challenging; and
- Read-outs can deliver markers and targets candidates without pre-conception, i.e., knowing how personal conditions and risk factors may affect the liver disease predisposition.

The knowledge representation and reasoning techniques presented above are very versatile and capable of covering almost every possible instance, namely by considering incomplete, contradictory, and even unknown data, a marker that is not present in existing systems. Indeed, this method brings a new approach that can revolutionize prediction tools in all its variants, making it more complete than the existing methodologies and tools for problem solving. The new paradigm of knowledge

representation and reasoning enables the use of the normalized values of the interval boundaries and their *DoC* values, as inputs to the ANN. The output translates a diagnosis of liver disease and the confidence that one has on such a happening.

The main contribution of this work relies on the fact that at the end, the extensions of the predicates that make the universe of discourse are given in terms of *DoCs* values that stand for one's confidence that the predicates arguments values fit into their observable ranges, taking into account their domains. It also encapsulates in itself a new vision of Multi-value Logics, once a proof of a theorem in a conventional way, is evaluated to the interval [0, 1]. Future work may recommend that the same problem must be approached using others computational frameworks like Case Based Reasoning [27], Genetic Programming [14] or Particle Swarm [28], just to name a few.

Acknowledgments. This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

References

1. Lombard, M.: Liver disease. In: Howard, S. (ed.) Annual Report of the Chief Medical Officer, Surveillance. On the State of the Public's Health, vol. 2012, pp. 95–108 (2012)
2. Day, C.P.: Genes or environment to determine alcoholic liver disease and nonalcoholic fatty liver disease. Liver Int. **26**, 1021–1028 (2006)
3. Koteish, A., Diehi, A.M.: Obesity and liver disease. Curr. Treat. Options Gastroenterol. **4**, 101–105 (2001)
4. Corey, K.E., Kaplan, L.M.: Obesity and liver disease: the epidemic of the twenty-first century. Clin. Liver Dis. **18**, 1–18 (2014)
5. Ahmed, M.H., Byrne, C.D.: Non-alcoholic fatty liver disease. In: Byrne, C.D., Wild, S.H. (eds.) Metabolic Syndrome, pp. 245–277. Wiley-Blackwell, Chichester (2011)
6. Yeh, M.M., Brunt, E.M.: Pathological features of fatty liver disease. Gastroenterology **147**, 754–764 (2014)
7. Luedde, T., Kaplowitz, N., Schwabe, R.F.: Cell death and cell death responses in liver disease: mechanisms and clinical relevance. Gastroenterology **147**, 765–783 (2014)
8. Martin, P., Friedman, L.S.: Assessment of liver function and diagnostic studies. In: Friedman, L.S., Keeffe, E.B. (eds.) Handbook of Liver Disease, 3rd edn, pp. 1–19. Elsevier Saunders, Philadelphia (2011)
9. Maruyama, S., Hirayama, C., Yamamoto, S., Koda, M., Udagawa, A., Kadowaki, Y., Inoue, M., Sagayama, A., Umeki, K.: Red blood cell status in alcoholic and non-alcoholic liver disease. J. Lab. Clin. Med. **138**, 332–337 (2001)
10. Rockey, D.C., Caldwell, S.H., Goodman, Z.D., Nelson, R.C., Smith, A.D.: Liver biopsy. Hepatology **49**, 1017–1044 (2009)
11. Chen, W.Y., Rosner, B., Hankinson, S.E., Graham, A., Colditz, G.A., Willett, W.C.: Moderate alcohol consumption during adult life, drinking patterns, and breast cancer risk. J. Am. Med. Assoc. **306**, 1884–1890 (2011)

12. Go, A.S., Mozaffarian, D., Roger, V.L., Benjamin, E.J., Berry, J.D., Blaha, M.J., Dai, S., Ford, E.S., Fox, C.S., Franco, S., Fullerton, H.J., Gillespie, C., Hailpern, S.M., Heit, J.A., Howard, V.J., Huffman, M.D., Judd, S.E., Kissela, B.M., Kittner, S.J., Lackland, D.T., Lichtman, J.H., Lisabeth, L.D., Mackey, R.H., Magid, D.J., Marcus, G.M., Marelli, A., Matchar, D.B., McGuire, D.K., Mohler 3rd, E.R., Moy, C.S., Mussolino, M.E., Neumar, R. W., Nichol, G., Pandey, D.K., Paynter, N.P., Reeves, M.J., Sorlie, P.D., Stein, J., Towfighi, A., Turan, T.N., Virani, S.S., Wong, N.D., Woo, D., Turner, M.B.: on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee: Heart disease and stroke statistics — 2014 update: a report from the American Heart Association. Circulation **129**, e28–e292 (2014)
13. Neves, J.: A logic interpreter to handle time and negation in logic databases. In: Muller, R.L., Pottmyer, J.J. (eds.) Proceedings of the 1984 Annual Conference of the ACM on The Fifth Generation Challenge, pp. 50–54. Association for Computing Machinery, New York (1984)
14. Neves, J., Machado, J., Analide, C., Abelha, A., Brito, L.: The halt condition in genetic programming. In: Neves, J., Santos, M.F., Machado, J.M. (eds.) EPIA 2007. LNCS (LNAI), vol. 4874, pp. 160–169. Springer, Heidelberg (2007)
15. Cortez, P., Rocha, M., Neves, J.: Evolving time series forecasting ARMA models. J. Heuristics **10**, 415–429 (2004)
16. Kakas, A., Kowalski, R., Toni, F.: The role of abduction in logic programming. In: Gabbay, D., Hogger, C., Robinson, I. (eds.) Handbook of Logic in Artificial Intelligence and Logic Programming, vol. 5, pp. 235–324. Oxford University Press, Oxford (1998)
17. Pereira, L.M., Anh, H.T.: Evolution prospection. In: Nakamatsu, K., Phillips-Wren, G., Jain, L.C., Howlett, R.J. (eds.) New Advances in Intelligent Decision Technologies. SCI, vol. 199, pp. 51–63. Springer, Heidelberg (2009)
18. Lucas, P.: Quality checking of medical guidelines through logical abduction. In: Coenen, F., Preece, A., Mackintosh, A. (eds.) Proceedings of AI-2003 (Research and Developments in Intelligent Systems XX), pp. 309–321. Springer, London (2003)
19. Machado, J., Abelha, A., Novais, P., Neves, J., Neves, J.: Quality of service in healthcare units. Int. J. Comput. Aided Eng. Technol. **2**, 436–449 (2010)
20. Cardoso, L., Marins, F., Magalhães, R., Marins, N., Oliveira, T., Vicente, H., Abelha, A., Machado, J., Neves, J.: Abstract computation in schizophrenia detection through artificial neural network based systems. Sci. World J. **2015**, 1–10 (2015). Article ID 467178
21. World Health Organization: Obesity and overweight. Fact Sheet Number 311. http://www.who.int/mediacentre/factsheets/fs311/en/
22. Heyward, V.H., Wagner, D.R.: Applied Body Composition Assessment, 2nd edn. Human Kinetics, Champaign (2004)
23. National Health Service. http://www.nhs.uk/Livewell/alcohol/Pages/alcohol-units.aspx
24. Kerr, W.C., Stockwell, T.: Understanding standard drinks and drinking guidelines. Drug Alcohol Rev. **31**, 200–205 (2012)
25. Vicente, H., Dias, S., Fernandes, A., Abelha, A., Machado, J., Neves, J.: Prediction of the quality of public water supply using artificial neural networks. J. Water Supply Res. Technol. AQUA **61**, 446–459 (2012)
26. Salvador, C., Martins, M.R., Vicente, H., Neves, J., Arteiro, J.M., Caldeira, A.T.: Modelling molecular and inorganic data of amanita ponderosa mushrooms using artificial neural networks. Agrofor. Syst. **87**, 295–302 (2013)
27. Carneiro, D., Novais, P., Andrade, F., Zeleznikow, J., Neves, J.: Using case-based reasoning and principled negotiation to provide decision support for dispute resolution. Knowl. Inf. Syst. **36**, 789–826 (2013)
28. Mendes, R., Kennedy, J., Neves, J.: The fully informed particle swarm: simpler, maybe better. IEEE Trans. Evol. Comput. **8**, 204–210 (2004)

How to Increase the Effectiveness of the Hepatitis Diagnostics by Means of Appropriate Machine Learning Methods

Alexandra Lukáčová[1], František Babič[1], Zuzana Paraličová[2], and Ján Paralič[1(✉)]

[1] Faculty of Electrical Engineering and Informatics, Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Letná 9/B, 042 00 Košice, Slovakia
{alexandra.lukacova, frantisek.babic, jan.paralic}@tuke.sk
[2] Medical Faculty, University of Pavol Jozef Šafárik, Tr. SNP 1, Košice, Slovakia
zuzana.paralicova@unlp.sk

Abstract. This paper presents how to improve the diagnostic process of hepatitis B and C based on collected questionnaires from patients hospitalized in all regional departments of infectology in Slovakia. Performed experiments were oriented in two directions: economic demands of the recommended treatment based on realized diagnostics and possible improvement of hepatitis diagnostics by means of exploratory and predictive analysis of additional information provided by patients. Exploratory data analysis was used to confirm or to reject some expected relationships between input attributes (e.g. ager or gender) and target diagnosis. Also, predictive mining resulted into interesting decision rules that can be used in medical practice as supporting information at an early stage of the diagnostic process. Finally, analysis of the treatment economic demands based on the estimated costs showed the need for timely and quality diagnostics to minimize the percentage of patients for which was hepatitis diagnosed late.

Keywords: Hepatitis · CHAID · Cost-benefit method

1 Introduction

Traditional diagnostic process is characterized by the manual evaluation of the necessary biomarkers or symptoms by the doctors or medical experts. Available technologies, methods and algorithms create the conditions to make this approach more automatic, interactive and complex. This change is not easy, because it is necessary to consider the amount of assumptions and dependencies. But the expected result can provide more effective diagnosis, data in better quality and more comprehensive view of the patient's medical history.

World Health Organization[1] defines hepatitis as an inflammation of the liver. The condition can be self-limiting or can progress to fibrosis (scarring), cirrhosis or liver cancer.

[1] http://www.who.int/features/qa/76/en/.

M.E. Renda et al. (Eds.): ITBAM 2015, LNCS 9267, pp. 81–94, 2015.
DOI: 10.1007/978-3-319-22741-2_8

Hepatitis viruses are the most common cause of hepatitis in the world but other infections, toxic substances (e.g. alcohol, certain drugs), and autoimmune diseases can also cause hepatitis. Hepatitis B virus (HBV) is transmitted through exposure to infective blood, semen, and other body fluids. HBV can be transmitted from infected mothers to infants at the time of birth or from family member to infant in early childhood. Transmission may also occur through transfusions of HBV-contaminated blood and blood products, contaminated injections during medical procedures, and through injection drug use. HBV also poses a risk to healthcare workers who sustain accidental needle stick injuries while caring for infected-HBV patients. Safe and effective vaccines are available to prevent HBV. Hepatitis C virus (HCV) is mostly transmitted through exposure to infective blood. This may happen through transfusions of HCV-contaminated blood and blood products, contaminated injections during medical procedures, and through injection drug use. Sexual transmission is also possible, but is much less common.

Presented experiments do not represent typical example of hepatitis diagnosis through appropriate biomarkers. Source dataset is represented by collected anonymous questionnaire fulfilled by patients involved in a prospective multicenter study that was organized by the Slovak Infectologists SLS co-financing by the educational grant Roche Slovakia, s.r.o. The study lasted 12 months from April 2010 to March 2011 and involved all regional departments of infectology in Slovakia (Bratislava, Trnava, Nitra, Trenčín, Banská Bystrica, Martin, Košice, Prešov).

1.1 Motivation

The hepatitis C virus infects about 0.67 % of the Slovakian population, representing 35 thousand people according to estimates resulting from epidemiological studies [1]. Despite the increase in the number of diagnosed diseases in recent years, remain more than 90 % of infections with hepatitis C unrecognized. Results of this study are similar. Prevalence of positive antiHCV was 1.4 %. Prevalence of HCV RNA reflecting chronic hepatitis C was 0,70 %. The problem is that the natural course of the disease is long time without clinical symptoms and remains therefore undiagnosed. The disease progresses in approximately 20 % of untreated patients to the liver cirrhosis. Patients with chronic hepatitis C (HC) have an increased risk of developing hepatocellular carcinoma (HCC), whose incidence is estimated to 1–3 % in patients infected with HCV (20 % of all HCC). Treatment costs of chronic hepatitis C consequences markedly increase with the progression of the disease. By experts' opinion, the costs of HCV therapy for patients with early stage disease (i.e. F0–F2 according to Metavir score) is estimated 370€. The costs raises to 740€ in stage F3 and stage of compensated cirrhosis (F4) up to 1.400€. The treatment cost of the decompensation cirrhosis increases rapidly, and it is estimated at 4.000€ for the treatment of ascites, a similar amount is estimated for the treatment of bleeding esophageal varices and treatment of hepatic encephalopathy. Treatment cost of HCC is estimated at 43.700€ and liver transplantation for 82.700€. Similarly, the cost of antiviral therapy significantly increases during disease progression. Standard treatment cost with pegylated interferon + ribavirin is around 12 to 15.000€, while indication of triple therapy increases costs to 35–40.000€. Price for new

treatment regimens without interferon, which are for the time being recommended for patients after failure of first-line treatment in advanced disease climbs to 50 to 90.000€. Similar calculations of treatment delay costs can be found in [2].

Raising costs increase importance of early disease diagnosis and treatment. Approximate price of screening test for anti-HCV antibodies is 7€. In medical practice, the anti-HCV testing is usually indicated in the condition with elevated liver enzymes, and in the presence of risk factors of epidemiological importance. Alaninaminotranspherasis (ALT) is the most commonly investigated liver enzyme. Its normal level reference value range for women from 0,2 to 0,6 μkat/l and for men from 0,2 to 0,8 μkat/l.

The situation with the occurrence of hepatitis B (HB) is even more pronounced. Chronic HB is the most common cause of cirrhosis associated with a high rate of complications. It is responsible for 75 % of HCC with high mortality and for 5–10 % of all liver transplantations. Up to one million people in the world die each year in the context of HB. Past epidemiological surveys indicate that the prevalence of HBsAg (antigen which is evidence of HBV infection) in Slovakia is less than 2 %. Prevalence of positive HBsAg in this study was 1.7 %. The cost of treatment rise continuously with liver disease progression. Annual treatment costs per patient range according to [3] from 761 US dollars for chronic hepatitis B to 86,552 US dollars for liver transplantation.

1.2 Related Work

Medical diagnosis using appropriate methods of data processing and analysis is becoming one of the key research bridges between data mining analysts and medicine experts or doctors. This cooperation can result into interesting experimental studies or decision support systems to optimize the values hidden in the data not only for patients, but for overall improvement of the diseases diagnostics and treatment. Following selected articles describe various approaches how to solve these issues in an effective way.

Using of a new discriminant diagnosis model constructed by attribute selection, algorithm C5.0 for decision trees generation and discrimination analysis is proposed in [4]. The authors used this two level approach for diagnosis of chronic hepatitis B in traditional Chinese medicine (TCM) in order to investigate the relationship of typical symptoms or lab factors with TCM syndromes. In the first level authors selected the critical attributes from the original set of attributes characterizing 1 015 inpatient and outpatient hepatitis B patients from three hospitals in China. Next level dealt with modelling within C5.0 algorithm that used the most significant set of attributes from "western medicine" and TCM. Resulted integrated discriminant (combination of TCM attributes and western medicine indicators) involved 7 attributes with classification accuracy 94.4 %. Our case is different because we are using general population data.

Team of authors from Iran in their work [5] presented an interesting combination of hybridized Support Vector Machine and simulated annealing for hepatitis diagnosis. They used for experiments a dataset from UCI machine learning database included 155 samples and 10-fold cross validation for verification. The best obtained accuracy of the proposed combination was 96.25 % on the selected dataset.

Sathyadevi in his work [6] applied three algorithms from machine learning area to the UCI machine learning database containing records about 473 patients characterized by 19 attributes (e.g. age, bilirubin, phosphate, albumin, etc.). ID3, C4.5 and CART was implemented and tested over this database and the best classification accuracy 83.2 % was obtained by CART algorithm.

Some relevant research activities are oriented on the use of mathematical method called rough sets to improve processing phase of medical data. Experiments described in [7] were performed over the UCI machine learning database containing 155 records characterized by 20 attributes as age, sex, steroid, antivirals, etc. The source dataset was divided into training and testing set in the ratio of 50 to 50 %. The maximal obtained diagnostic accuracy was 94 % using LEM2 algorithm to generate classification rules. Similar approach is described in [8], the authors proposed a new hybrid medical decision support system based on rough set and extreme learning machine (ELM) for the diagnosis of hepatitis disease. At the first level, redundant factors were removed through rough set methods and then classification procedure was implemented through ELM. With the same dataset, the obtained accuracy ranged around 100 % depending on the ratio of the records distribution between training and testing set.

The same UCI machine learning dataset containing 155 records and 20 attributes was used by Roslina and Noraziah [9] to evaluate a combination of Support Vector Machine and Wrapper methods for hepatitis diagnostics. Wrapper method was used to remove noise features before classification. In the case of classifier using SVM without feature selection, the obtained accuracy was around 72 %. On the other hand, use of feature selection method (only 10 attributes from 20 previous) improves the accuracy to 74 %.

Three classification methods, i.e. decision trees, neural networks and Naïve Bayes were used in [10] to extract hidden knowledge relevant for HCV diagnosis from historical data. Authors used dataset containing 859 records characterized by 13 medical attributes from national liver diseases referral center in Egypt. 602 records were used for training and 257 for testing purposes (random selection). The most accurate classification model 95 % was generated by decision trees methods. Also, authors investigated the impact and relationship between input medical attributes and target diagnosis in the form of rules.

Above described related studies show that there have been various attempts to approach hepatitis patient records by machine learning methods. Most of them used UCI repository, which in fact focuses on the survival of patients with hepatitis and can therefore hardly be compared with our goal of early prediction of hepatitis B and C patients taking into account the associated costs of diagnosis as well as various treatments. We did not find any similar study in the literature.

1.3 Methods

Selection of appropriate methods depends on the specified objectives, which in our case were exploratory data analysis in order to identify interesting hidden relationships or knowledge in the processed dataset and predictive mining, taking into account

associated costs (or future expected costs) represented by classification models with the aim to improve the overall effectiveness of the hepatitis diagnostics.

We used the Chi-squared Automatic Interaction Detector (CHAID) which is one of the oldest tree classification methods originally proposed by Kass [11]. Despite the fact that in related work (see Sect. 1.2) are mostly used algorithms like C5.0 and CART, we decided for CHAID because obtained output is highly visual and easy to interpret. It usually creates simple conservative trees that give good results on the test set. This algorithm constructs non-binary trees and uses the Chi-square test to determine the best next split at each step. CHAID is recursive partitioning method.

The optimal cut-off points of a diagnostic test are defined as the points at which the expected utility of this test is maximized [12]. We opted for cost-benefit method, which is based on calculating of the ROC curve slope at the optimal cut-off points [13, 14]. It weighs the relative costs of the different predictions in the diagnosis by (1), where:

- C_{FP} – costs of false positive records.
- C_{TP} – costs of true positive records.
- C_{TN} – costs of true negative records.
- C_{FN} – costs of false negative records.

$$S = \frac{1-p}{p} CR \tag{1}$$

The abbreviation CR in (1) means cost ratio, which is given as:

$$CR = \frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}} \tag{2}$$

Coefficient p is disease prevalence. R package OptimalCutpoints was used for this purpose [15].

2 Exploratory Data Analysis

In our experiments we used data from the study focused on the prevalence of chronic viral hepatitis B and C in Slovakia. This study included hospitalized patients older than 15 years, except for those with a known diagnosis of chronic hepatitis B or C and for which it was possible to ensure the collection of blood. A total of 4 598 patients were examined after that each filled in the anonymous questionnaire consisting of 3 sections (see Table 1). First 5 questions monitored demographic data of the patient. The next part concerned on the collection of epidemiological data, where patients were asked at main risk factors causing HB and HC.

Last section included primary patient's diagnosis, the blood test results from examination of basic markers HBsAg, anti-HCV and level of ALT. Final set for processing and modeling phase contained 39 attributes; distribution of female and male patients was almost similar (52:48 in %).

Table 1. Description of all variables included in experiments

Variable code	Variable description (incl. possible values for nominal attributes)
Demographic part	
County town	town, where the patient was hospitalized
Month	month, when the patient was hospitalized
District	district, where the patient comes from
Age	age of the patient
Sex	man/woman
Education	elementary/secondary/higher
Risk factors	
Overcoming the hepatitis in the past	A/B/C/do not know what type/none
If yes, what type of hepatitis?	acute/chronic/unspecified
If yes, treated?	treated/untreated
Contact with hepatitis	A/B/C/do not know what type/none
If yes, what type of contact?	family/work/sexual partner/friend/other
Hepathopathy diagnosed in the past	yes/no
If yes, since when?	the year of the identification
Surgical operation	yes/no
If yes, number?	the number of operations
If yes, when?	year(s) when the operation was
Biopsy, endoscopy	yes, no
If yes, number?	the number of procedures
If yes, when?	year(s) when the procedure was
Dialysis	yes/no
If yes, since when?	the year of the identification
Blood transfusion	yes/no
If yes, number?	the number of transfusions
If yes, when?	year(s) when the transfusion was received
Tattoo	yes/no
If yes, number?	the number of tattoos
If yes, when?	year(s) when the tattoo was made
Piercing	yes/no
If yes, number?	the number of piercings
If yes, when?	year(s) when the piercing was made
Alcohol	abstainer/rarely/once a week/several times a week
Drugs - intravenous	never/repeatedly/only one experience
Sexual behavior	heterosexual/homosexual/bisexual
Number of sexual partners	the number of sexual partners
Blood test results	
Primary diagnosis	the type of diagnosis
ALT	liver diagnostic test (μkat/l)
HBsAg	hepatitis B surface antigen (yes/no)
Anti-HCV	antibodies against hepatitis C virus (yes/no)

The largest number of patients was examined in the Košice department (943), next Bratislava (737) and the minimum number was from Martin (350). Numbers from all other departments vary around 500 patients. Whereas the transmission of body fluids is one of the most important HBV transmission paths, we investigated the relationship between attribute *Number of sexual partners* and HBV positive diagnosis. At first, previous numeric attribute was transformed to binary one with separation boundary 6 partners. But the hypothesis about the potential of significant value of this attribute for the diagnosis of HBV was not confirmed. Next we investigated relationship between gender of patients and resulting diagnosis. In the case of HBV, number of male patients with a positive diagnosis is significantly higher than in case of female patients (see Fig. 1).

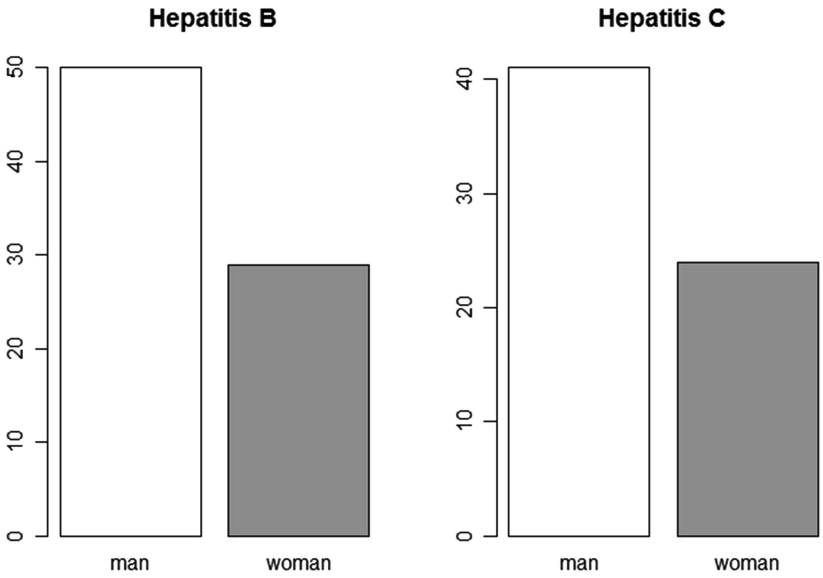


Fig. 1. Distribution of healthy and patients suffered from HB and HC through gender.

In the next step we wanted to find out whether there is some age group with the majority of ill patients. We applied the ChiMerge algorithm, which uses the χ^2 statistic to determine if the relative class frequencies of intervals are distinctly different or if they are similar enough to justify merging them into a single interval [16]. In the case of HBV, age was divided into four categories (see dashed horizontal lines on Fig. 2), where most of the patients with positive diagnosis were older than 43 years (30 positive to 2077 negative patients). However, most interesting is the category between 24 and 31 years, containing 22 positive and 588 negative patients.

In the case of HCV, the category between 27 and 46 was the one with the highest relative number of patients with positive diagnosis (see dashed horizontal lines on Fig. 3).

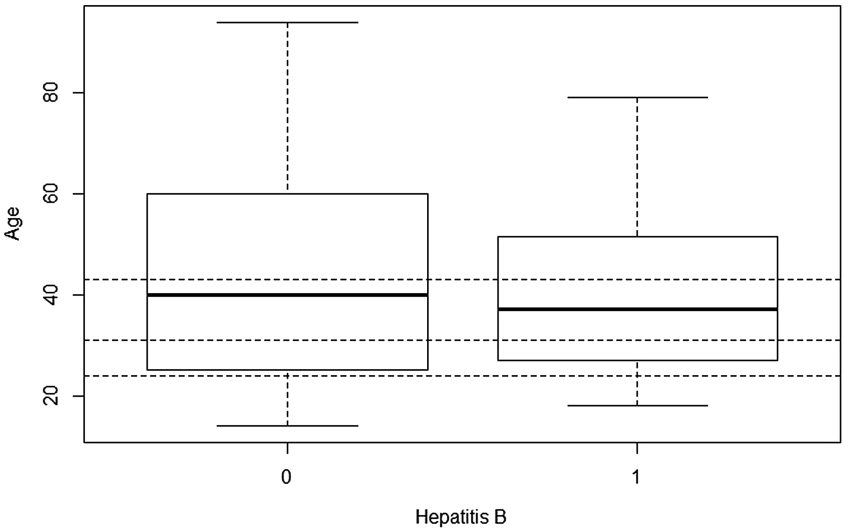


Fig. 2. Visualization of the relationship between patient age and the diagnosis HB.

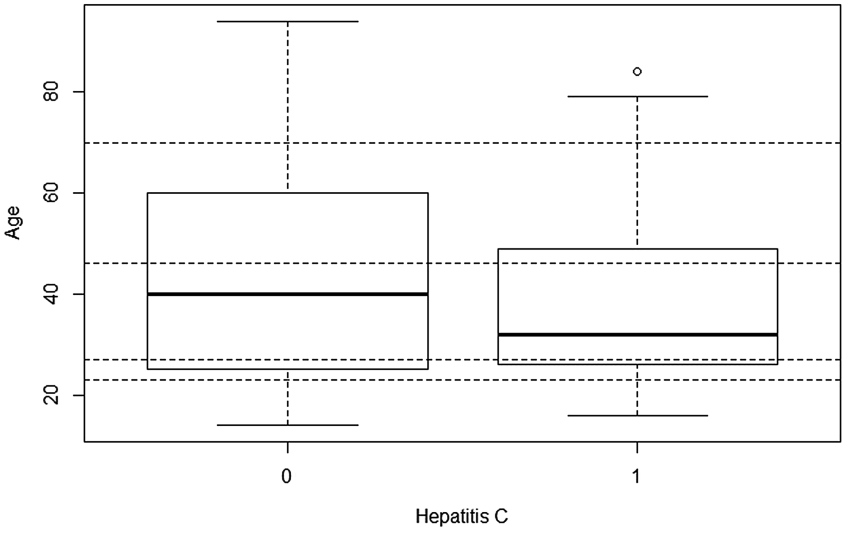


Fig. 3. Visualization of the relationship between patient age and the diagnosis HC.

Raw dataset contained many typos and spelling errors, so it was necessary to clean the data and delete some rows with unreliable information. This operation reduced number of records to 4 565 patients. As three questions ("overcoming the hepatitis in the past", "contact with hepatitis" and "if yes, what type of contact?") could have more than one answer, so we transformed them into binary attributes, i.e. in the case of multiple answer ("A" and "B"), the value of both binary attributes "A" and "B" was set

to 1. In case of questions type "if yes, when?" we always left only one (the first) year from answers, what allowed understand these attributes as numerical. Dataset in this stage was very imbalanced.

Since screening of blood transfusions for HCV had begun in 1992, we transformed the attribute "The year of blood transfusion" to new binary attribute, which followed this information.

Dataset contained only 79 patients with confirmed HBV and 65 patients with confirmed HCV. This fact required to use appropriate approaches in our performed experiments, e.g. experiments with the whole datasets, experiments with separate male and female patients, experiments with different ratios division into training and test set, experiments with random or stratified sampling, experiments with oversampling or subsampling dataset based on target attribute distribution. Following section presents the most interesting results from this extensive set of experiments.

3 Predictive Data Mining

Performed experiments were divided into several groups based on the partial objectives and in the following text we present only selected part of these experiments, the results of which were identified by participating domain expert as interesting.

3.1 HBV Decision Tree for Male Patients with Normal Level of ALT

Resulting decision tree generated by means of CHAID method is visualized on Fig. 4. Due to markedly unbalanced dataset (see Sect. 2) we used cost sensitive learning method with cost matrix from Table 2 (9€ cost for false positive mistake and 1400€ cost for false negative mistake). The presented model corresponds to the following decision rules:

1. *Overcoming the any hepatitis in the past* = yes AND *ALT* <= 0.56 THEN *hepatitis B* = negative. (Nodes: 0, 7 and 8)
2. *Overcoming the any hepatitis in the past* = yes AND *ALT* > 0.56 THEN *hepatitis B* = positive. (Nodes: 0, 7 and 9)
3. *Overcoming the any hepatitis in the past* = no AND *Age* <= 39 THEN *hepatitis B* = negative. (Nodes: 0, 1 and 2)
4. *Overcoming the any hepatitis in the past* = no AND 39 < *Age* <= 56 AND *Surgical operation* = yes THEN *hepatitis B* = positive. (Nodes: 0, 1, 3 and 5)
5. *Overcoming the any hepatitis in the past* = no AND 39 < *Age* <= 56 AND *Surgical operation* = no THEN *hepatitis B* = negative. (Nodes: 0, 1, 3 and 4)
6. *Overcoming the any hepatitis in the past* = no AND *Age* > 56 THEN *hepatitis B* = negative. (Nodes: 0, 1, 6)

Classification accuracy for this model was around 88 % and just for curiosity, the same experiment for female patients resulted into classification model with more than 90 % accuracy, also for HBV diagnosis. One of the medically interesting results (rule Nr. 2 above) is that not only elevated values of ALT but even those between 0,56 and

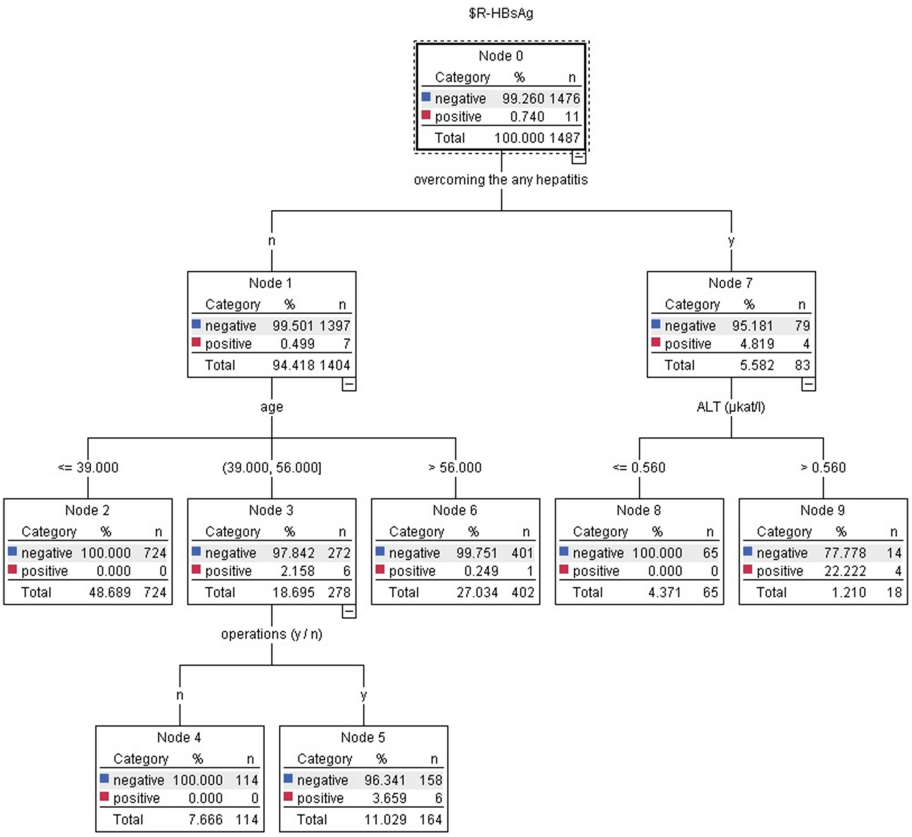


Fig. 4. Generated CHAID decision tree for male patients.

Table 2. Used cost matrix.

	Predicted negative	Predictive positive
Actual negative	0	7€ (for HCV) or 9€ (for HBV)
Actual positive	1 400€ (treatment of cirrhosis)/740€ (F3 stage)	0

0,8 µkat/l can be associated with unrecognized HBV infection. People who overcame any hepatitis and have ALT below 0,56 µkat/l, do not need to be screened for HBV (rule Nr. 1). In patients, who did not overcome any hepatitis, the highest probability to be HBsAg positive is if they are between 39 and 56 (rules Nr. 4 to 5).

3.2 Optimal Cut-off Points for Selected Attribute

These experiments were oriented to find the optimal cut-off points within cost-benefit method. Based on previous experiments, we selected attribute *ALT* whose typical

cut-off points used in medical practice are 0.8 for men and 0.6 for women. We used following cost matrix to evaluate the effectiveness of the generated models (see Table 2). False positive records represented those patients who were identified by generated model as suffering from HBV or HCV, but in fact they were healthy people. This fact was subsequently confirmed by the primary test which approximate price is 9 € or 7€, resp. False negative records represented those patients whose model classified as healthy, but in reality they were infected with HBV or HCV. Their subsequent treatment would be more expensive. Therefore, for comparison we used two different cost values,q first 1 400€ (which corresponds to cost of F4 stage - cirrhosis treatment - for details see Sect. 1.1 above) and second, more optimistic value, 740€ (which corresponds to cost of F3 stage treatment - for details see also Sect. 1.1 above).

Table 3. New cut-off points for hepatitis B.

Treatment of cirrhosis cost		F3 stage cost	
New cut-off point for ALT (male patients)	0.59	New cut-off point for ALT (male patients)	0.59
New cut-off point for ALT (female patients)	0.24	New cut-off point for ALT (female patients)	3.94
Number of False positive records	2 758	Number of False positive records	1 128
Number of False negative records	6	Number of False negative records	20

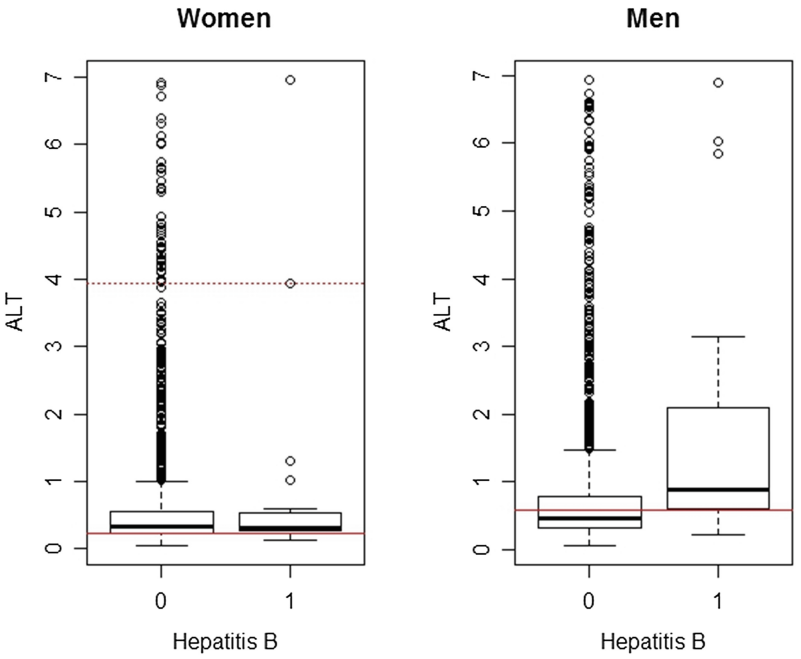


Fig. 5. Box plot of patients with normal ALT level according to the diagnosis HB (red line means new cut-off point with treatment cirrhosis cost, red dotted line with F3 stage cost) (Color figure online).

In the case of HBV, our experiments brought following results (see Table 3). It is interesting that the calculated cut-off points for male population are in both experiments the same, but there is a big difference in female population. Visualization of these results can be seen in Fig. 5.

If we set our baseline to total costs, which represent the case when all patients are tested for HBsAg (40 374€), the overall costs decreases to 33 222€ with calculation of cirrhosis treatment costs (F4 stage) and to 24 952€ with calculation of F3 stage costs. Total cost with actual (old) cut-off points of ALT are 47 668€ with calculation of cirrhosis treatment costs (F4 stage) and 30 508€ with calculation of F3 stage costs.

The same results presentation we use also for HCV (see Table 4 and Fig. 6):

Table 4. New cut-off points for hepatitis C.

Treatment of cirrhosis cost		F3 stage cost	
New cut-off point for ALT (male patients)	0.18	New cut-off point for ALT (male patients)	0.67
New cut-off point for ALT (female patients)	0.2	New cut-off point for ALT (female patients)	1
Number of False positive records	4 155	Number of False positive records	1 217
Number of False negative records	1	Number of False negative records	22

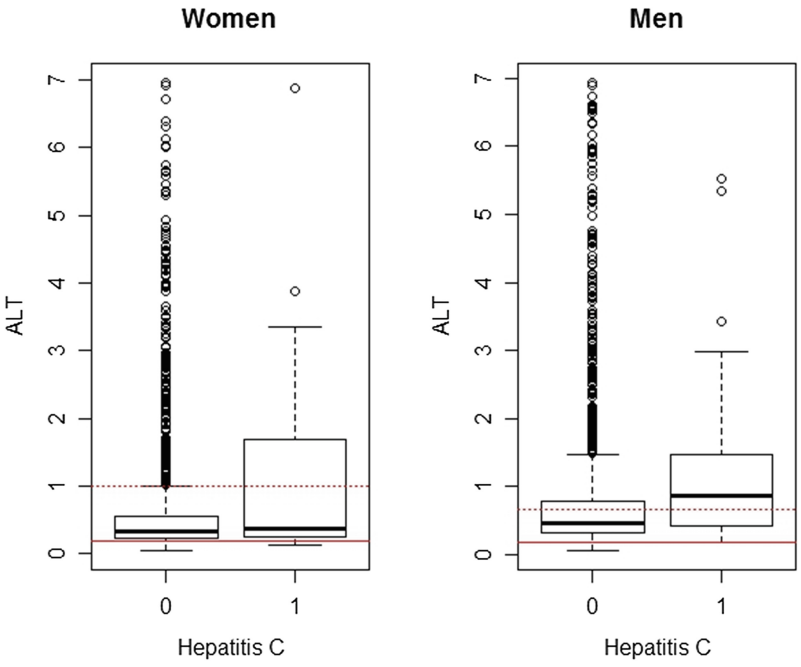


Fig. 6. Box plot of patients with normal ALT level according to the diagnosis HC (red line means new cut-off point with treatment cirrhosis cost, red dotted line with F3 stage cost) (Color figure online).

The overall costs of the Anti-HCV testing are 30 485€ with calculation of cirrhosis treatment costs and 24 799€ with calculation of F3 stage costs, which is slightly lower than the baseline for HCV (31 500€). Total costs with actual (old) cut-off points of ALT are 43 855€ with calculation of cirrhosis treatment costs and 27 355€ with calculation of F3 stage costs. This means that in case of HCV the calculation of new cut-off values did not bring any practical contribution.

4 Conclusion

Presented results show that machine learning models can be useful in medical practice to predict patients who should be screened for chronic hepatitis B or C. Especially extracted new cut-off points for HBV represent more sensitive diagnosis, i.e. they should lower the costs implied by late diagnosis of hepatitis B, which is associated with much higher treatment costs. The results of our study show that the probability of HBV infection is higher in patients with ALT ≥ 0,56 μkat/l, which could lead physicians to test patients for HBsAg antigen even with normal levels of ALT. This approach can early state diagnosis, which would significantly reduce the costs of their later treatment, and prevent further disease progression and worsening health of patients.

Acknowledgment. This publication is the result of the Project implementation: University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS: 26220220182, supported by the Research & Development Operational Programme funded by the ERDF (50 %); supported also by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/1147/12 (50 %).

References

1. Schréter, I., Kristian, P., Klement, C., Kohútová, D., Jarčuška, P., Maďarová, L., Avdičová, M., Máderová, E.: Prevalencia infekcie vírusom hepatitídy C v Slovenskej republike. Klin Mikrobiol Inf Lék **13**(2), 54–58 (2007)
2. Lee, T.A., Veenstra, D.L., Iloeje, U.H., Sullivan, S.D.: Cost of chronic hepatitis B infection in the United States. J. Clin. Gastroenterol. **38**, 144–147 (2004)
3. Leidner, A.J., Chesson, H.W., Xu, F., Ward, J.W., Spradling, P.R., Holmberg, S.D.: Cost-effectiveness of hepatitis C treatment for patients in early stages of liver disease. Hepatology **61**, 1860–1869 (2015). doi:10.1002/hep.27736
4. Chen, X., Ma, L., Chu, N., Hu, Y.: Diagnosis based on decision tree and discrimination analysis for chronic hepatitis b in TCM. In: Proceedings of 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW 2011), pp. 817–822. doi:10.1109/BIBMW.2011.6112478
5. Sartakht, J.S., Zangooei, M.H., Mozafari, K.: Hepatitis disease diagnosis using a novel hybrid method. Comput. Methods Programs Biomed. **108**(2), 570–579 (2011). Elsevier
6. Sathyadevi, G.: Application of CART algorithm in hepatitis disease diagnosis. In: Proceedings of IEEE Recent Trends in Information Technology (ICRTIT 2011), pp. 1283–1287 (2011)
7. Kanik, T.: Hepatitis disease diagnosis using Rough Set - modification of the pre-processing algorithm. Inf. Commun. Technol. **1**(1), 47–50 (2012). International Conference 2012

8. Kaya, Y., Uyar, M.: A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. Appl. Soft Comput. **13**(8), 3429–3438 (2013). Elsevier
9. Roslina, A.H., Noraziah, A.: Prediction of Hepatitis Prognosis using support vector machine and wrapper method. In: Proceeding of IEEE Fuzzy Systems and Knowledge Discovery (FSKD 2010), pp. 2209–2211 (2010)
10. Rarwan, A.A.A., Hafeez, T.E., Mamdouh, H.: An analysis of hepatitis C virus prediction using different data mining techniques. Int. J. Comput. Sci. Eng. Inf. Technol. Res. (IJCSEITR) **3**(4), 209–220 (2013)
11. Kass, G.V.: An exploratory technique for investigating large quantities of categorical data. Appl. Stat. **29**(2), 119–127 (1980)
12. Vranova, J., Horak, J., Kratka, K., Hendrichova, M., Kovarikova, K.: ROC analysis and the use of cost-benefit analysis for determination of the optimal cut-point. J. Czech Phys. **148**, 410–415 (2009)
13. McNeill, B.J., Keeler, E., Adelstein, S.J.: Primer on certain elements of medical decision making, with comments on analysis ROC. N. Engl. J. Med. **5**, 211–215 (1975)
14. Metz, C.E.: Basic principles of ROC analysis. Seminars Nucl. Med. **8**, 283–298 (1978)
15. Lopez-Raton, M., Rodriguez-Alvarez, M.X., Suarez, C.C., Sampedro, F.G.: OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. J. Stat. Softw. **61**(8), 1–36 (2014)
16. Kerber, R.: ChiMerge: discretization of numeric attributes. In: Proceedings of the Tenth National Conference on Artificial Intelligence, pp. 123–128 (1992)

Ant-Inspired Algorithms for Decision Tree Induction

An Evaluation on Biomedical Signals

Miroslav Bursa[1(✉)] and Lenka Lhotská[2]

[1] Czech Institute of Informatics, Robotics and Cybernetics,
Czech Technical University in Prague, Prague, Czech Republic
bursam@ciirc.cvut.cz
[2] Department of Cybernetics, Faculty of Electrical Engineering,
Czech Technical University in Prague, Prague, Czech Republic
lhotska@fel.cvut.cz

Abstract. In this paper we present an evaluation of ant-inspired method called *ACO_DTree* over biomedical data. The algorithm maintains and evolves a population of decision trees induced from data. The core of the algorithm is inspired by the Min-Max Ant System.

In order to increase the speed of the algorithm we have introduced a local optimization phase. The generalization ability has been improved using error based pruning of the solutions.

After parameter tuning, we have conducted experimental evaluation of the *ACO_DTree* method over the total of 32 different datasets versus 41 distinct classifiers. We conducted 10-fold crossvalidation and for each experiment obtained about 20 quantitative objective measures. The averaged and best-so-far values of the selected measures (precision, recall, f-measure, ...) have been statistically evaluated using Friedman test with Holm and Hochberg post-hoc procedures (on the levels of $\alpha = 0.05$ and $\alpha = 0.10$). The *ACO_DTree* algorithm performed significantly better ($\alpha = 0.05$) in 29 test cases for the averaged f-measure and in 14 cases for the best-so-far f-measure.

The best results have been obtained for various subsets of the UCI database and for the dataset combining cardiotocography data and data of myocardial infarction.

1 Introduction

In many industrial, business, healthcare and scientific areas we can see still growing use of computers and computational resources. Together with the boom of high-speed networks and increasing storage capacity of database clusters, data warehouses and cloud technologies, a huge amount of various data can be stored. Such data are often mixed from different source, containing different data types, unusual coding schemes, and seldom come without any errors (or noise). *Data mining* is not only an important scientific area, but also an important tool in industry, business and healthcare as the people making decision are becoming

M.E. Renda et al. (Eds.): ITBAM 2015, LNCS 9267, pp. 95–106, 2015.
DOI: 10.1007/978-3-319-22741-2_9

overwhelmed by the data. As an example we can mention the Human Genome Project, Hubble Space Telescope and the Human Brain mapping Project. A nice overview of swarm intelligence for data mining can be found in [21].

Automatic construction of simple rules in the form of decision trees has been attempted virtually in all disciplines in which data exploration methods have been developed. It has been traditionally developed in the fields of statistics, engineering (pattern recognition) and decision theory (decision table programming). Recently renewed interest has been generated by research in artificial intelligence (machine learning) and neurosciences (neural networks) [22].

Multiple metaheuristics have been used for inducing decision trees. However, to our best knowledge we did found only a few successful attempts to use ant colony optimization method for the induction of decision trees (see Sect. 2.2 for more information).

2 Background Information

2.1 Ant Colony Optimization

The high number of individuals and the decentralized approach to task coordination in many ant species shows that ant colonies show high degree of parallelism, self-organization and fault tolerance. In studying these paradigms, there is a high chance to discover inspiration concepts for many successful metaheuristics.

Simple Ant Colony Optimization. The first ant colony optimization algorithm was the ant system [12]. This system has been improved and the first optimization algorithm available was the simple ACO (SACO) [10]. The SACO in fact implements the double bridge experiment [9]. It has been used for the TSP problem, where each edge is assigned a cost value (length) L. For the SACO each edge is also assigned a pheromone value τ_{ij} that is updated during the run. The pheromone value is used for inducing the potential solutions.

It has been observed that initial experiments on the binary bridge have rapidly converged to a solution and the ants did not explore alternative paths. Therefore an pheromone evaporation at the end of each iteration has been introduced.

Ant System. Ant System (AS) [11] improves SACO by changing the transition probability between nodes to include heuristic information and by the inclusion of a tabu list (thus adding a memory to the algorithm). The AS in fact balances the exploration and exploitation tradeoff that is controlled by the parameters α and β. If $\alpha = 0$, no pheromone is used, when $\beta = 0$, the algorithm degrades to SACO.

A different formulation of the transition probability has been defined in [20] where the α parameter defines a relative importance of the pheromone and removes the β parameter.

Ant Colony System. The Ant Colony System [13] has been developed as an improvement to performance of the AS. It uses a different transition rule, different pheromone update rule, local pheromone update and a candidate list to promote specific nodes.

The pseudorandom-proportional action rule is defined with a random number $r \sim U(0,1)$ and an user-specified parameter $r_0 \in [0,1]$. The r_0 parameter balances exploration ($r > r_0$) and exploitation ($r \leq r_0$).

In case of ACS, only the globally best ant is allowed to reinforce pheromone concentrations on the links (cf. AS). The authors implemented two methods of selecting the best path: (1) *iteration-best* where the pheromone is updated according to the best solution found in current iteration or (2) *global-best* where the pheromone is updated according to the best solution found ever.

Max-Min Ant System. It has been discovered that the AS prematurely stagnates for complex problems. All the ants followed exactly the same path and there was a very little exploration and too rapid exploitation of the highest pheromone concentrations. Therefore the Max-Min Ant System (MMAS) has been developed [28]. The main difference between AS and MMAS is that there is a limitation on the pheromone intensities, $\tau_{ij} \in [\tau_{min}, \tau_{max}]$ where the τ_{min} and τ_{max} are problem dependent static parameters. A positive lower bound τ_{min} facilitates better exploration as all the links have positive probability of being selected into solution. Additionally, only the best ant may reinforce pheromone. The firsts version of MMAS focused on the iteration-best path, later versions used different strategies (using combination of both strategies, reinitialization in case of stagnation, etc.).

We have tried all the ant metaheuristics mentioned above and finally, ended with a modified version of MMAS algorithm, including adaptive change of pheromone deposit and evaporation rate and the elitism.

More variants can be found in relevant literature, i.e. in the series of [14].

2.2 Decision Trees

Decision tree induction algorithms are used in many areas, such as astronomy, character recognition, finance, marketing, medicine, natural language processing, software engineering, etc. ([1]). The *decision trees* are hierarchical, sequential classification structures that recursively partition the set of observations (data). Each node of the decision tree is either a test that partitions the data, or a decision (classification) about the object. In this paper the decision trees are constructed automatically from the data.

Automated construction of decision trees (or decision rules) has been performed in many areas where data exploration methods have been developed.

The Problem of Construction. The construction of optimal decision trees (in the sense of minimal number of tests performed) is a NP-Complete problem ([17]). Authors [8] proved that even finding the root node in an optimal strategy is a NP-hard problem.

The algorithms for decision trees induction using ant colonies can be found in recent publications only, such as [24].

A comprehensive overview of decision trees can be found in [27].

Use of Metaheuristics for Tree Induction. The CART algorithm [5] is fully deterministic and it is prone to be stuck in local optima. The problem of traditional decision tree constructing algorithms is that it usually uses optimal split for a node only, but does not consider a group of data. Therefore a heuristic approach is suitable and provides more general results.

In 1990, Koza [18] introduced genetic programming for induction of decision trees. In 1993 [16] introduced an algorithm called SADT that uses a simulated annealing process to find oblique splits at decision node. Later in 1997, [22] introduced an OC1 system. It does not use the brute-force approach of SADT, but uses a random search only to improve the existing solution. Therefore it may be regarded as a hybridization of the CART and SADT technique. The classical approaches are very sensitive to incomplete and noisy data. A situation that is very usual in medical environment.

Evolutionary techniques have also been successfully implemented, i.e. [19]. More information can be found in [26]. It usually uses a population of trees, selecting the best according to a certain fitness function.

The authors in [1] propose to distinguish between (1) evolutionary induction of decision trees and (2) evolutionary design/improvement of decision tree components. In (1) the individuals are decision trees, in (2) the individuals are components of decision tree classifiers. Another logical possibility is to use evolutionary algorithms to optimize the parameters of an algorithm. In out methodology we use the first approach, therefore the decision trees are sometimes referred to as *individuals*.

An adaptive algorithm for ant-inspired induction of decision trees (ACDT) [3] uses so called twoing (binary) decision criterion or other decision rules that are not able to cope with continuous attributes. The authors proved that the algorithm is comparable to the classical CART algorithm. The authors later presented a modification using inequality rule for coping with continuous attributes [4]. However, similar approach has already been used in our publication few years earlier [6, 7]. For comparison the authors have used only 11 datasets from UCI repository and compared with an Ant-Miner algorithm [25] and the modification using continuous attributes [23].

The second ant-colony inspired induction algorithm can be found in [24]. It uses a virtual start node and two types of edges: nominal or continuous. In case of continuous attributes a dynamic discretization is used (as in C4.5 or cAnt Miner rule induction algorithm (of the same authors)). The authors also use the same heuristics as in C4.5 algorithm (information gain ratio for generating threshold values and further dividing the training sets into subsets). The pheromone is used as an accurate feedback of the quality. Authors evaluate their algorithm on 22 UCI datasets and compare them to weka's J48 and SimpleCART algorithm.

3 The *ACO_DTree* Method Description

We have used population approach with elitism. The potential solutions (decision trees) represent individuals. The pseudocode of the algorithm can be described as follows. The trees are induced from an incidence matrix (representing a memory of the algorithm) containing pheromone values. In the initial phase, the pheromone matrix is initialized either with random or constant values and the first population of trees is induced. In the iterative phase the population is first optimized (decision values in tree nodes are optimized) and then the population is increased and the worst solutions are removed. According to the elitist ratio, the best solutions contribute back to the pheromone matrix. Pheromone evaporation and adaptive measures keep the population diversified while keeping the pheromone matrix from being saturated (premature convergence).

Algorithm 1. Pseudocode of the creation and training of the *ACO_DTree* classifier

```
initialize ACO pheromone matrix
create initial population of decision trees
for all NUM_ITERATIONS do
    update adaptive values that depend on the iteration number
    if PSO_TRAIN_ITERATIONS elapsed then optimize population using the PSO algorithm [Parallel]
    end if
    increase the population with new PSO trained individuals [Parallel]
    prune the population [Parallel]
    remove the worst individuals too keep the population count constant
    update pheromone (elitist mode)
    evaporatePheromone
    if stopping criteria reached then
        BREAK
    end if
end for
save resulting classifier and related data
```

Fitness Measures of the Decision Trees. For each algorithm (over a dataset) the following outputs are saved for each partial crossvalidation run and for the final result (i.e. for 10-fold crossvalidation, $10 + 1$ results are available). Only the most important are listed, the total count of measures saved is about 70:

- Classified/misclassified/unclassified instances (absolute and relative value)
- FN/FP/TN/TP (absolute and relative value)
- F-Measure
- Precision/Recall
- Area under ROC/PRC (AUROC/AUPRC)
- Confusion matrix
- Kappa statistic
- Various information scores (Kononenko & Bratko)
- Various SF measures (entropy gain, etc.)

4 Experimental Part

4.1 Methodology Overview

As we are comparing 73 different groups of classifiers over 31 datasets, we need to divide them into some natural clusters. The classifier groups are shown in the Table 1 (the *ACO_DTree* method is not listed as it is compared with all other classifiers). The data groups are shown in Table 2. The reason for such division is to obtain a detailed information about the proposed method behavior instead of displaying only the *All Classifiers* over *All datasets* result (displayed in the rightmost bottom cell of the evaluation tables) as it is usual in other publications.

We have used the 10-fold crossvalidation. The data were first randomized (shuffled) and then divided into 10-folds (using stratification).

Evaluation. There are many measures available for evaluation. In this paper evaluate the performance using the *precision* and *recall* measure. We can use (1) the average value (over 10 crossvalidation runs) or (2) maximal (BSF, best-so-far) value.

The statistical significance of the results will be assessed as follows. First we will use Friedman nonparametrical test to verify the hypothesis that mean values of the measures are equal (H_0). If the null hypothesis is rejected, we continue with post-hoc tests – Nemenyi test for comparing the classifiers against the others. If a statistically significant difference between performance of two classifiers is found, we proceed with Holm and Hochberg post-hoc tests. If we cannot reject the H_0 at $\alpha = 0.05$ we try at $\alpha = 0.1$.

4.2 Classifiers

The classifier groups are shown in the Table 1. The *Distinct instances* column shows a number of distinct classifiers implemented as a separate class. The *Total instances* contains also variations containing different strategies (i.e. different kernel function, etc.). The grouping is according to the WEKA datamining software Java classes.

4.3 Datasets

We have divided the datasets into the following groups for detailed evaluation: See Table 2.

- **UCI** – Contains datasets from the UCI database. This is a selected part of UCI database that is used in many articles from the domain of artificial intelligence. It contains 23 datasets in total.
- **ECG** – Contains MIT and AHA ECG databases, in total 6 datasets. These datasets are prepended with the `MIT_` or `AHA_` prefix.
- **CTG + MI** – Contains data of myocardial infarction (2 datasets) and cardiotocography data (two datasets).

Table 1. Overview of the classifiers grouping for evaluation. Note that the *ACO_DTree* method is included in each group and is not listed explicitly.

Classifier	Distinct instances	Total instances
NaïveBayes	4	6
Functions	8	15
Lazy	4	15
Misc	4	8
Rules	11	12
Trees	10	16
Total	**41**	**72**

- **MIT** – Contains selected signals of the MIT ECG database together with the whole MIT database (total of 4 datasets). Does not contain the AHA ECG database. These datasets are prepended with the `MIT_` prefix.
- **All** – Includes all datasets available. In total it is 32 datasets.

Table 2. Overview of the dataset grouping for evaluation. The groups are not exclusive, overlapping occurs. Therefore the total sum is not equal to the sum of all the subgroups.

Classifier	No. datasets
ECG	6
CTG and MI	4
ECG_MIT	3
All	16
Total	**32**

5 Results

This section summarizes the statistical evaluation of the experiments. Only the final results are included in this part (obtained after parameter tuning).

We have evaluated the average (mean) and BSF measures from 10-fold cross-validation. In order to present the results in compact form, the following symbols are used in the result tables:

- The $\emptyset$ symbol means that the Friedman test did not reject the H_0 hypothesis[1].
- The symbols $\underset{05}{R}$ and $\underset{10}{R}$ mean that the Friedman test rejected the H_0 hypothesis on the level of α equal to 0.05 or 0.1 respectively. To save space we do not use decimals.

[1] The H_0 hypothesis states that all the classifiers performed the same.

– The symbols $\underset{05}{\surd}$ and $\underset{10}{\surd}$ indicate a significantly better results for the *ACO_DTree* method when compared to the other classifiers (by the Holm or Hochberg post-hoc test) on the level of α equal to 0.05 or 0.1 respectively.

Each cell of the table contains left and right part: The left side of each cell contains the statistical result symbol for an average value of the measure. BSF value of the measure is located in the right part of the cell (Table 3).

This kind of presentation is quite unusual, however it provides much more information than traditionally used approach when only the bottom-right cell is actually presented – the result for all classifiers vs. all datasets.

Note that each cell stands for the run of multiple classifiers (included the *ACO_DTree* method) over multiple datasets in order for the statistical tests to be valid.

Table 3. Statistical results for the *precision* measure (avg. and max) for the classifiers with optimized parameters.

Data\Classifier: precision	Bayes	Funct.	Lazy	Misc	Rules	Trees	All
ECG dataset	$\underset{10}{\surd}$ $\underset{}{\emptyset}$	$\underset{05}{R}$ $\underset{}{\emptyset}$	$\underset{05}{R}$ $\underset{05}{R}$	$\underset{}{\emptyset}$ $\underset{}{\emptyset}$	$\underset{05}{R}$ $\underset{}{\emptyset}$	$\underset{05}{R}$ $\underset{}{\emptyset}$	$\underset{05}{R}$ $\underset{}{\emptyset}$
CTG and MI	$\underset{05}{\surd}$ $\underset{}{\emptyset}$	$\underset{05}{\surd}$ $\underset{}{\emptyset}$	$\underset{05}{\surd}$ $\underset{05}{R}$	$\underset{05}{\surd}$ $\underset{}{\emptyset}$	$\underset{05}{\surd}$ $\underset{05}{R}$	$\underset{05}{R}$ $\underset{}{\emptyset}$	$\underset{05}{\surd}$ $\underset{05}{R}$
ECG MIT	$\underset{05}{R}$ $\underset{}{\emptyset}$	$\underset{05}{R}$ $\underset{}{\emptyset}$	$\underset{05}{R}$ $\underset{05}{R}$	$\underset{05}{R}$ $\underset{}{\emptyset}$	$\underset{05}{R}$ $\underset{}{\emptyset}$	$\underset{05}{R}$ $\underset{}{\emptyset}$	$\underset{05}{R}$ $\underset{}{\emptyset}$
All datasets	$\underset{05}{\surd}$ $\underset{05}{R}$	$\underset{05}{\surd}$ $\underset{}{\emptyset}$	$\underset{05}{\surd}$ $\underset{05}{\surd}$	$\underset{05}{\surd}$ $\underset{}{\emptyset}$	$\underset{05}{\surd}$ $\underset{05}{\surd}$	$\underset{05}{R}$ $\underset{}{\emptyset}$	$\underset{05}{\surd}$ $\underset{05}{\surd}$

5.1 Summary

Based on the statistical evaluation that is presented in Sect. 4 we can conclude that the *ACO_DTree* methods performs significantly better on the level of $\alpha = 0.05$ when compared over 32 different datasets and 41 distinct instances of classifiers. This overall result could be further divided for different measures and for various groupings of the datasets and/or classifiers (Table 4).

We conducted 10-fold crossvalidation and for each experiment saved and statistically evaluated about 20 quantitative objective measures. The averaged and best-so-far values of the selected measures have been statistically evaluated using Friedman test with Holm and Hochberg post-hoc procedures (on the levels of $\alpha = 0.05$ and $\alpha = 0.10$). The Nemenyi statistical test has also been conducted.

The *ACO_DTree* algorithm performed significantly better ($\alpha = 0.05$) in 29 test cases for the averaged f-measure and in 14 cases for the averaged precision measure. In case of the accuracy measure it was 13 and 15 cases respectively. The best results have been obtained for various subsets of the UCI database and for the dataset combining cardiotocography data and for the data of myocardial infarction.

Table 4. Statistical results for the *recall* measure (avg. and max) for the classifiers with optimized parameters.

Data\Classifier: recall	Bayes	Funct.	Lazy	Misc	Rules	Trees	All
ECG dataset	R_{05} $\emptyset_{05}$	R_{05} $\emptyset_{05}$	R_{05} $\emptyset_{05}$	R_{05} $\emptyset_{05}$	R_{05} $\emptyset_{05}$	R_{05} $\emptyset_{05}$	R_{05} $\emptyset_{05}$
CTG and MI	$\surd_{05}$ $\emptyset_{05}$	$\surd_{05}$ $\emptyset_{05}$	$\surd_{05}$ $\emptyset_{05}$	$\surd_{05}$ $\emptyset_{05}$	$\surd_{05}$ R_{10}	R_{05} $\emptyset_{05}$	$\surd_{05}$ $\emptyset_{05}$
ECG MIT	R_{05} $\emptyset_{05}$	R_{05} $\emptyset_{05}$	R_{05} $\emptyset_{05}$	R_{05} $\emptyset_{05}$	R_{05} $\emptyset_{05}$	R_{05} $\emptyset_{05}$	R_{05} $\emptyset_{05}$
All datasets	$\surd_{05}$ $\surd_{05}$	$\surd_{05}$ $\emptyset_{05}$	$\surd_{05}$ $\emptyset_{05}$	$\surd_{05}$ $\emptyset_{05}$	$\surd_{05}$ $\emptyset_{05}$	R_{05} $\emptyset_{05}$	$\surd_{05}$ R_{05}

6 Conclusion and Discussion

We have proposed, implemented and experimentally evaluated an ant-colony inspired method for induction of decision trees. We have also identified the areas where the *ACO_DTree* algorithm performed significantly better when compared to other state-of-the art classifier implementations. In addition to the accuracy measure we have evaluated also other quantitative measures. The advantage of the proposed method lies in the robustness achieved by replicating the self-organization behavior of ants that has been incorporated into an algorithm for induction of decision trees.

The algorithm uses randomization of the axis-parallel tree to improve the final solution. We have introduced a local search phase that leads to faster convergence of the algorithm. The local phase can be either random (jittering) or using another nature-inspired approach (PSO).

The algorithm has shown competitive results when compared to other classifiers. For the percent of correctly classified results (an average value) it has been statistically better in 4 cases for the whole UCI database and once for the ECG MIT and (CTG and MI) dataset on significance level $\alpha = 0.05$ and in one case for the whole ECG dataset ($\alpha = 0.10$).

More than 12400 experiments over 73 classifiers and 32 datasets has been run in the final evaluation phase.

6.1 Discussion

We are aware that in the branch of classifiers and decision trees, it is a common solution to use only the classifier accuracy for evaluation. In case of biomedical signals it is important to measure inter- and intra- patient results, together with measures used in medicine, such as SE and SP. In this project we have obtained about 20 different quantitative measures that can be used for evaluation. In addition, apart from the mean average value we can evaluate the best/worst-so-far solution found and even the standard deviation. The question is how this measures and their features should contribute to the final evaluation.

6.2 Future Work

It would be interesting to study (and extend) the algorithm to use hypercurves for splitting as opposed to axis-parallel splitting. The problem is that the task of finding such trees is much more complicated and the readability and comprehensibility of such tree is reduced. It is desirable to use sophisticated splits only when the extra complexity of the split is compensated by the contribution to tree quality.

Acknowledgment. This research project has been supported by the project number NT11124-6/2010 "Cardiotocography evaluation by means of artificial intelligence" of the Ministry of Health Care.

The research is supported by the project No. 15-31398A Features of Electromechanical Dyssynchrony that Predict Effect of Cardiac Resynchronization Therapy of the Agency for Health Care Research of the Czech Republic.

And I would like to acknowledge the UCI repository [15] and the relevant donors and creators of the datasets. This breast cancer data (Parp_BRLj) was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklic for providing the data. The blood transfusion data (UCI_bloodt) has been provided by Prof. I-Cheng Yeh [29]. The Breast Cancer Wisconsin dataset (Parp_BRWis) was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg [2]. CTG data originate from Faculty Hospital in Brno, Czech Republic. The CTG database is freely available online [30].

References

1. Barros, R., Basgalupp, M., de Carvalho, A., Freitas, A.: A survey of evolutionary algorithms for decision-tree induction. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **42**(3), 291–312 (2012)
2. Bennett, K., Mangasarian, O.: Robust linear programming discrimination of two linearly inseparable sets. Optim. Methods Softw. **1**(1), 23–34 (1992)
3. Boryczka, U., Kozak, J.: Ant colony decision trees – a new method for constructing decision trees based on ant colony optimization. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part I. LNCS, vol. 6421, pp. 373–382. Springer, Heidelberg (2010)
4. Boryczka, U., Kozak, J.: An adaptive discretization in the ACDT algorithm for continuous attributes. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part II. LNCS, vol. 6923, pp. 475–484. Springer, Heidelberg (2011)
5. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth International Group, Belmont (1984)
6. Bursa, M., Lhotska, L.: Automated classification tree evolution through hybrid metaheuristics. In: Corchado, E., Corchado, J.M., Abraham, A. (eds.) Innovations in Hybrid Intelligent Systems. AISC, vol. 44, pp. 191–198. Springer, Heidelberg (2007)
7. Bursa, M., Lhotska, L., Macas, M.: Hybridized swarm metaheuristics for evolutionary random forest generation. In: Proceedings of the 7th International Conference on Hybrid Intelligent Systems 2007, IEEE CSP, pp. 150–155 (2007)

8. Cox, L.A., Quiu, Y., Kuehner, W.: Heuristic least-cost computation of discrete classification functions with uncertain argument values. Ann. Oper. Res. **21**(1), 1–30 (1989)
9. Deneubourg, J.L., Aron, S., Goss, S., Pasteels, J.M.: The self-organizing exploratory pattern of the argentine ant. J. Insect Behav. **3**, 159–168 (1990)
10. Dorigo, M., Di Caro, G.D.: New ideas in optimization. In: The Ant Colony Optimization Meta-Heuristic, pp. 11–32. McGraw-Hill (1999)
11. Dorigo, M., Maniezzo, V., Colorni, A.: Ant system: optimization by a colony of cooperating agents. IEEE Trans. Syst. Man Cybern. B **26**(1), 29–41 (1996)
12. Dorigo, M.: Optimization, learning and natural algorithms. Ph.D. thesis, Dipartimento di Elettronica, Politecnico di Milano, Italy (1992)
13. Dorigo, M., Gambardella, L.M.: Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Trans. Evol. Comput. **1**(1), 53–66 (1997)
14. Engelbrecht, A.P.: Computational Intelligence: An Introduction, 2nd edn. Wiley, New York (2007)
15. Frank, A., Asuncion, A.: UCI machine learning repository (2010). http://archive.ics.uci.edu/ml
16. Heath, D., Kasif, S., Salzberg, S.: Learning oblique decision trees. In: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence IJCAI (1993)
17. Hyafil, L., Rivest, R.L.: Constructing optimal binary decision trees is np-complete. Inf. Process. Lett. **5**(1), 15–17 (1976)
18. Koza, J.R.: Concept formation and decision tree induction using the genetic programming paradigm. In: Schwefel, H.-P., Männer, R. (eds.) Parallel Problem Solving from Nature. LNCS, vol. 496, pp. 124–128. Springer, Heidelberg (1991)
19. Llor, X., Garrell, J.M.: Automatic classification and artificial life models. In: Proceedings of Learning Workshop (2000)
20. Maniezzo, V., Colorni, A.: The ant system applied to the quadratic assignment problem. IEEE Trans. Knowl. Data Eng. **11**(5), 769–778 (1999)
21. Martens, D., Baesens, B., Fawcett, T.: Editorial survey: swarm intelligence for data mining. Mach. Learn. **82**, 1–42 (2011)
22. Murthy, K.: On Growing Better Decision Trees from Data. Ph.D. thesis, Johns Hopkins University, Baltimore, MD (1997)
23. Otero, F., Freitas, A., Johnson, C.: cAnt-Miner: an ant colony classificationalgorithm to cope with continous attributes. In: CIDM, pp. 225–231 (2009)
24. Otero, F.E.B., Freitas, A.A., Johnson, C.G.: Inducing decision trees with an ant colony optimization algorithm. Appl. Soft Comput. **12**(11), 3615–3626 (2012)
25. Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: Data mining with an ant colony optimization algorithm. IEEE Trans. Evol. Comput. **6**(4), 321–332 (2002)
26. Podgorelec, V., Kokol, P., Stiglic, B., Rozman, I.: Decision trees: an overview and their use in medicine. J. Med. Syst. **26**(5), 445–463 (2002)
27. Rokach, L., Maimon, O.: Top-down induction of decision trees classifiers - a survey. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **1**(11), 1–12 (2002)
28. Stutzle, T., Hoos, H.: Max-min ant system. Future Gener. Comput. Syst. **16**(8), 889–914 (2000)

29. Yeh, I.C., Yang, K.J., Ting, T.M.: Knowledge discovery on RFM model using Bernoulli sequence. Expert Syst. Appl. **36**(3), 5866–5871 (2009). http://dx.doi.org/10.1016/j.eswa.2008.07.018
30. Chudacek, V., Spilka, J., Bursa, M., Janku, P., Hruban, L., Huptych, M., Lhotska, L.: Open access intrapartum ctg database. BMC Pregnancy Childbirth **14**, 16 (2014)

Poster Session

Microsleep Classifier Using EOG Channel Recording: A Feasibility Study

Martin Holub(✉), Martina Šrutová, and Lenka Lhotská

Department of Cybernetics, Faculty of Electrical Engineering,
Czech Technical University in Prague, Prague, Czech Republic
{holubma6, srutomar, lhotska}@fel.cvut.cz

Abstract. The microsleeps (MS) cause many accidents and can have a huge social impact. Automated prediction or early detection of the MS states could help to monitor level of fatigue. An automated MS classifier based on the EOG signal is proposed. There were analysed 28 episodes of MS. We observed slow eye movements without rapid changes during MS episodes. An automated feature extraction and classification using EOG channels showed promising results (sensitivity 93 %, positive predictivity 57 %). To confirm the hypothesis it is crucial to extend the study and to analyse larger amount of MS data.

Keywords: Microsleep · Electrooculogram · Automatic detection · Classifier

1 Introduction

An automated detection of human behavioral microsleeps (MS) has been already studied from many perspectives. These states of extreme drowsiness can take from several tenths of a second to several tens of seconds. MS is often manifested by absent or prolonged responses, or by the short suspension of performance. It seems that the people fall asleep momentarily during the MS [1]. These symptoms are accompanied by droopy eyes, slow eyelid-closure, and head nodding [1]. MS can be often observed in individuals, who are sleep-deprived, or who are performing extended monotonous tasks. It is typical for drivers, air traffic controllers or other control workers. The MS can be also related to stress, depression, various sleep disorders such as narcolepsy or sleep apnea, or to use of certain drugs. Early detection and prediction of these states using an automated MS detection can for example reduce accidents caused by the MS.

In the state of the art of MS automated analysis, there are different approaches to the detection of MS proposed. The simplest methods are mostly based on the speech tests, psychological tests and behavioral tests. The advanced methods use physiological signals such as EEG (electroencephalography) [2], EOG, fMRI (functional magnetic resonance imaging) [3] and EMG (electromyography) [4], or their combinations [5–8]. A study using ECG [9] was also presented.

The aim of our study is to test feasibility of electrooculography (EOG) signal classifier to automatically detect episodes of MS.

M.E. Renda et al. (Eds.): ITBAM 2015, LNCS 9267, pp. 109–113, 2015.
DOI: 10.1007/978-3-319-22741-2_10

2 Methods

2.1 Dataset

The measurements were taken on six healthy volunteers (4 men and 2 women, aged 23-35 years). These subjects were not aware of any mental, neurological or sleep disorders. They were non-smokers, and they were currently without any medication at least within a year. Measured subjects were asked to abstain of drinking alcohol and coffee during 7 days before measurement. They were asked to reduce sleep to 4 h per night during the two days. Activity of measured individuals was not limited during the measurement. EOG signals were acquired with pair electrodes placed on temporal bone and referenced to the center of frontal bone. The sampling frequency was 256 Hz. Records were scored by a neurologist to obtain sleep staging and MS scoring. Only records with MS were selected for analysis. The records of the length 3.5 - 10 min with MS were selected from each subject (in total MS 28 episodes).

2.2 EOG Signal Feature Extraction

The parametrization was made from the EOG signal in the time domain. A sliding window of 0.5 s width with the 0.2 s shift was applied to the signal E1 and the maximum ($E1_{max}$) value and the minimum ($E1_{min}$) value for the each position of the window were found. The parameter $EOG1_{amp}$, shown in Fig. 1, was counted as the *E1max-E1 min* in each window. $EOG1_{amp}$ was filtered by Savitzky-Golay FIR smoothing filter of the second order with the window length of 15 samples and then squared. This processing was independently applied to both signals E1 and E2, so that $EOG1_{amp}$ and $EOG2_{amp}$ were calculated. Then Ramp feature was calculated as follows: $R_{amp} = EOG1_{amp} * EOG2_{amp}$.

The feature R_{amp} was used as classifier input to detect MS from EOG. The following detection rule was used. R_{amp} was thresholded and the signal was declared as a MS in case it was below the threshold for longer than 3 s and shorter than 15 s. The EOG signal was analysed only during scored awake stages. There is used a fix

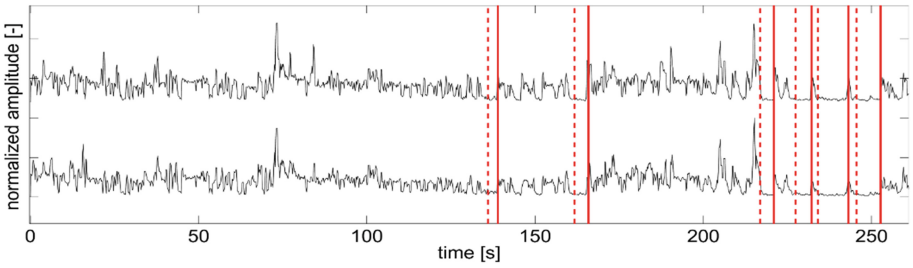


Fig. 1. The figure shows the parameter $EOG1_{amp}$ (top) and $EOG2_{amp}$ (bottom), the vertical interrupted lines mark the beginnings of the scored MS and vertical solid lines mark the ends of the scored MS.

threshold *THF*, which was set as the same value for all signals. The value was selected based on the analysis from the Precision-Recall (PR) curve.

3 Results

The amplitude of the EOG feature R_{amp} was observed as a very low during the episodes of MS. It distinguishes MS episodes from the other signal (Fig. 2).

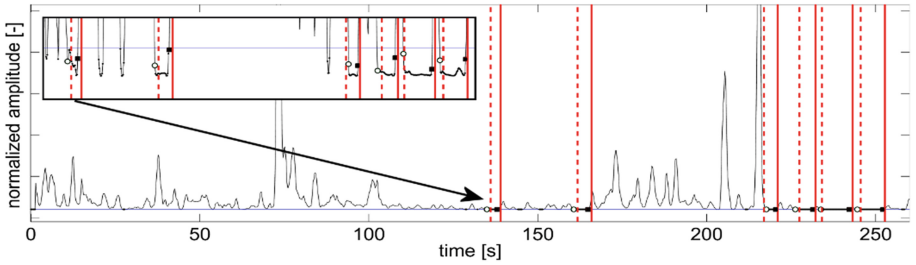


Fig. 2. The figure shows MS detection (*THF*-horizontal solid line) from the parameter R_{amp}, where white circles are the beginnings of the detected MS episodes and black squares are the ends of the automatically detected MS episodes. The vertical interrupted lines mark the beginnings of the scored MS and vertical solid lines mark the ends of scored MS. The detail of the MS detection is shown in the left up corner.

The curve PR (see Fig. 3) was calculated by the sweeping *THF* in the chosen range. The threshold, for which the inherent trade-off between precision and recall was the best, was chosen as a final threshold for the statistics. The best value of PR curve is the nearest to the right up corner (Fig. 3), where the precision and recall reach maximum values. Statistical values for each signal were calculated and presented in the Table 1.

Table 1. Statistical evaluation (TP – true positive, FP – false positive, FN – false negative).

Signal	Recall	Precision	TP	FP	FN
1	1.00	0.80	4	1	0
2	1.00	0.63	5	3	0
3	0.75	1.00	6	0	2
4	1.00	0.29	2	5	0
5	1.00	0.13	1	7	0
6	1.00	0.67	8	4	0
all	**0.93**	**0.57**	**26**	**20**	**2**

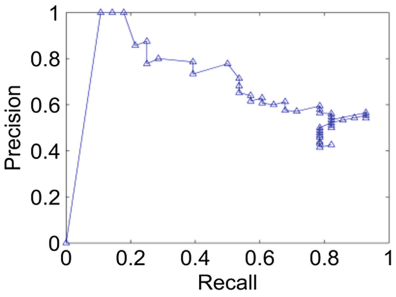


Fig. 3. Precision-Recall curve.

4 Discussions and Conclusions

The aim of presented feasibility study was to test whether EOG signal analysis is able to predict MS. Automated MS classifier based on EOG thresholding was presented. The episodes of MS correspond to the signal episodes without significant eye movements and fast changes. The feature R_{amp} was proposed with the motivation to find rapid changes in the eye movements and to distinguish specific high frequency waves of blink and arousal, which follows usually at the end of MS, from the flat parts without eye movements. This wave is highly differentiated in most of the MS episodes. The width of the sliding window was chosen 0.5 s. It was set to the value higher than maximum duration of physiological blinking, being 0.3-0.4 s. The shift of 0.2 s is used to improve time resolution of analysis. The detection reached the precision 57 % and recall 93 %. The detection results using such small amount of data can be considered promising and motivating to extend the study and to analyse large amount of MS data. This study shows, that two channels of EOG could be sufficient to classify MS episodes. In most of the published approaches [5, 7, 8] there are the EOG signals used in combination with other signals such as EEG or more channels of EOG. More data will be important to test proposed approach thoroughly. Since the EOG signal is relatively easy to record, the detection from EOG signal has very wide utilization not only in the sleep centres, but also for the safety of transport and industry.

Acknowledgment. This work has been supported by the project No.SGS13/203/OHK3/3T/13 of the Czech Technical University in Prague.

References

1. Peiris, M.T., Jones, R.D., Davidson, P.R., Carroll, G.J., Bones, P.J.: Frequent lapses of responsiveness during an extended visuomotor tracking task in non-sleep-deprived subjects. J. Sleep Res. **15**(3), 291–300 (2006)
2. Peiris, M.T., Jones, R.D., Davidson, P.R., Bones, P.J.: Detecting behavioral microsleeps from EEG power spectra. In: Engineering in Medicine and Biology Society, EMBS 2006, 28th Annual International Conference of the IEEE, pp. 5723–5726. IEEE (2006)
3. Poudel, G.R., Jones, R.D., Innes, C.R., Watts, R., Signal, T.L., Bones, P.J.: fMRI correlates of behavioural microsleeps during a continuous visuomotor task. In: Engineering in Medicine and Biology Society, EMBC 2009. Annual International Conference of the IEEE, pp. 2919–2922. IEEE, September 2009
4. Balasubramanian, V., Adalarasu, K.: EMG-based analysis of change in muscle activity during simulated driving. J. Bodywork Mov. Therapies **11**(2), 151–158 (2007)
5. Golz, M., Sommer, D., Krajewski, J., Trutschel, U., Edwards, D.: Microsleep episodes and related crashes during overnight driving simulations. In: Proceedings of the Sixth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design (2011)
6. Czisch, M., Wehrle, R., Harsay, H.A., Wetter, T.C., Holsboer, F., Sämann, P.G., Drummond, S.P.: On the need of objective vigilance monitoring: effects of sleep loss on target detection and task-negative activity using combined EEG/fMRI. Frontiers in neurology, vol. 3 (2012)

7. Leong, W.Y., Mandic, D.P., Golz, M., Sommer, D.: Blind extraction of microsleep events. In 15th International Conference on Digital Signal Processing, pp. 207–210. IEEE, July 2007
8. Rimini-Doering, M., Altmueller, T., Ladstaetter, U., Rossmeier, M.: Effects of lane departure warning on drowsy drivers' performance and state in a simulator. In: Proceedings of the third international driving symposium on human factors in driver assessment, training, and vehicle design, pp. 88–95, June 2005
9. Furman, G.D., Baharav, A., Cahan, C., Akselrod, S.: Early detection of falling asleep at the wheel: A heart rate variability approach. In: Computers in Cardiology, pp. 1109–1112. IEEE, September 2008

Author Index

Abelha, António 71
Almeida, Ana 71

Babič, František 81
Bae, Jang Hwang 49
Bhaskar, Pinaki 3
Bochicchio, Mario A. 25
Bursa, Miroslav 95
Buzzi, Marina 3

Carchiolo, Vincenza 16
Carvalho, André 71
Cunha, Adriana 71

Fröhlingsdorf, Christian 56

Geraci, Filippo 3
Greco, Marilena 25

Hales, Alan 56
Holub, Martin 109

Ishag, Ibrahim M. 49

Keech, Malcolm 56

Lhotská, Lenka 95, 109
Li, Peipei 37
Lobreglio, Giambattista 25
Longheu, Alessandro 16
Lu, Jing 56
Lukáčová, Alexandra 81

Machado, José 71
Malgeri, Michele 16
Mills-Mullett, Alex 56

Neves, João 71
Neves, José 71

Paralič, Ján 81
Paraličová, Zuzana 81
Park, Hyun Woo 49
Park, Soo Ho 49
Pellegrini, Marco 3

Rew, David 56
Ryu, Keun Ho 37, 49
Ryu, Kwang Sun 49

Šrutová, Martina 109

Vaira, Lucia 25
Vicente, Henrique 71

Wette, Christian 56

Zappatore, Marco 25