Carlo Combi
Yuval Shahar
Ameen Abu-Hanna (Eds.)

LNAI 5651

Artificial Intelligence in Medicine

12th Conference on Artificial Intelligence in Medicine, AIME 2009
Verona, Italy, July 2009, Proceedings

Springer

Lecture Notes in Artificial Intelligence 5651

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Carlo Combi Yuval Shahar
Ameen Abu-Hanna (Eds.)

Artificial Intelligence in Medicine

12th Conference on Artificial Intelligence
in Medicine, AIME 2009
Verona, Italy, July 18-22, 2009
Proceedings

Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Carlo Combi
University of Verona, Department of Computer Science
Ca' Vignal 2, strada le Grazie 15, 37134 Verona, Italy
E-mail: carlo.combi@univr.it

Yuval Shahar
Ben Gurion University of the Negev
Department of Information Systems Engineering
P.O. Box 653, Beer-Sheva 84105, Israel
E-mail: yshahar@bgumail.bgu.ac.il

Ameen Abu-Hanna
University of Amsterdam, Academic Medical Center
Department of Medical Informatics
Meibergdreef 15, 1105 AZ Amsterdam, The Netherlands
E-mail: a.abu-hanna@amc.uva.nl

Library of Congress Control Number: 2009930527

CR Subject Classification (1998): I.2, I.4, J.3, H.2.8, H.4, H.3

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-642-02975-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-02975-2 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12717941 06/3180 5 4 3 2 1 0

Preface

The European Society for Artificial Intelligence in Medicine (AIME) was established in 1986 following a very successful workshop held in Pavia, Italy, the year before. The principal aims of AIME are to foster fundamental and applied research in the application of artificial intelligence (AI) techniques to medical care and medical research, and to provide a forum at biennial conferences for discussing any progress made. For this reason the main activity of the society was the organization of a series of biennial conferences, held in Marseilles, France (1987), London, UK (1989), Maastricht, The Netherlands (1991), Munich, Germany (1993), Pavia, Italy (1995), Grenoble, France (1997), Aalborg, Denmark (1999), Cascais, Portugal (2001), Protaras, Cyprus (2003), Aberdeen, UK (2005), and Amsterdam, The Netherlands (2007). This volume contains the proceedings of AIME 2009, the 12th Conference on Articial Intelligence in Medicine, held in Verona, Italy, July 18-22, 2009.

The AIME 2009 goals were to present and consolidate the international state of the art of AI in biomedical research from the perspectives of theory, methodology, and application. The conference included two invited lectures, full and short papers, tutorials, workshops, and a doctoral consortium. In the conference announcement, authors were solicited to submit original contributions regarding the development of theory, techniques, and applications of AI in biomedicine, including the exploitation of AI approaches to molecular medicine and biomedical informatics and to healthcare organizational aspects. Authors of papers addressing theory were requested to describe the properties of novel AI methodologies potentially useful for solving biomedical problems. Authors of papers addressing techniques and methodologies were asked to describe the development or the extension of AI methods and their implementation, and to discuss the assumptions and limitations of the proposed methods and their novelty with respect to the state of the art. Authors of papers addressing systems were asked to describe the requirements, design, and implementation of new AI-inspired tools and systems, and discuss their applicability in the medical field. Finally, authors of application papers were asked to present the implementation of AI systems to solve significant medical problems, and to provide sufficient information to allow the evaluation of the practical benefits of such systems.

AIME 2009 received 140 abstract submissions, 126 thereof were eventually submitted as complete papers. Submissions came from 34 different countries, including 16 outside Europe. These numbers confirm the high relevance of AIME in attracting the interest of several research groups around the globe. All papers were carefully peer-reviewed by experts from the Program Committee with the support of additional reviewers. Each submission was reviewed by at least two, and on average three reviewers. Several submissions received four and even five reviews. The reviewers judged the quality and originality of the submitted pa-

pers, together with their relevance to the AIME conference. Seven criteria were taken into consideration in judging submissions: the reviewers' overall recommendation, the appropriateness, the technical correctness, the quality of presentation, the originality, the reviewers' detailed comments, and the reviewers' confidence in the subject area.

In a full-day skype-based virtual meeting held on April 2, 2009 and in several short discussions in subsequent days, a small committee consisting of the AIME 2009 Scientific Co-chair and Organizing Committee Chair, Carlo Combi, the AIME President, Ameen Abu-Hanna, and the AIME 2009 Scientific Co-chair, Yuval Shahar, made the final decisions regarding the AIME 2009 scientific program. As a result, 24 long papers (with an acceptance rate of about 19%) and 36 short papers were accepted. Each long paper was presented in a 25-minute oral presentation during the conference. Each short paper was presented in a 5-minute presentation and by a poster. The papers were organized according to their topics in the following main themes: (1) Temporal Reasoning and Temporal Data Mining; (2) Therapy Planning, Scheduling, and Guideline-Based Care; (3) Case-Based Reasoning; (4) Medical Imaging;(5) Knowledge-Based and Decision Support Systems; (6) Ontologies, Terminologies and Natural Language; (7) Data Mining, Machine Learning, Classification and Prediction; (8) Probabilistic Modeling and Reasoning; (9) Gene and Protein Data.

Moreover, AIME 2009 had the privilege of hosting two invited speakers: Carol Friedman, from Columbia University, New York, and Catherine Garbay, from the CNRS-Université de Grenoble. Carol Friedman gave a talk on "Discovering Novel Adverse Drug Events Using Natural Language Processing and Mining of the Electronic Health Record," and Catherine Garbay on "Computer Vision: A Plea for a Constructivist View."

Continuing a tradition started at AIME 2005, a doctoral consortium, organized on this occasion by Ameen Abu-Hanna, was held again this year and included a tutorial given by Ameen Abu-Hanna and Niels Peek on how to evaluate probabilistic models. A scientific panel consisting of Riccardo Bellazzi, Michel Dojat, Jim Hunter, Elpida Keravnou, Peter Lucas, Silvia Miksch, Niels Peek, Silvana Quaglini, Yuval Shahar, and Blaz Zupan discussed the contents of the students' doctoral theses.

As a novelty of AIME 2009, a significant number of full-day workshops were organized prior to the AIME 2009 main conference: the workshop entitled "KR4HC 2009, Knowledge Representation for Health-Care: Data, Processes and Guidelines," chaired by David Riaño (Universitat Rovira i Virgili, Spain) and Annette ten Teije (Vrije Universiteit Amsterdam, The Netherlands); the workshop "IDAMAP 2009, Intelligent Data Analysis in Biomedicine and Pharmacology," chaired by Tomaz Curk (University of Ljubljana, Slovenia), John H. Holmes (University of Pennsylvania School of Medicine, USA), and Lucia Sacchi (University of Pavia, Italy); the workshop "Personalization for e-Health 2009," chaired by Floriana Grasso (University of Liverpool, UK) and Cécile Paris (CSIRO, Sydney, Australia); and the workshop "Neuro-Sharing 2009, Sharing Data and Tools in Neuroimaging," chaired by Michel Dojat (Grenoble Insti-

tut des Neurosciences, France), Bernard Gibaud (VISAGES, France), and Gilles Kassel (MIS, Université d'Amiens, France). To complete this exciting set of scientific events, a full-day interactive tutorial was given by John H. Holmes, University of Pennsylvania, USA entitled "Introduction to Clinical Data Mining."

We would like to thank everyone who contributed to AIME 2009. First of all we would like to thank the authors of the papers submitted and the members of the Program Committee together with the additional reviewers. Thanks are also due to the invited speakers as well as to the organizers of the workshops and the tutorial and doctoral consortium. Final thanks go to the Organizing Committee, who managed all the work making this conference possible. The free EasyChair conference Web system (http://www.easychair.org/) was an important tool supporting us in the management of submissions, reviews, selection of accepted papers, and preparation of the overall material for the final proceedings. We would like to thank the University of Verona, the Department of Computer Science of the University of Verona, and the Faculty of Mathematical, Physical and Natural Sciences of the same university, which hosted and sponsored the conference. Finally, we thank the Springer team for helping us in the final preparation of this LNCS book.

May 2009

Carlo Combi
Yuval Shahar
Ameen Abu-Hanna

Organization

Scientific Co-chairs

Carlo Combi
Yuval Shahar

Program Committee

Ameen Abu-Hanna, The Netherlands
Klaus-Peter Adlassnig, Austria
Steen Andreassen, Denmark
Pedro Barahona, Portugal
Riccardo Bellazzi, Italy
Petr Berka, Czech Republic
Isabelle Bichindaritz, USA
Carlo Combi, Italy
Michel Dojat, France
Henrik Eriksson, Sweden
Catherine Garbay, France
Peter Haddawy, Thailand
Arie Hasman, The Netherlands
Reinhold Haux, Germany
John Holmes, USA
Werner Horn, Austria
Jim Hunter, UK
Hidde de Jong, France
Elpida Keravnou, Cyprus
Nada Lavrac, Slovenia
Xiaohui Liu, UK
Peter Lucas, The Netherlands
Roque Marin, Spain
Paola Mello, Italy
Gloria Menegaz, Italy
Silvia Miksch, Austria
Stefania Montani, Italy
Mark Musen, USA
Barbara Oliboni, Italy
Niels Peek, The Netherlands
Mor Peleg, Israel
Christian Popow, Austria
Silvana Quaglini, Italy
Marco Ramoni, USA
Stephen Rees, Denmark
Lucia Sacchi, Italy
Rainer Schmidt, Germany
Brigitte Seroussi, France
Yuval Shahar, Israel
Basilio Sierra, Spain
Costas Spyropoulos, Greece
Mario Stefanelli, Italy
Paolo Terenziani, Italy
Samson Tu, USA
Allan Tucker, UK
Frans Voorbraak, The Netherlands
Dongwen Wang, USA
Thomas Wetter, Germany
Blaz Zupan, Slovenia
Pierre Zweigenbaum, France

Organizing Committee

Carlo Combi
Mauro Gambini
Sara Migliorini
Aurora Miorelli
Barbara Oliboni
Roberto Posenato
Gabriele Pozzani

Additional Reviewers

Luca Anselma
Nicola Barbarini
Manuele Bicego
Alessio Bottrighi
Ernst Buchberger
Manuel Campos
Umberto Castellani
Hsun-Hsien Chang
Marco Cristani
Matteo Cristani
Fulvia Ferrazzi
Nivea Ferreira
Mauro Gambini
Albert Gatt
Marco Gavanelli
Adela Grando
Theresia Gschwandtner
Arjen Hommersom
Fernando Jimenez
Jimenez, Fernando
Jose M. Juarez
Katharina Kaiser
Anastasia Krithara
Cristiana Larizza
Giorgio Leonardi
Alberto Malovini
Michael McGeachie
Marco Montali
Robert Moskovitch
Igor Mozetic
Natasha Noy
Giulia Paggetti
Jose Palma
Silvia Panzarasa
Georgios Petasis
Francesca Pizzorni Ferrarese
Francois Portet
Gerasimos Potamianos
Gabriele Pozzani
Giuseppe Pozzi
Phattanapon Rhienmora
Fabrizio Riguzzi
Graeme Ritchie
Lina Maria Rojas
hspace*4mmBarahona
Andrea Roli
Stefania Rubrichi
Andreas Seyfang
Nigam Shah
Davide Sottara
Paolo Torroni
Harald Trost
George Tsatsaronis
Marina Velikova
Rattapoom Waranusast

Doctoral Consortium

Chair: Ameen Abu-Hanna, The Netherlands

Workshops

KR4HC 2009, Knowledge Representation for Health-Care: Data, Processes and Guidelines

Co-chairs: David Riaño (Universitat Rovira i Virgili, Spain) and Annette ten Teije (Vrije Universiteit Amsterdam, The Netherlands)

IDAMAP 2009, Intelligent Data Analysis in Biomedicine and Pharmacology

Co-chairs: Tomaz Curk (University of Ljubljana, Slovenia), John H. Holmes (University of Pennsylvania School of Medicine, USA), and Lucia Sacchi (University of Pavia, Italy)

Personalisation for e-Health 2009

Co-chairs: Floriana Grasso (University of Liverpool, UK) and Cécile Paris (CSIRO, Sydney, Australia)

Neuro-Sharing 2009, Sharing Data and Tools in Neuroimaging

Co-chairs: Michel Dojat (Grenoble Institut des Neurosciences, FR), Bernard Gibaud (VISAGES, FR), and Gilles Kassel (MIS, Université d'Amiens, FR)

Tutorial

Introduction to Clinical Data Mining

John H. Holmes, University of Pennsylvania School of Medicine, USA

Table of Contents

Invited Talks

1. Temporal Reasoning and Temporal Data Mining

2. Therapy Planning, Scheduling and Guideline-Based Care

3. Case-Based Reasoning

4. Medical Imaging

5. Knowledge-Based and Decision-Support Systems

6. Ontologies, Terminologies and Natural Language

7. Data Mining, Machine Learning, Classification and Prediction

8. Probabilistic Modeling and Reasoning

9. Gene and Protein Data

Discovering Novel Adverse Drug Events Using Natural Language Processing and Mining of the Electronic Health Record

Carol Friedman

Department of Biomedical Informatics, Columbia University, New York, US

Abstract. This talk presents an overview of our research in use of medical knowledge, natural language processing, the electronic health record, and statistical methods to automatically discover novel adverse drug events, which are serious problems world-wide.

Keywords: Pharmacovigilance, natural language processing, electronic health records, patient safety, adverse drug events.

1 Introduction

Natural language processing (NLP) of narrative electronic health records (EHR) is an enabling technology for the clinical domain because it is high throughput, and can automatically generate vast amounts of comprehensive and structured coded data that can be used by many different clinical applications to drastically improve healthcare. This technology provides an enormous opportunity for the biomedical research community because a wide range of coded clinical data can be made available for new and/or improved development of clinical applications. This talk will focus on a specific aspect of our research concerned with use of NLP, the EHR, biomedical knowledge, and statistical methods in order to detect undiscovered adverse drug events (ADEs), which are serious problems world-wide. In the United States alone, they result in more than 770,000 injuries and deaths each year [1], cost between $1.56 and $5.6 billion annually [2], and lead to increased hospital care [3]. In 1994, an estimated overall 2,216,000 hospitalized patients had serious ADEs, making ADEs approximately the fifth leading cause of death. Prior to approval, clinical trials study new drugs using relatively small test populations, which on average amount to less than 2,000 people for each trial [4], and, thus, the trials cannot account for the wide range of diverse conditions and populations that are necessary for a more thorough study. Therefore, continued post-market monitoring of drug reactions is necessary for patient safety [5]. Traditional approaches to pharmacovigilance rely on data from Spontaneous Reporting Systems (SPRSs) [6], but underreporting to these databases is widespread and erratic, delaying or preventing the detection of ADEs, and leaving many patients who are already taking the drugs at risk [7].

An alternative approach to use of SPRSs to detect ADE signals would be the use of data in EHRs, because they contain clinical information captured during the process of care and can be associated with enormous and diverse populations. There has been

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 1–5, 2009.

related work regarding the use of EHRs for pharmacovigilance. For example, a database, called the General Practice Research Database (GPRD), containing over 3.6 million active patients from primary care practices throughout the UK, was used to validate potential ADEs [8]. The GPRD contains anonymized longitudinal medical data, which is mainly coded and consists of co-prescription, co-morbidity, dosage details, off-label prescription, and patient demographics. In the United States, the Department of Veterans Affairs (VA) patient databases were used to explore the effect of beta-blockers for patients with chronic heart failure who were on warfarin [9]. Our work differs from related work in that we use comprehensive clinical information that occurs in the narrative reports whereas others primarily use manually coded data only. Additionally, our long-term aim is to discover novel ADE signals prospectively, in a timely fashion, and not to validate signals detected using SPRS data.

2 Overview

Our method for ADE discovery builds upon our work concerned with automatic acquisition of associations between clinical entities, such as disease-symptom and disease-drug associations. The underlying technique for generating associations is based on utilization of the MedLEE NLP system [10], which is used to process narrative discharge summaries so that the clinical events in the notes are structured and coded based on UMLS codes [11], and is followed by use of statistical methods. The statistical method that obtained disease-symptom associations for a selected set of diseases is described more fully by Cao and colleagues [12]. The coded output, consisting of MedLEE generated UMLS codes and corresponding modifiers was used to obtain diseases and symptoms. Certain UMLS codes were selected based on their semantic categories, and then some of the selected events were filtered out to remove those associated with modifiers, such as negations, past history, or family history. Associations were obtained using statistical methods on the selected and filtered data because disease symptom relationships are not usually explicitly stated in the patient record. To identify the associations, the chi-square (χ^2) statistic with a volume test adjustment was used. The numbers of occurrences of disease and symptom events that remained after filtering were recorded as well as the numbers of pairs that co-occurred in one discharge summary, and a modified χ^2 method was then used to determine disease-symptom associations based on appropriate cutoff thresholds. Evaluation was performed and the results demonstrated that the method was effective. The method was subsequently used to create an executable knowledge base, which is available online (http://www.dbmi.columbia.edu/xiw7002/DS-KB), of disease-symptom associations [13]. This knowledge base is also used by us to improve performance for ADE detection. Subsequent research [14] used similar methods to acquire knowledge regarding disease-drug associations using information in discharge summaries.

Aspects of the method for ADE discovery are described more fully by Wang and colleagues [15, 16], and we provide a summary here. It consists of five main steps, which are illustrated in Figure 1: 1) processing narrative reports to extract and code clinical data using the MedLEE NLP system, which enables reliable retrieval because the output obtains UMLS codes along with fine-grained modifiers representing certainty, temporal, anatomical, severity, quantitative, and qualitative types of

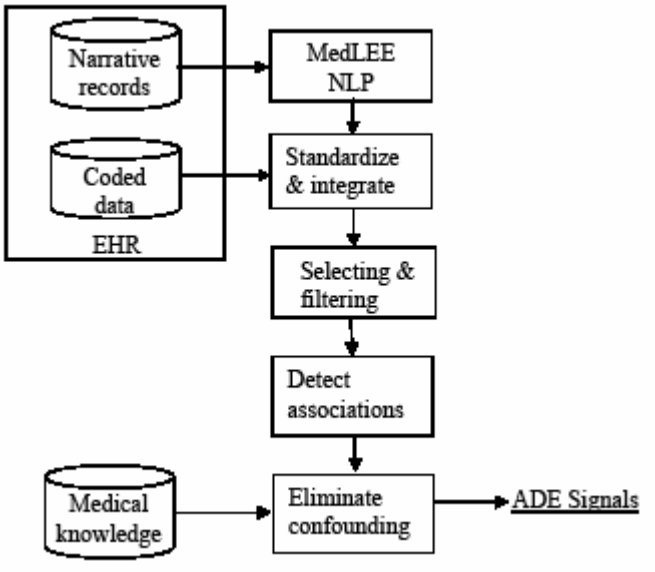


Fig. 1. Overview of ADE detection method

information that may modify the meaning of the primary UMLS concepts; 2) obtaining heterogeneous coded data in the EHR, such as laboratory tests and results, and standardizing the coded data so that it is consistent with the MedLEE output; 3) selecting specific types of clinical entities (medications, diseases, pathological conditions), filtering events (removing negated events, past events, and events occurring in certain sections or in the wrong temporal order), and obtaining their frequencies as well as frequencies of co-occurring pairs; 4) computing statistical associations; and 5) using knowledge as well as statistical and information theoretical methods to remove confounding associations.

There are a number of challenges involved in each of the main steps, which are summarized below:

1. The narrative records are often telegraphic, and frequently contain atypical abbreviations and ungrammatical sentences; their heterogeneous nature means that there are usually different formats for each different type of note, such as use of new lines, indentation, and tabular formats, which do not have the usual sentence endings resulting in run on sentences; additionally, the notes omit or have unusual section headers, which are often important for accurate retrieval.
2. Coded EHR data may have nonstandard codes, which require mapping to the standard codes, which, in our case, is the UMLS. Additionally, abnormal ranges of values have to be specified for each test in order to enable filtering out of normal results, and each combination of laboratory test-abnormal value must be mapped to the appropriate codes denoting the abnormality. For example, a creatinine measurement that is high should be mapped to the UMLS concept C0235431 denoting *Blood creatinine increased*. Finally, the granularity of the coding system is usually problematic because fine-grained codes may result in the distribution of equivalent concepts over many values, causing a dilution of the signal. For example, in the UMLS there are many different highly specific concepts associated with cough, such as *brassy cough, aggravated cough, non-productive cough,* and *nocturnal cough,* which should all be considered equivalent to *cough* for signal detection.
3. Selecting specific types of entities, such as medications and pathological conditions, depends on the accuracy of the coding system or ontology and the granularity of the

semantic classification. For example, the UMLS has a class called **Finding**, which consists of normal, abnormal, and irrelevant conditions (e.g. *family planning*), and refinement of the classification is usually necessary, but is costly. Standardization of medications presents another problem. For example, brand name medications should be mapped to generic names, but the process is not straightforward. Contextual filtering according to temporality is critical, because, ideally, only conditions that occur after medications are administered should be considered as candidates for ADEs; however, temporal processing is very complex, and temporal information is often incomplete.

4. Statistical associations may not be clinically relevant. Although two entities are associated, the exact relation cannot be determined based on statistics, and statistical associations may not be clinically interesting. For example, a drug-disease association may represent a relation where the drug treats/prevents the disease, where the drug causes the disease, or where the association is a confounder because it is indirectly caused by another relation. Since there are many interdependencies in the data, confounding is a substantial problem.
5. Executable knowledge concerning indications for medications, manifestations of diseases, and known ADEs would be helpful in reducing confounding but most of the information is available only through proprietary databases, which cannot be mined. Some knowledge in the UMLS exists concerned with specification of indications for medications, but there are many medications that are not in that knowledge base. Use of information theory to classify an association as indirect using mutual information and the data processing inequality is relatively effective but has limitations because it only signifies that a direct link has not been found based on the data that was provided. Further research is needed to develop more sophisticated methods that provide a means of inferring confounding effects of variables.

3 Conclusions

Access to comprehensive information in the EHR offers many interesting research possibilities within the clinical domain. This talk focuses on a specific application, concerned with mining the EHR to discover novel ADEs for the purpose of improving patient safety and reducing costs, but many more automated applications are possible. An overview of the method under development was summarized and many challenging research issues that need to be explored further was described. NLP is a critical technology for the clinical domain, and leads to the need for more research opportunities in the field concerned with knowledge acquisition and discovery, patient management, health care management, and biosurveillance.

References

1. Classen, D.C., Pestotnik, S.L., Evans, R.S., Lloyd, J.F., Burke, J.P.: Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. JAMA 277(4), 301–306 (1997)
2. Bates, D.W., Spell, N., Cullen, D.J., et al.: The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. JAMA 277(4), 307–311 (1997)

3. Schneeweiss, S., Hasford, J., Gottler, M., Hoffmann, A., Riethling, A.K., Avorn, J.: Admissions caused by adverse drug events to internal medicine and emergency departments in hospitals: a longitudinal population-based study. Eur. J. Clin. Pharmacol. 58(4), 285–291 (2002)
4. Chiang, A.P., Butte, A.J.: Data-driven methods to discover molecular determinants of serious adverse drug events. Clin. Pharmacol. Ther. 85(3), 259–268 (2009)
5. Amery, W.K.: Why there is a need for pharmacovigilance. Pharmacoepidemiol Drug Saf. 8(1), 61–64 (1999)
6. Goldman, S., Kennedy, D., Graham, D., et al.: The clinical impact of adverse event reporting. Center for Drug Evaluation and Research. Food and Drug Administration (1996)
7. Moride, Y., Haramburu, F., Requejo, A.A., Begaud, B.: Under-reporting of adverse drug reactions in general practice. Br. J. Clin. Pharmacol. 43(2), 177–181 (1997)
8. Wood, L., Martinez, C.: The general practice research database: role in pharmacovigilance. Drug Saf. 27(12), 871–881 (2004)
9. Berlowitz, D.R., Miller, D.R., Oliveria, S.A., Cunningham, F., Gomez-Caminero, A., Rothendler, J.A.: Differential associations of beta-blockers with hemorrhagic events for chronic heart failure patients on warfarin. Pharmacoepidemiol. Drug Saf. 15(11), 799–807 (2006)
10. Friedman, C., Shagina, L., Lussier, Y., Hripcsak, G.: Automated encoding of clinical documents based on natural language processing. J. Am. Med. Inform. Assoc. 11(5), 392–402 (2004)
11. Lindberg, D., Humphreys, B., McCray, A.T.: The Unified Medical Language System. Meth. Inform. Med. 32, 281–291 (1993)
12. Cao, H., Hripcsak, G., Markatou, M.: A statistical methodology for analyzing co-occurrence data from a large sample. J. Biomed. Inform. 40(3), 343–352 (2007)
13. Wang, X., Friedman, C., Chused, A., Markatou, M., Elhadad, N.: Automated knowledge acquisition from clinical narrative reports. AMIA Annu. Symp. Proc., 783–777 (2008)
14. Chen, E.S., Hripcsak, G., Xu, H., Markatou, M., Friedman, C.: Automated acquisition of disease-drug knowledge from biomedical and clinical documents: an initial study. J. Am. Med. Inf. Assoc. 15(1), 87–98 (2008)
15. Wang, X., Hripcsak, G., Markatou, M., Friedman, C.: Active Computerized Pharmacovigilance using Natural Language Processing, Statistics, and Electronic Health Records: a Feasibility Study. J. Am. Med. Inform. Assoc. 16(3), 328–337 (2009)
16. Wang, X., Hripcsak, G., Friedman, C.: Characterizing environmental and phenotypic associations using information theory and electronic health records. In: 2009 AMIA Summit, March 15, p. 134 (full paper selected for publication in BMC Bioinformatics) (2009)

Computer Vision: A Plea for a Constructivist View

Catherine Garbay

Laboratoire d'Informatique de Grenoble, CNRS-Université de Grenoble, Batiment IMAG B, 385 avenue de la Bibliothèque, 38400, Saint Martin d'Hères, France
Catherine.Garbay@imag.fr

Abstract. Computer vision is presented and discussed under two complementary views. The positivist view provides a formal background under which vision is approached as a problem-solving task. By contrast, the constructivist view considers vision as the opportunistic exploration of a realm of data. The former view is rather well supported by evidence in neurophysiology while the latter view rather relies on recent trends in the field of distributed and situated cognition. The notion of situated agent is presented as a way to design computer vision systems under a constructivist hypothesis. Various applications in the medical domain are presented to support the discussion.

Keywords: Computer Vision, Distributed Cognition, Situated Agents, Medical Image Processing.

1 Introduction

The objective of computer vision is to support the coherent description and interpretation of visual scenes, whether autonomously, or semi-automatically. This objective is known to raise challenging issues, for several reasons:

- vision systems have to operate in incompletely specified and partially known universes
- the appearance of objects may vary depending on several conditions (illumination, capture, occlusions…)
- the interpretation task highly depends on context : actual grey level values do not bring significant information when taken in isolation
- the tools at hand are difficult to evaluate, their robustness and adequacy is only partially known
- there is a gap (so-called semantic gap) between the symbolic apprehension of high level concepts and their concrete instantiation in images
- goals are ill-defined, since the purpose of computer vision systems is precisely to provide a description of the environment in which it is meant to operate

Computer vision appears as a scientific domain at the crossroads of multiple influences, from mathematics to situated cognition. Recent work has focalized on the mathematical view on vision and hence on a rather positivist view, according to

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 6–15, 2009.

which vision is seen as an optimization problem. I argue in this paper for a complementary view, where vision is seen as a joint construction process, involving the mutual elaboration of goals, actions and descriptions.

2 A Positivist View

Mathematical methods have attracted considerable attention in the last decades with amazing results in varied application domains, from MR image interpretation to real-time 3D or video image processing. Three main categories of approaches can be distinguished, namely variational, statistical and combinatorial methods [1]. Among the strengths of these methods, apart their computing performance and their formal basis, one can quote the possibility to integrate many terms of variations in complex objective functions, to learn dedicated visual tasks from complex conditional, multi-dimensional distributions, or to cope with varied levels of modelling in a distinct way.

Whatever the approach, there is still no universal method to address visual perception issues and the choice of the most appropriate technique is rather task-driven.

Among the guiding lines of these modern approaches, one can quote:

- to capture variability: the goal is not to represent a "mean view" of an object, but rather to capture the variations of its appearance
- to integrate heterogeneous knowledge domains together with contextual information : a region is not delineated based on a single characteristic obeying a universal law, but as an arrangement of several variables, and because it differentiates from others regions in its surrounding
- to minimize the *a priori* needed to recognize a scene, and to avoid the use of intuitive representations, by looking closer to the realm of data and its internal consistency (look for regularities and for problem sensitive descriptors, model only the variations that are useful)
- to deconstruct the notion of object: consider the object not as a "unity" nor as a "whole" but as a combination of patches, or singular points
- to deconstruct the notion of category : do not model a category by its essence, but through its marginal elements

These approaches have known a wide development in the last decade, and their formal bases are well elaborated. In addition, they have recently been demonstrated to model efficiently various neurophysiological and perceptive phenomena.

An attempt to translate the Gestalt theory program - what are the laws and the combination of laws that govern the grouping of distinct entities into consistent objects - into a mathematical and computer vision program has been proposed recently [2]. The idea is more precisely to translate the qualitative geometric phenomenological observations provided by this theory into quantitative laws that might support numerical simulations and be used for analysis. Gestalts may also be seen as sets of elements whose arrangement could not occur in noise. In other terms, any structure showing too much consistency to be found by chance in noise may be seen as a coherent perception. A contrario methods have evolved recently on this basis.

Combining ideas and approaches from biological and computer vision is another recent trend. In [3] a simple and efficient architecture for boundary detection is

proposed that is inspired from models of biological vision and establish a link with variational techniques. Very efficient object identification in real-time has been obtained by SpikeNet [4], an image-processing system that uses very large-scale networks of asynchronously firing neurons working on a single feed-forward pass. Such comparisons, in addition to bring novel image processing paradigms, allow checking the biological plausibility of experimentally suggested hypotheses.

3 A Constructivist View

3.1 Vision as a Distributed, Situated and Prescriptive Process

Distributed, situated and embodied cognition [5], [6], [7] are complementary views that have evolved in the late seventies to emphasize the role of interaction in cognitive processes. Roughly speaking, the conjecture is that cognition does not reduce to local information processing but rather develops and operates within complex dynamics tying human and its physical, mental and social environment. A “constructive” vision of the relation between mind and the world is proposed, according to which cognitive processes are not only influenced by the context in which they operate, but more deeply active constructors of the environment in which they evolve.

The theory of activity has been supporting this idea, that "the structuring of activity is not something that precedes it but can only grow directly out of the immediacy of the situation" [8]. Stated differently, the involvement in action creates circumstances that might not be predicted beforehand [6].

Going a step further, one may approach human vision as a situated and active process [9] evolving in a physically distributed universe (images). Context is known to play a central role in the orientation of visual processes, and vision may not be considered as an abstract and desembodied process : "the context in which cognitive activity takes place is an integral part of that activity, not just the surrounding context for it" [11].

Therefore, vision is considered as an incremental process, operating transformations on the world that in turn modify the way the environment is perceived and interpreted. Vision, as intelligence, is not a representation activity but rather an active process guided by information collection and knowledge acquisition. It does not obey any predefined goal but rather appears as a prescriptive process according to which past perception drives the formation of future perception. This process may be further characterized by the intrinsic mediation tying the goal, the activity and its context, and the result this activity [8].

3.2 Situated Agents

The field of Distributed Artificial Intelligence has known considerable development in the same decades, particularly under the vision of Minsky of a “Society of Mind" [12].

Agents are autonomous entities sharing a common environment and working in a cooperative way to achieve a common goal. They are provided with limited perception abilities and local knowledge and are meant to evolve to adapt to novel situations, by modifying either their own behaviour or the whole system organization or the way they cooperate with others. Some advantages of multi-agent systems are among others

[13]: handling knowledge from different domains, designing reliable systems able to recover from incomplete or noisy information and wrong knowledge, focusing spatially and semantically on relevant knowledge, cooperating and sharing tasks between agents in various domains, reducing computation time through distributed and asynchronous implementation. They have been applied rather widely to medical image analysis [14].

In situated MASs [15], the environment embeds the agents and other domain objects, each on an individual position. The environment can have different topologies, depending on the application. Agent and environments constitute complementary parts of a complex system that can mutually affect each other.

In computer vision, the environment is anchored in the spatial space of the image (2D, 3D or 4D if time is to be considered). It is meant to contain the initial data at hand as well as information computed and collected by the agents, knowledge and models learn by experience, or traces of their behaviours. As a consequence, any result, any knowledge shunk may be shared by the community of agents; any individual action may influence other's actions.

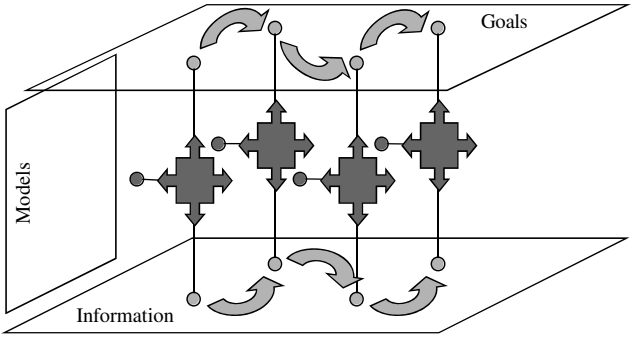


Fig. 1. The agents are situated physically (at a given spatial or temporal location), semantically (for a given goal or task) and functionnally (with given models or competences). Data, information and knowledge are shared through the common environment.

The agents are situated, that is anchored at a given position in the problem space, in terms of data to analyze, goals to be pursued and models to proceed (Fig. 1). These agents work in a specialized and local way, they produce partial results that are shared via the environment. The agents are provided with estimation and learning capabilities and perform a dual adaptation : internal adaptation by the selection of adequate processing models, according to the situations to be faced and to the goals to be reached, external adaptation by the dynamical generation of constraints, e.g. of new sets of data and goals to explore ; such adaptation may require the creation of new agents, modifying as a consequence the structure of the analyzing system itself.

Three modes of cooperation can be distinguished in this framework, namely augmentative, integrative, and confrontational, depending whether a task is (i) distributed to agents working in parallel on partial sets of data, (ii) distributed to agents with complementary competences, or (iii) executed by competing agents with similar competences

[16]. Besides its adaptation and cooperation abilities, situated agents are provided with focusing capabilities which allows exploring new locations in the image, goal or activity space, either by modifying the agent "location" or by creating new agents in given locations. Such design promotes a coupling between (i) the agent perception, activity and intention and (ii) the dynamics of the agent and its environment.

3.3 Facing the Co-determination Issues

What appears difficult to face in computer vision is the co-determination between goals, actions and situations : goals are conditioned by action results as well as encountered situations, while action selection and focusing depends on goals and action results depends on the situations at hand.

As a matter of fact, the hypothesis that there is a rationale for action that exists separately and independently from the action itself does not hold for the considered niche [10]. There is no predefined external goal, rather there is a constant interleaving of mutually dependent analyses occurring at different levels (production of goals, selection of actions, focusing, detection…), each of them being situated in the context of others.

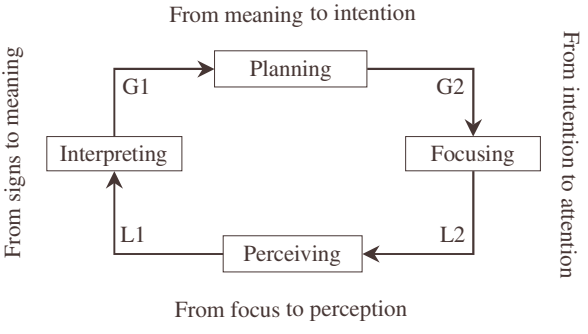


Fig. 2. Scene interpretation as interleaving semantic and praxiological gap issues

Elucidating the dynamics at hand raises difficult issues. It may be considered as involving two distinct sub-problems with specific characteristics (Fig. 2):

- semantic gap: how to build a global and consistent interpretation (G1) from local and inconsistent percepts (L1) acquired in the framework of given focus of attention (L2)
- praxiological gap: how to derive local focus of attention (L2) from a global intention (G2) formulated as the result of the perceived scene understanding (G1)

According to this framework, vision is defined as the ability to establish a viable coupling between an intentional dynamic, an attentional dynamic, and an external environment on which to act. The raised complexity issues may be approached under the situated agent framework described in the previous section, according to which the scope of perception, goal formation and focusing is restricted while action takes place on a situated basis.

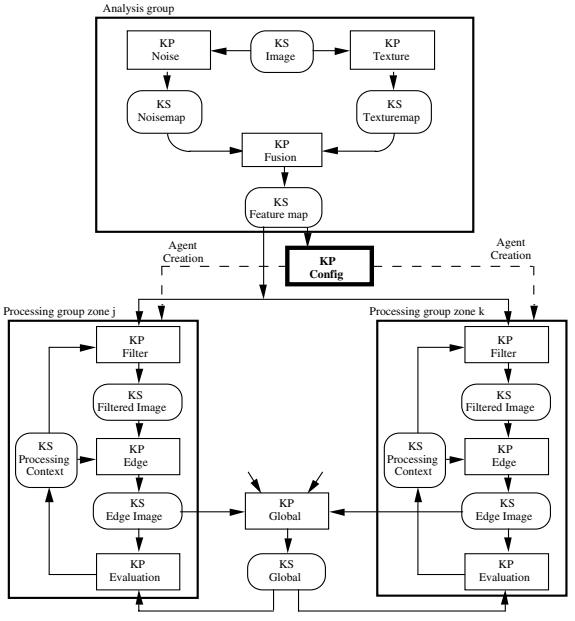


Fig. 3. System architecture: it is based on a distinction between two type of agents, namely Knowledge Sources (KS agents) and Knowledge Processors (KP agents)

4 Design Examples

4.1 Context-Based Adaptation of Image Processing Styles

The purpose of this work [17] was to discuss the potential of a multi-agent approach to adapt edge detection operators to low-level image characteristics. The system, whose architecture is presented in Fig. 3, proceeds in two steps (characterization and processing). In a first step, noise and texture maps are computed, which allow partitioning the image into zones presenting similar noise/texture features. In a second step, and for each zone, a filtering/edge detection strategy is applied, based on considering the zone characteristics and evaluating the resulting edge map. Heuristic knowledge is provided to the processing KP agent, which allows it (i) to select a processing operator given the zone characteristic and (ii) to adjust the parameters given a quality evaluation criterion. The process is iterated independently for each zone until a convergence criterion is reached. The results are finally combined in a global segmentation image. The potential of the approach has been experimented on synthetic as well as natural images.

4.2 Combining Various Cooperation Strategies for MRI Brain Scan Segmentation

Automatic segmentation of MRI brain scans is a complex task because of the variability of the human brain anatomy, which limits the use of general knowledge, and because of the artifacts inherent to MRI acquisition, which results in biased and noisy

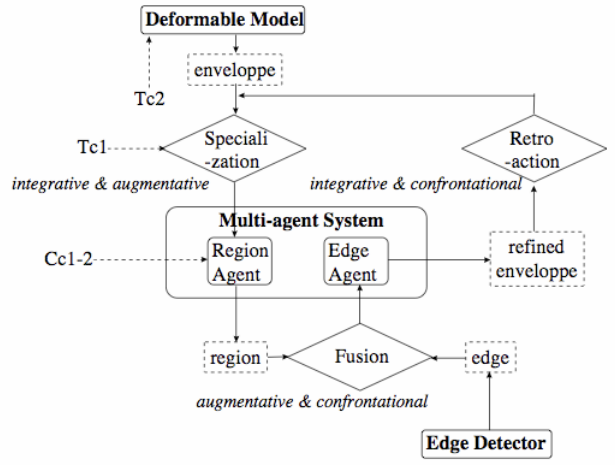


Fig. 4. Global view of the processing steps and information flow. The three systems used in our framework are represented by rectangles. Dashed rectangles show the results provided. Each triangle indicates a step in the global segmentation process. For each step, the modes of cooperation are indicated. The arrows indicate the information flow and the dashed arrows the introduction of domain knowledge : Tc1 (topology criterion tying white matter and grey matter), Tc2 (topology criterion tying grey matter and cerebro-spinal fluid), Cc1 and Cc2 (classification criteria to distinguish between grey matter and white matter).

images. To tackle these difficulties we have proposed a framework [16] to support the cooperation between several heterogeneous sources of information and several processing styles (deformable model, region growing and edge detection). The agents operate under automatically and dynamically generated constraints involving the gray levels specific to the considered image, statistical models of the brain structures and general knowledge about MRI brain scans. Integrative, augmentative and confrontational cooperation styles are distinguished and combined during the three steps of the segmentation process (Fig. 4) : specialization of the seeded-region-growing agents, fusion of heterogeneous information and retroaction over slices. These modes allow the mixing of heterogeneous information (model based or data driven) to dynamically constraint pixel growing agents. The described cooperative framework allows the dynamic adaptation of the segmentation process to the individual characteristics of each MRI brain scan.

4.3 Situated Agent-Based Processing for Cell Migration Analysis

A multi-agent model has been developed for the analysis of *in vitro* cell motion from image sequences. A generic agent model has been proposed [18], where agents integrate perception, interaction and reproduction behaviours (Fig. 5). Perception behaviour classifies pixels based on static and motion based criteria. Interaction behaviour allows two agents to merge or to negotiate parts of regions. Reproduction behaviour specifies ways to explore the images. Agent's behaviours are specialized at execution time, depending on the goals to achieve, that is on the cell image component to be

processed. These goal-oriented agents are created and located dynamically on given parts of the image ; they may in turn launch new agents when needed. An internal manager is provided to each agent to control the behaviour's execution. It makes use of an event-driven scheme to manage the behaviour priorities. The frames are processed in pipeline, and information from previous frame is used to treat the current frame.

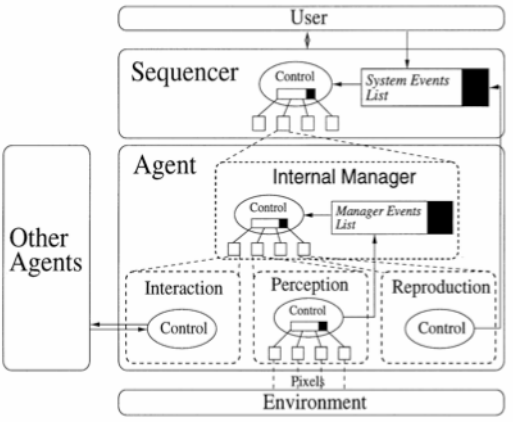


Fig. 5. Global view of the system architecture. Perception, interaction and reproduction behaviours are interleaved in an opportunistic way, depending on the encountered events (detection of some information, presence of an other agent, life time of an agent…)

4.4 Distributed Markovian Processing for MRI Segmentation

We have proposed in [19] an original markovian approach from MRI segmentation. Markovian segmentation is a widely used probabilistic framework, which introduces local regularization via spatial dependencies between voxels, thus resulting in robustness to local perturbations such as noise. A major obstacle to accurate MRI segmentation comes from the presence of heavy spatial intensity variations within each tissue, which results in difficulties to process the entire volume with single tissue models. Bias field modelling is often considered in addition to cope with such difficulty. Our approach involves a set of distributed agents operating in a local way : each individual agent proceeds to the autonomous estimation of tissue models, which results in the construction of a set of models representing the intensity variations over the volume. Inter-agent cooperation is provided, to ensure the local model consistency. In addition, structure segmentation is performed via dedicated agents, which integrate anatomical spatial constraints provided by an a priori fuzzy description of brain anatomy. Structure agents cooperate with tissue agents in order that gradually more accurate modelling is provided (Fig. 6). Experiments conducted over synthetic as well as natural images have demonstrated the computational efficiency as well as the accuracy of the approach.

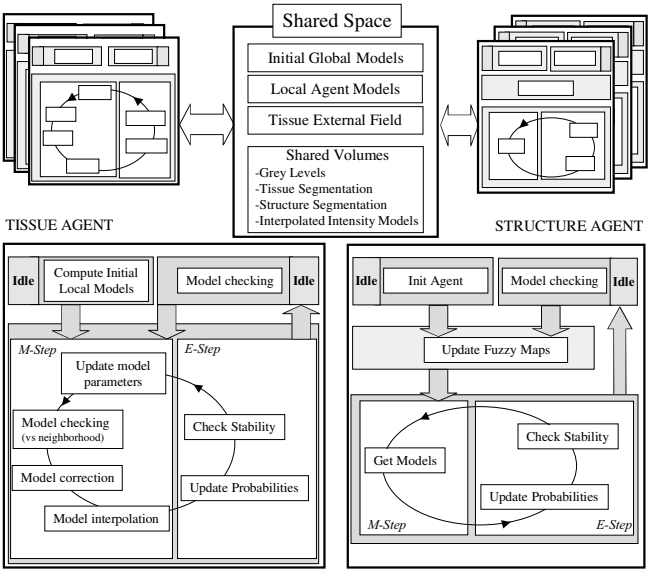


Fig. 6. The proposed system architecture : structure and tissue agents cooperated via the MRF framework to ensure accurate image segmentation. The agents operate on a local basis to cope with intensity variations. Cooperation between neighbouring tissue agents is provided to ensure the consistency of local models.

5 Conclusion

We have opposed in this paper two complementary views on computer vision : a positivist view, which provides a formal basis on which to ground processing algorithms, and a constructivist view, which results in the design of distributed architectures involving situated cooperating agents. Various examples from the latter approach have been presented, to emphasize the wide range of design that may be envisaged. Both views must be considered as complementary : explicitating and analyzing their differences is crucial and further work is needed to envisage their mixing within coherent problem solving framework [20]. Mixing markov-based joint modelling and agent-based distributed processing is a step in this direction.

References

1. Paragios, N., Chen, Y., Faugeras, O.: Preface. In: Paragios, N., Chen, Y., Faugeras, O. (eds.) The Handbook of Mathematical Models in Computer Vision, pp. 1–5. Springer, Heidelberg (2005)
2. Desolneux, A., Moisan, L.N., Morel, J.M.: From Gestalt Theory to Image Analysis : A Probabilistic Approach. Springer, Heidelberg (2008)
3. Kokkinos, I., Deriche, R., Faugeras, O., Maragos, P.: A Computational Analysis and Learning for a Biologically Motivated Model of Boundary Detection. Neurocomputing 71, 1798–1812 (2008)

4. Thorpe, S.J., Guyonneaua, R., Guilbauda, N., Allegrauda, J.-M., VanRullen, R.: SpikeNet: Real-Time Visual Processing with one Spike per Neuron. Neurocomputing 58–60, 857–864 (2004)
5. Hutchins, E.: Cognition in the Wild. MIT Press, Cambridge (1995)
6. Suchman, L.A.: Plans and Situated Action: the Problem of Human-Machine Interaction. In: Pea, R., Brown, J.S. (eds.). Cambridge University Press, Cambridge (1987)
7. Varela, F., Thomson, E., Rosch, E.: L'Inscription Corporelle de l'Esprit, Le Seuil, Paris (1993)
8. Nardi, B.A.: Studying Context: a Comparison of Activity Theory, Situated Action Models and Distributed Cognition. In: Nardi, B.A. (ed.) Context and Consciousness: Activity Theory and Human-Computer Interaction. MIT Press, Cambridge (1996)
9. Weyns, D., Steegmans, E., Holvoet, T.: A Model for Active Perception in Situated Multi-Agent Systems. Applied Artificial Intelligence 18, 867–883 (2004)
10. Bianchi, N., Bottoni, P., Spinu, C., Garbay, C., Mussio, P.: Situated Image Understanding in a Multi-Agent Framework. International Journal of Pattern Recognition and Artificial Intelligence 12, 595–624 (1998)
11. Resnick, L.B.: Shared Cognition: Thinking as Social Practice, Perspectives on Socially Shared Cognition. In: Resnick, L.B., Levine, J.M., Thistle, S.D. (eds.), pp. 1–20. American Psychological Association, Washington (1991)
12. Minsky, M.: The Society of Mind. Simon and Shuster, New York (1985)
13. Shariatpanahi, H.F., Batmanghelich, N., Kermani, A.R.M., Ahmadabadi, M.N., Soltanian-Zadeh, H.: Distributed Behavior-Based Multi-Agent System for Automatic Segmentation of Brain MR Images. In: International Joint Conference on Neural Networks, Vancouver, BC, Canada, pp. 4535–4542 (2006)
14. Bovenkamp, E.G.P., Dijkstra, J., Bosch, J.G., Reiber, J.H.C.: Multi-Agent Segmentation of IVUS Images. Pattern Recognition 4, 647–663 (2004)
15. Maes, P.: Situated Agents can have Goals. In: Maes, P. (ed.) Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back, pp. 49–70. MIT Press, London (1990)
16. Germond, L., Dojat, M., Taylor, C.J., Garbay, C.: A Cooperative Framework for Segmentation of MRI Brain Scans. Artificial Intelligence in Medicine 20, 77–93 (2000)
17. Spinu, C., Garbay, C., Chassery, J.M.: A Cooperative and Adaptive Approach to Medical Image Segmentation. In: Wyatt, J.C., Stefanelli, M., Barahona, P. (eds.) AIME 1995. LNCS, vol. 934, pp. 379–390. Springer, Heidelberg (1995)
18. Boucher, A., Doisy, A., Ronot, X., Garbay, C.: A society of Goal-Oriented Agents for the Analysis of Living Cells. Artificial Intelligence in Medecine 14, 183–199 (1998)
19. Scherrer, B., Dojat, M., Forbes, F., Garbay, C.: Agentifiation of Markov Model Based Segmentation: Application to MRI Brain Scans. Artificial Intelligence in Medicine 46, 81–95 (2009)
20. Oulasvirta, A., Tamminen, S., Höök, K.: Comparing Two Approaches to Context: Realism and Constructivism. In: Proc of the 4th Decennial Conference on Critical Computing: between Sense and Sensibility, Aarhus, Denmar, pp. 195–198 (2005)

Mining Healthcare Data with Temporal Association Rules: Improvements and Assessment for a Practical Use

Stefano Concaro[1,2], Lucia Sacchi[1], Carlo Cerra[2], Pietro Fratino[3], and Riccardo Bellazzi[1]

[1] Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy
[2] Sistema Informativo Aziendale e Controllo di Gestione, ASL di Pavia, Italy
[3] Università di Pavia, Italy
stefano.concaro@unipv.it, lucia.sacchi@unipv.it, carlo_cerra@asl.pavia.it, pietrofratino@virgilio.it, riccardo.bellazzi@unipv.it

Abstract. The Regional Healthcare Agency (ASL) of Pavia has been maintaining a central data repository which stores healthcare data about the population of Pavia area. The analysis of such data can be fruitful for the assessment of healthcare activities. Given the crucial role of time in such databases, we developed a general methodology for the mining of Temporal Association Rules on sequences of hybrid events. In this paper we show how the method can be extended to suitably manage the integration of both clinical and administrative data. Moreover, we address the problem of developing an automated strategy for the filtering of output rules, exploiting the taxonomy underlying the drug coding system and considering the relationships between clinical variables and drug effects. The results show that the method could find a practical use for the evaluation of the pertinence of the care delivery flow for specific pathologies.

Keywords: Temporal data mining, temporal association rules, hybrid events, healthcare data, diabetes mellitus.

1 Introduction

Health care organizations are increasingly collecting large amounts of data related to their day-by-day activities. Since 2002, the Regional Healthcare Agency (ASL) of Pavia has been collecting and maintaining a central data repository to trace all the main healthcare expenditures of the population of Pavia area (about 530000 people) in charge of the National Healthcare System (SSN). Since the main purposes of the project are related to economic reimbursement, the repository stores *administrative* healthcare data, mainly concerning physicians' workload, patients' hospital admissions, drug prescriptions and lab tests. In 2007 the ASL of Pavia started to collect also *clinical* healthcare data related to subgroups of patients selected on the basis of the most prevalent pathologies in the population. In particular diabetes, hypertension, cardiovascular diseases and several types of cancer were considered. The analysis of such healthcare databases, which integrate both administrative and clinical data, could greatly help to gain a deeper insight into the health condition of the population and to

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 16–25, 2009.

extract useful information that can be exploited in the assessment of health care processes and in organizational learning [1].

Given the large dimension of these databases and the primary role played by the temporal features therein stored, we exploited Temporal Data Mining techniques [2] to analyze this kind of data. In particular, we recently developed an algorithm for the extraction of Temporal Association Rules (TAR) over a set of multivariate temporal sequences of hybrid events [3]. In this paper we show how the method can be extended to suitably manage the integration of both clinical and administrative data, in order to mine interesting frequent temporal associations between diagnostic or therapeutic patterns. Moreover, we address the problem of developing an automated strategy for the filtering of output rules, exploiting the taxonomy underlying the drug coding system and considering the relationships between clinical variables and drug effects. As an example of the application, we focus our analysis on a sample of patients suffering from *Diabetes Mellitus*. The results show that, considering the perspective of a Regional Healthcare Agency, the method could be put into practice to monitor the quality of the care delivery flow for specific pathologies and for specific clinical conditions characterizing the analyzed population.

2 Data Overview

In the context of the Italian National Healthcare System (SSN) the Regional Healthcare Agencies have a central role in the coordination of the care delivery process to the assisted population. In particular, they can provide recommendations and protocols to General Practitioners (GPs) for the care of the most prevalent pathologies in the population. In this context, during 2007 the ASL of Pavia started a collaboration with general practitioners aimed at collecting clinical data related to a selected subgroup of about 1300 diabetic patients living in the Pavia area. This data collection is aimed at providing a feedback about the efficacy of the care delivery process for primary care. The selected patients periodically undergo a medical visit and their GP is responsible for the transmission of their personal clinical data to the ASL. In the period between January 2007 and October 2008 (22 months) a total of about 5000 inspections was recorded. The clinical variables considered in this study are mainly pato-physiological parameters, derived from the results of clinical tests, and information about the current medical care, as summarized in Table 1. The temporal location of the data stored in the database is defined by the date of the visit.

The central repository of the ASL of Pavia offers the opportunity to join data from the GPs' inspections with the data stored into the healthcare administrative DataWarehouse (DW), which records all the main healthcare expenditures of the assisted population. Since this DW is mainly devoted to reimbursement purposes, the reported information is in the form of "process" data. In particular, the main administrative data refer to: hospital admissions (provided through the hospital discharge record including DRG codes and ICD9-CM diagnoses and procedures), drug prescriptions (provided through the ATC[1] code) and ambulatory visits (defined by a specific Italian national code, DGR n. VIII/5743 - 31/10/2007). Despite the administrative nature of these kind of data, several

[1] Anatomical Therapeutic Chemical classification system.

epidemiological details can be inferred from diagnoses related to hospital admissions or from the type of drugs used to treat specific diseases. On the other hand, in the case of ambulatory visits, many clinical details are not reported. For example, we know that a patient underwent a blood glucose test on a specific day, but we have no information about the outcome of the test.

Table 1. List of the clinical variables considered in the analysis on diabetic patients. The range of possible values and the Interquartile (IQ) range are reported for variables 1-9, while for attributes "Anti-Hypertensive Therapy" and "Care Intervention" the allowed categorical values are listed.

Variable	Range	IQ Range	Unit
1. Body Mass Index (BMI)	[10-80]	[25.15-31.28]	Kg/m^2
2. Systolic Blood Pressure (SBP)	[60-240]	[130-150]	mmHg
3. Diastolic Blood Pressure (DBP)	[30-150]	[75-85]	mmHg
4. Glycaemia	[50-500]	[112-162]	mg/dl
5. Glycated Haemoglobin (HbA1c)	[3-20]	[6.3-7.9]	%
6. Total Cholesterol	[80-500]	[175-232]	mg/dl
7. HDL Cholesterol	[10-120]	[43-62]	mg/dl
8. Triglycerides	[10-2000]	[91-177]	mg/dl
9. Cardio-Vascular Risk (CVR)	[0-100]	[8.57-30.33]	%
10. Anti-Hypertensive Therapy	{Yes; No}	-	-
11. Care Intervention	{Diet; Health training; None}	-	-

3 Methods

3.1 Integrating Administrative and Clinical Data through Temporal Abstractions

The mining of the databases described in the previous section requires to handle the integration of both administrative and clinical data.

On the one hand, healthcare administrative data are by nature represented by sequences of events. A sequence of events can be defined as a time ordered succession of *episodes*, where an episode formally identifies a single instance of a specific *event*. In more detail, each episode: i) represents a single occurrence of an event (for example the prescription of a specific drug); ii) is related to a subject (for example a specific patient) and iii) is characterized by its temporal coordinates within an observation period.

On the other hand, clinical data are usually a set of time series of numeric values (e.g. the time series of systolic pressure values). In order to get a uniform representation of these data as temporal sequences of events, the clinical data needed first to undergo to a pre-processing procedure. In particular, we exploited knowledge-based Temporal Abstractions (TAs) to shift from a time point quantitative representation of the time series to a qualitative interval-based description of the available data [4]. To this end both state and trend TAs were exploited in the procedure. *State detection* was applied to discretize clinical continuous values in state intervals. These intervals were determined

on the basis of physiological thresholds selected by an expert clinician on the basis of his experience and according to the most recent European guidelines on diabetes care. For example, a value of 200 mmHg for systolic blood pressure is represented as an episode of "high Blood Pressure" state. *Trend detection* was then used to describe state changes in two consecutive visits, thus allowing the definition of *Increasing*, *Steady* or *Decreasing* intervals for each variable. Given for example the case where an "high glycaemia" state was detected during a first visit and a "normal glycaemia" state was detected in the following visit, this determines a trend episode of "Decreasing glycaemia".

The integration of the available data sources, properly pre-processed as discussed, leads to a uniform representation of the data as temporal sequences of healthcare events (Figure 1).

In the following we will exploit the administrative data related to drug prescriptions only, even if the procedure could be easily extended to represent also events of hospital admissions or lab tests. In this analysis the length of the time interval related to each episode is estimated on the basis of the days of Defined Daily Dose (DDD). The DDD is defined as the assumed average maintenance dose per day for a drug used in its main indication in adults. In our case, given a prescription including a total drug amount x, the length of the episode interval is estimated by the quantity x/DDD, which is expressed in days.

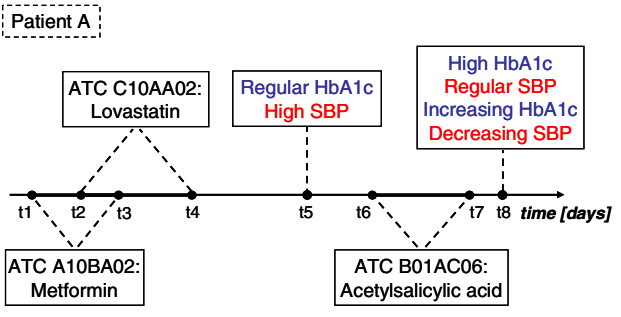


Fig. 1. Example of the integration of clinical and administrative healthcare events through a uniform representation of temporal sequence. The temporal granularity is fixed to 1 day.

The represented healthcare events considered in this analysis are then included in the following categories:

- *State Abstractions*: e.g. Glycaemia<65 (low), Glycaemia 110-180 (high), etc..;
- *Trend Abstractions*: e.g. Glycaemia Increasing, Glycaemia Decreasing, etc..;
- *Drug prescriptions*: e.g. C10BA02 (Metformin), B01AB01 (Heparin), etc..

In this context the temporal representation of the events is a central issue [5]. The temporal nature of a single episode is strongly dependent on the choice of the temporal *granularity*, which can be defined as the maximum temporal resolution used for the representation of all the sequences of events [6]. In our case, since both pharmaceutical archives and clinical databases store data with a resolution of one day,

the granularity was set equal to this value. Under this assumption, it is easy to note that the analyzed temporal sequences are made up of *hybrid* events, i.e. events characterized by an heterogeneous temporal nature. Considering the time dimension, the represented episodes can thus be classified as either with a duration (identified by a *time interval*) or without a duration (identified by a *time point*). Events like drug assumption, typically lasting a few days, can be represented by time intervals, while lab tests, performed in correspondence of a medical visit, can be represented by time points (Figure 1). Moreover, the overlapping episodes related to the prescription of the same drug for a specific patient were joined into a single continuous episode interval.

For each patient, the observation period for the generation of the temporal sequences was set between three months before the first visit and three months after the last one. Although this choice doesn't involve the same observation period for all the patients, it ensures the homogeneity of the analysis with respect to the frequencies of the different visits. Given these constraints, a total number of about 110000 healthcare episodes, partitioned in 1217 different healthcare events, was detected in the diabetic sample.

3.2 Temporal Association Rules

To carry on our analysis we run an algorithm devoted to the mining of Temporal Association Rules (TARs) over a set of temporal sequences of hybrid events [3]. Herein a TAR is defined as a relationship specified through a temporal operator which holds between an *antecedent*, consisting in a pattern of single or multiple cardinality, and a *consequent*, consisting in a pattern with single cardinality. Herein a pattern is defined as the occurrence of one or more contemporary events (e.g. "Antacids" together with "Beta-blockers"). The allowed temporal relationships are specified by Vilain [7] and Allen's [8] operators, with the addition of the PRECEDES operator [9]. Besides the mentioned temporal operators, the exploited algorithm is provided with three temporal constraints (*left shift, right shift* and *gap*) which are used to properly control the mutual distance of the antecedent and the consequent of a rule [10]. The rules extraction algorithm is designed following an Apriori-like strategy [11], where the rule search and selection is performed on the basis of thresholds on *confidence* and *support*. The definition of those two parameters was properly adapted to be applied in a temporal context with hybrid data [3, 9, 10].

Moreover, the algorithm offers the additional opportunity to select specific *rule templates*, defining the event classes allowed for the antecedent and the consequent selection respectively. This feature helps to focus the search only on relationships between the members of the classes that the user wants to investigate, and may be particularly useful to present the resulting rules to the users (e.g. clinicians). Considering the different types of represented events, the choice of the rule template allows several possibilities which lead to the extraction of knowledge related to different clinical scenarios. As a representative example of the method, in this analysis we focus on a specific rule template, where the antecedent selection is limited to state and trend abstractions, while the events allowed for the consequent selection are limited to drug prescriptions. This configuration aims at establishing which abstractions on physiological

parameters frequently show a significant temporal association with subsequent prescriptions of drugs, in order to interpret the medical prescriptions of specific drugs as a reaction of the clinical conditions observed during the visit.

In this analysis we investigated also the opportunity to exploit the taxonomical information that is included into the coding systems used to classify the administrative healthcare events. The common feature of these coding systems, concerning mainly the classification of drugs, diagnoses, procedures and ambulatory visits, lies in their hierarchical structure, which defines a taxonomy of concepts from a general to a specific level. Since we are focusing on drug prescriptions, we have to deal with the ATC coding system, which is made up of 5 hierarchical levels. We then tested a multiple-level (or hierarchical) mining strategy [12, 13] by choosing different levels of specificity along the taxonomy of the ATC classification system, adapting the application to our context where the temporal representation plays an essential role. Table 2 shows an example of ATC codes. The level number goes from the more general (1) to the more specific (5), and the codes increase the number of digits from one to seven. The rule mining algorithm may work with any level of aggregation of ATC codes.

Table 2. An example of the ATC classification system: Metformin is a biguanide, a particular kind of oral anti-diabetic drug, used to treat type II Diabetic patients and insulin resistance

Level	Code	Content	Description
1	A	Alimentary tract and metabolism	Anatomical main group
2	A10	Drugs used in Diabetes	Therapeutic subgroup
3	A10B	Oral Blood Glucose Lowering drugs	Pharmacological subgroup
4	A10BA	Biguanides	Chemical subgroup
5	A10BA02	Metformin	Chemical Substance

4 Results

In this section we present the results obtained on the temporal sequences of healthcare events extracted for the sample of diabetic patients. The thresholds for minimum support and confidence were herein set to *minsup=0,01* and *minconf=0,3*. The selection of thresholds for confidence and support particularly low with respect to the values used in typical data mining applications was primarly oriented to minimize the loss of potentially useful information. However, selecting these values for *minsup* and *minconf*, the depth of the analysis was limited to minimal samples of about 13 patients, which was considered a significant size to avoid the extraction of information related only to outliers. Following this approach the algorithm extracts a large set of rules. The crucial issue of reducing the output information was subsequently carried out through a post-processing procedure.

Since the target of the analysis was the investigation of precedence relationships, we chose to use the BEFORE temporal operator. The *gap* parameter, which defines the maximum allowed distance between antecedent and consequent, was set to *90* days, according to the knowledge about the usual temporal occurrence of drug prescriptions.

In order to avoid the extraction of a great amount of output information which is often irrelevant, uninteresting or redundant, the application of the algorithm was followed by a post-processing step, oriented to shrink the whole set of frequent rules generated by the mining step to a reduced set of "interesting" rules. The procedure is based on a objective measure of interestingness, as the rules are filtered according to the minimum improvement constraint (*minimp*) [14]. The minimum improvement of a rule of cardinality K>1, is defined as the ratio between the rule confidence and the maximum confidence value of all its subrules (cardinality={1,…,K-1}). The pruning step is then performed by keeping only the rules with *minimp>1*, i.e. those rules which increase the confidence value with respect to *all* their subrules. Several measures to evaluate rule interestingness, such as lift, leverage, Gini index, chi-squared, correlation indexes, etc.., were proposed in literature. Herein we chose to use the *minimp* constraint because it allows to perform a very restrictive reduction of the rule set, since it progressively exploits the predictive information provided by the confidence of the subrules as the rule cardinality increases. This post-processing step thus provides a solid and effective solution to the redundancy reduction. The rule ranking is then performed according to the decreasing order of respectively confidence and support values.

Table 3. A representative set of TARs defined by the BEFORE operator. The table reports the events included in the antecedent and in the consequent, the values for confidence and support, and the dummy variable (ClinR) used to classify the relationship between clinical variables and drug effects.

#	Antecedent	Consequent	Conf	Supp	ClinR
1	-HbA1c 7-8 (high) -CVR 5-10%	ATC A10: Drugs used in diabetes	0.635	0.021	1
		ATC A10B: Oral Blood Glucose Lowering drugs	0.635	0.021	
		ATC A10BA: Biguanides	0.5	0.017	
		ATC A10BA02: Metformin	0.5	0.017	
2	-BMI 25-30 (overweight) -Glycaemia 110-180 (high) -HbA1c > 9 (excessively high)	ATC B01: Antithrombotic agents	0.537	0.013	0
		ATC B01A: Antithrombotic agents	0.537	0.013	
		ATC B01AC: Platelet aggregation inhibitors, excluding heparin	0.464	0.011	
3	-BMI 30-40 (moderate obesity) -SBP 140-160 (moderate hypertension) -SBP Decreasing	ATC C09B: ACE inhibitors, combinations	0.373	0.015	1
		ATC C09BA: ACE inhibitors and diuretics	0.373	0.015	
4	-Glycaemia > 180 (very high) -Triglycerides > 350 (very high)	ATC C10A: Lipid modifying agents, plain	0.571	0.012	1

A reduced subset of 100 rules (20 rules for each ATC level), consisting in the best confidence ranked rules, was then submitted to a clinician who gave a medical evaluation of the discovery method. Each rule was classified on the basis of the presence or absence of a clinical evidence of a relationship between the physiological variables included in the antecedent and the effects of the prescribed drugs represented in the consequent. A subset of TARs obtained after the application of the mining step and the post-processing procedure is reported in Table 3.

For each rule (*#*) the table reports the events included in the antecedent and in the consequent, the values for confidence and support, and the classification performed by the clinician (*ClinR*)[2]. The variable ClinR is set to 1 if the clinician found a clinical evidence in the extracted rule, and 0 otherwise. The table shows also the results of the multi-level mining performed on the hierarchy of the ATC coding system, as highlighted in correspondence of the rules which span over multiple table rows. In particular, if we consider the rules labelled as #1, these correspond to all the rules that can be generated from the ATC classification A10, also reported in the example of Table 2. The different rules are related to different levels of the hierarchy.

5 Discussion

From the clinical viewpoint, the majority of the rules reported in Table 3 has a clear clinical meaning.

Rule 1 supports the evidence that a diabetic subject showing a Cardio-Vascular Risk within the interval 5-10% and an high glycated haemoglobin value, has a 63% probability to undergo a prescription of oral hypoglycemic agents in the following 90 days. More specifically, a patient satisfying this rule has a 50% probability to undergo a prescription of Metformin (Rule 1, 4th row). This association is verified in the 2% of the diabetic sample. This rule was considered to reflect a clinical relationship between antecedent and consequent concepts, as blood glucose lowering drugs have an effect on the glycated haemoglobin.

Rule 2 states that a subject found to be overweight and showing both an high glycemic and high glycated haemoglobin values has a 53% probability to undergo a prescription of antithrombotic agents in the following three months, and more specifically a 46% probability to undergo a prescription of platelet aggregation inhibitors. Apparently there is no direct clinical relationship between the physiological observations and the drug effects, as shown by the clinician's judgement set to 0.

Rule 3 shows that, given the simultaneous observation of moderate obesity, moderate hypertension and decrease in systolic blood pressure during an inspection, a diabetic subject has a 37% probability to undergo a prescription of ACE inhibitors in the following 90 days. This rule holds a clinical relationship between the involved concepts, since ACE inhibitors certainly have an effect on blood pressure regulation.

Rule 4 shows that, given a clinical condition characterized by very high glycaemia and very high triglycerides levels, a diabetic patient has a 57% probability to undergo a prescription of lipid modifying agents in the following three months. This rule

[2] The algorithm may also provide the average length of the antecedent episodes, the consequent episodes and the gap between them. This information, not shown here, can be useful to evaluate the temporal aspects of the discovered associations.

supports a close clinical relationship between the concepts in the antecedent and in the consequent, as lipid modifying agents have an effect on the level of triglycerides.

To summarize, the rule classification system based on the clinical meaning of the concepts proved to be effective. When considering the perspective of introducing this temporal data mining strategy into the clinical practice, this procedure helps to discriminate rules that have a compliance with available a-priori medical knowledge from rules that can suggest new findings. In the former case, the rule can simply be interpreted as a quantitative verification of the medical knowledge. In the latter case, strange or surprising rules can be considered as the starting point for deeper and more detailed analyses, oriented to find out whether the extracted associations are able to suggest unknown knowledge.

6 Conclusions

The analysis presented in this paper highlights the main potentials of the application of temporal associations rules for the mining of healthcare databases. The applied algorithm has shown to be an effective general method which allows to properly manage different heterogeneities of the data. First, the method allows to handle hybrid events, i.e. events characterized by heterogeneous temporal nature. Second, it allows to exploit the integration of different healthcare information sources, such as administrative and clinical data, through a uniform representation.

In this work we developed some features able to support the application of the algorithm and the management of the output results. A first feature consists in a post-processing pruning strategy, aimed at filtering the rules according to a statistical objective measure, the minimum improvement. This helps to reduce the problem of redundancy highlighting only the most interesting rules from a statistical viewpoint. A second feature is represented by a multiple-level data mining approach, which allows to select different levels of specificity on the hierarchical taxonomies of administrative events. The application was illustrated for the ATC coding system, even if the procedure can be easily extended to include also diagnoses and procedures (ICD9) or ambulatory visits. Finally, a classification of the rules was performed by a clinician on the basis of the presence or the absence of a clinical relationship between the concepts involved in the antecedent and the consequent of the rules. This feature greatly helps to discriminate rules that have a compliance with available a-priori medical knowledge from rules that can suggest new findings. As a future extension, this procedure could be introduced into an ontology-driven method [13], in order to automatically interpret and filter out the output rules according to the already available knowledge structured within a clinical ontology. A further extension will be aimed at estimating the true drug dosage from the available data of drug prescriptions, which intrinsically suffers from the variability of the drug purchase rate and actually doesn't include any remark about the real GP's intention to treat. The detection of variations in the drug dosage could potentially allow to find temporal associations with clinical variables, drug interactions or healthcare events like hospitalizations.

On the whole the method has been shown to be capable to characterize groups of subjects, highlighting interesting frequent temporal associations between diagnostic or therapeutic patterns and patterns related to the patients' clinical condition. Considering

the perspective of a Regional Healthcare Agency, this method could be properly used for a qualitative and quantitative evaluation of the observed care-flow for specific pathologies and for specific clinical conditions with respect to the recommended or expected care delivery flow.

References

1. Stefanelli, M.: The socio-organizational age of artificial intelligence in medicine. Artif. Intell. Med. 23, 25–47 (2001)
2. Post, A.R., Harrison, J.H.: Temporal data mining. Clin. Lab. Med. 28, 83–100 (2008)
3. Concaro, S., Sacchi, L., Cerra, C., Fratino, P., Bellazzi, R.: Temporal Data Mining for the Analysis of Administrative Healthcare Data. In: IDAMAP Workshop, Washington (2008)
4. Shahar, Y.: A framework for knowledge-based temporal abstraction. Artif. Intell. 90, 79–133 (1997)
5. Adlassnig, K.P., Combi, C., Das, A.K., Keravnou, E.T., Pozzi, G.: Temporal representation and reasoning in medicine: Research directions and challenges. Artif. Intell. Med. 38, 101–113 (2006)
6. Combi, C., Franceschet, M., Peron, A.: Representing and Reasoning about Temporal Granularities. J. Logic Comput. 14, 51–77 (2004)
7. Vilain, M.B.: A system for reasoning about time. In: 2nd National Conference in Artificial Intelligence, Pittsburgh, pp. 197–201 (1982)
8. Allen, J.F.: Towards a general theory of action and time. Artif. Intell. 23, 123–154 (1984)
9. Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R.: Temporal data mining for the quality assessment of hemodialysis services. Artif. Intell. Med. 34, 25–39 (2005)
10. Sacchi, L., Larizza, C., Combi, C., Bellazzi, R.: Data mining with Temporal Abstractions: learning rules from time series. Data Min. Knowl. Disc. 15, 217–247 (2007)
11. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In: 20th International Conference on Very Large Data Bases, pp. 487–499. Morgan Kaufmann Publishers Inc., San Francisco (1994)
12. Han, J., Fu, Y.: Discovery of Multiple-Level Association Rules from Large Databases. In: 21th International Conference on Very Large Data Bases, pp. 420–431. Morgan Kaufmann Publishers Inc., San Francisco (1995)
13. Raj, R., O'Connor, M.J., Das, A.K.: An ontology-driven method for hierarchical mining of temporal patterns: application to HIV drug resistance research. In: AMIA Annual Symposium, Chicago, pp. 614–619 (2007)
14. Bayardo, R.J., Agrawal, R., Gunopulos, D.: Constraint-Based Rule Mining in Large, Dense Databases. In: 15th International Conference on Data Engineering, pp. 188–197. IEEE Computer Society Press, Los Alamitos (1999)

A Temporal Data Mining Approach for Discovering Knowledge on the Changes of the Patient's Physiology

Corrado Loglisci and Donato Malerba

Dipartimento di Informatica
Universita' degli Studi di Bari - Italy
Via Orabona 4, 70125, Bari
{loglisci,malerba}@di.uniba.it

Abstract. Physiological data represent the health conditions of a patient over time. They can be analyzed to gain knowledge on the course of a disease or, more generally, on the physiology of a patient. Typical approaches rely on background medical knowledge to track or recognize single stages of the disease. However, when no one domain knowledge is available these approaches become inapplicable. In this paper we describe a Temporal Data Mining approach to acquire knowledge about the possible causes which can trigger particular stages of the disease or, more generally, which can determine changes in the patient's physiology. The analysis is performed in two steps: first, identification of the states of the disease (namely, the stages through which the physiology evolves), then detection of the events which may determine the change from a state to the next one. Computational solutions to both issues are presented. The application to the scenario of the sleep disorders allows to discover events, in the form of breathing and cardiovascular disorders, which may trigger particular sleep stages. Results are evaluated and discussed.

Keywords: Temporal Data Mining, Physiological Data, States, Events.

1 Introduction

Physiological data is a particular kind of clinical data which consist of the measurements over time of some parameters (e.g., blood oxygen, respiration rate, heart rate, etc.) which describe the health conditions of a patient. They can convey relevant information since they reflect the course of a pathology (pathogenesis) or, more generally, the evolution of the physiology of a patient. Clinicians often attempt to understand these data, unearth information hidden within and exploit it in the care process. Interpreting physiological data is thus an activity of vital importance, which can become extremely difficult if not adequately supported. One of the successful strategies for this issue is the integration of information technologies with artificial intelligence tools. A particularly relevant role can be played by data mining techniques [8]. Indeed, clinical data have an informative potential from which new medical knowledge can be gained or

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 26–35, 2009.

knowledge previously acquired can be validated. Moreover, data mining tools can handle vast quantities of records and heterogeneous types of data.

Mining physiological data demands for methods which can properly deal with the temporal dimension of the data. Works reported in the literature can be clustered in two main research lines [12]. In the first line, the mining process aims to gain information about the patient's condition with respect to a limited set of measurements. The problem is usually faced by applying *trend detection* techniques aiming at recognizing patterns (expected trends) occurring in data [6]. The second line deals with data which describe the complete course of the pathology, and the goal is to track each stage of the disease with a pre-existing disease model [1, 5]. However, although these methodologies have proved effective in many scenarios, they still present two problems. First, the techniques based on disease pre-existing models become inapplicable when domain knowledge used to define these models may not be available or not easily accessible. Second, time is a information useful that clinicians can exploit for a deeper comprehension of a disease [7] (e.g., timing of heart failure or the duration of cardiac arrhythmia), but, nevertheless, it is still seldom investigated.

These considerations motivate the current work. In this paper we propose a method which analyzes time-varying physiological data in order to discover knowledge about the possible causes which can determine changes in the physiology of a patient. The method proceeds in two consecutive steps: first, the sequence of relevant stages (*states*) through which the physiology of the patient evolves is identified, then alterations (*events*) occurring in a state which can determine the transition to the next state are detected. A state corresponds to a specific situation of the patient's physiology which holds for a period of time: two consecutive states represent two different situations and together they denote a change in the physiology. The events rather reflect particular physiological variations (e.g., natural developments, drug administration) which together lead the patient from one state to another one.

The paper is organized as follows. In next section some related works are briefly presented and the contribution of the present work is pointed out. After having formulated the problem in Section 3, the method to mine time-varying physiological data is described. It is composed of two steps: the first one aims to identify the states through which the patient's physiology evolves (Section 4), while the second step aims to detect the events which would trigger the transition from one state to the next one (Section 5). In Section 6 the application to the case of Sleep Disorders is explored and some results are discussed. Finally, conclusions close this work.

2 Related Work and Contribution

Analyzing time-varying physiological data is not a novel methodology to understand the course of a disease or, more generally, the physiology of a patient. It has been mainly investigated through the studies of Adlassnig (e.g.,[1]) and Dojat (e.g.,[5]). In the former the purpose is to track the development of a disease

w.r.t. to a disease model which is previously defined by exploiting background knowledge concerning the specific disease. The model is represented in the form of a finite state automaton where the states and the relative transitions denote respectively the expected stages of the disease and the changes from a stage to another one. States and transitions, however, do not handle nor present temporal information with the result that the model-automaton does not permit to acquire time-related knowledge about the stages of the disease.

This is an issue indeed investigated in the works of Dojat [5] which exploit the temporal dimension to recognize a sequence of disease stages (session) in a set of predetermined sequences (scenarios). Sessions and scenarios are represented as temporal graphs where a couple of vertices and their linking edge denote a single stage of the disease. Although this solution permits to gain information about the occurrence and the duration of the stages, it exploits background information (i.e., previously identified scenarios), which, especially for new pathologies, could have been not acquired yet.

Our contribution is quite different and allows i) to interpret the course of a disease with an approach guided only by data and does not rely on domain medical knowledge, ii) to analyze physiological data in order to determine physiological events which can trigger particular states of the disease and iii) to discover temporal, quantitative and qualitative information about states and events: the first expresses the time-interval during which a state (or event) occurs, while the others provide a description (in terms numeric and symbolic respectively) of the physiology of the patient in that period of time.

3 Formulation of the Problem

In order to formulate the problem of interest in this work here we introduce some useful concepts and then describe the problem in formal terms.

Let $P : \{a_1, ..., a_m\}$ be the finite set of real-valued parameters which describe the physiology of the patient. Physiological data consists of a collection Mp of time-ordered measurements of the set P.

An *event* e is defined by the signature $\langle t_F, t_L, Ea, IEa, SEa\rangle$, where $[t_F, t_L]$ ($t_F < t_L$) indicates the time-period of the event (t_F, $t_L \in \tau$[1]). $Ea : \{ea_1, ..., ea_k, ..., ea_{m'}\}$ is a subset of P and contains m' distinct parameters which take values in the intervals $IEa : [inf_1, sup_1]$,..., $[inf_k, sup_k]$,..., $[inf_{m'}, sup_{m'}]$ respectively during the time-period $[t_F, t_L]$. Finally, $SEa : \{sv_1, ..., sv_k, ..., sv_{m'}\}$ is a set of m' symbolic values such that sv_k is associated to ea_k: while SEa is a qualitative representation, IEa is a quantitative description of the event e.

Two examples of events are $e_1 : \langle t_1, t_5, \{bloodoxygen\}, \{[6300, 6800]\}, \{decreasing\}\rangle$ and $e_2 : \langle t_6, t_{10}, \{bloodoxygen\}, \{[6600, 7000]\}, \{increasing\}\rangle$ which can interpreted as follows: e_1 (e_2) is associated with the time-period $[t_1, t_5]$ ($[t_6, t_{10}]$) during which the parameter *blood oxygen* ranges in $[6300, 6800]$ ($[6600, 7000]$) and has a decreasing (increasing) trend.

[1] τ is a finite totally ordered set of time-points. Henceforth, the corresponding order relation is denoted as $\leq$.

A *state* S_j is defined by the signature $\langle ts_j, te_j, C_j, \{sv_1, \ldots, sv_m\}\rangle$, where ts_j, $te_j \in \tau$ $(ts_j < te_j)$ represents the time-period of the state, $C_j : \{f_1, f_2, \ldots\}$ is a finite set of *fluents*, namely, facts in terms of the parameters P that are true during the time-period $[ts_j, te_j]$. The set $\{\mathrm{sv}_1, \ldots, \mathrm{sv}_h, \ldots \mathrm{sv}_m\}$ contains m symbolic values such that sv_h is a high-level description of the parameter $a_h \in P$ during the time-period $[ts_j, te_j]$.

An example of state is $S_1 : \langle t_1, t_{10}, \{$ *blood oxygen in [6500,6700], heart rate in [69,71], respiration rate in [2300,5500]*$\}$, $\{$*blood oxygen is increasing, heart rate is steady, respiration rate is increasing*$\}$ $\rangle$ which can be interpreted as follows: S_1 is associated with the period of time $[t_1, t_{10}]$ and is characterized by the fact (fluent) that the parameters *blood oxygen*, *heart rate* and *respiration rate* range respectively in $[6500, 6700]$, $[69, 71]$, $[2300, 5500]$ and have increasing, steady and increasing trend respectively.

The problem of interest in this work can be thus formulated as follows:

Given: a collection of time-ordered measurements of P, $Mp : \langle Mp_{t_1}, Mp_{t_2}, ..., Mp_{t_n}\rangle$,

Goal: i) identifying a finite sequence S:$\{\mathrm{S}_1, \mathrm{S}_2, \ldots,\}$ of states which represent distinct subsequences of Mp, ii) detecting events $\{\mathrm{e}_1, \mathrm{e}_2, \ldots\}$ which may trigger the transition from S_j into S_{j+1}, $j = 1, \ldots, s-1$, where S_j and S_{j+1} are known.

As a concrete example we consider a collection Mp where $\tau = \{t_1, \ldots, t_{35}\}$. Following the examples above reported, given two states S_1, S_2 associated respectively to the time-periods $[t_1, t_{10}]$, $[t_{11}, t_{35}]$, the events $\{e_1, e_2\}$ occurring during S_1 may trigger S_2 or, in other words, determine the transition of the patient's physiology from S_1 into S_2.

The problem here illustrated is solved through a two-stepped procedure presented in the two following sections.

4 Identification of States

As before introduced, a state $S_j : \langle ts_j, te_j, C_j, \{sv_1, ..., sv_h, ..., sv_m\}\rangle$ can be seen as one of the steps of the disease progression. It is characterized by a duration $([ts_j, te_j])$ and a finite set of fluents, holding in $[ts_j, te_j]$, described in numerical (C_j) and symbolic $(\{sv_1, ..., sv_h, ..., sv_m\})$ terms. In other words a state corresponds to one of the distinct segments of Mp, and this provides us some hints on the approach to follow in order to identify the states. At this aim we resort to our previous work [9] to segment Mp in a sequence of periods of time $([ts_j, te_j])$ and generate the fluents (C_j). Following the algorithm proposed in [9] segmentation splits Mp in a succession of multi-variate segments $([ts_j, te_j])$ which meet two conditions: the variability of each parameter a_i does not exceed a user-defined threshold ω and the duration of each segment has to be greater than a user-defined minimum threshold *minSD*.

This algorithm produces a sequence of segments different of each other guaranteeing that two consecutive segments have different fluents. Therefore, given three consecutive segments, $[ts_{j-1}, te_{j-1}]$, $[ts_j, te_j]$, $[ts_{j+1}, te_{j+1}]$, the fluents C_j

associated to $[ts_j, te_j]$ can be defined as the set of conjunctive conditions, defined on the parameters a_i, which hold in $[ts_j, te_j]$ but neither in the preceding nor in the subsequent time interval ($[ts_{j-1}, te_{j-1}]$ and $[ts_{j+1}, te_{j+1}]$ respectively). The generation of fluents C_j is solved by resorting to the inductive logic programming approach proposed in [10], which is able to find a set of conjunctive conditions $\{f_1, f_2, ...\}$ which logically entails the set $\{Mp_{ts_j}$,..., $Mp_{te_j}\}$ (i.e., $[ts_j, te_j]$) and do not the set $\{Mp_{te_{j-1}}$,..., $Mp_{te_{j-1}}$,..., $Mp_{ts_{j+1}}$,..., $Mp_{te_{j+1}}\}$ (i.e., $[ts_{j-1}, te_{j-1}]$, $[ts_{j+1}, te_{j+1}]$).

Finally, the high-level descriptions $\{\mathrm{sv}_1, \ldots, \mathrm{sv}_h, \ldots \mathrm{sv}_m\}$ of the parameters P are derived through a temporal abstraction technique [11]. Generally, a temporal abstraction technique is defined through a unary or multiple-argument function $\Theta : \Pi \rightarrow \Lambda$ which provides a high-level representation $\lambda \in \Lambda$ of the most relevant features $\pi \in \Pi$ of data. In our case Θ returns, for each parameter a_h, a representation familiar to the clinicians of the slope of the regression line built on the values taken by a_h in the time interval $[ts_j, te_j]$: for instance, the slope values ranging in the interval (0.2..1] are described as INCREASE.

5 Detection of Events

Once the sequence of states has been identified, we look for events which may determine the transition from a state S_j into the next state S_{j+1}. Our assumption is that events which cause the transition should occur during the time interval $[ts_j, te_j]$ (namely, when the patient is in the state S_j) and should not occur in $[ts_{j+1}, te_{j+1}]$ (namely, when the patient is in the state S_{j+1}). Moreover, we assume that natural developments or drug administrations associated to an event e_k are different from those associated to the next event e_{k+1}.

This hypothesis constraints the search and provides us some hints on the approach to follow in order to detect events. The basic idea consists of mining candidate events and, then, selecting from them those more *statistically interesting*.

The technique for mining the candidates proceeds by iteratively scanning the measurements included in the state S_j (i.e., $\{Mp_{ts_j}, \ldots, Mp_{te_j}\}$) and S_{j+1} (i.e., $\{Mp_{ts_{j+1}}, \ldots, Mp_{te_{j+1}}\}$) with two adjacent time-windows which slide back in time. In particular, given a couple of windows (w,w'), where w' immediately follows w, if a candidate is found, then the next candidate is sought for the pair (w'',w), where the new time window w'' has the same size of w (see Figure 1 (a)). Otherwise, the next candidate is sought for the pair (w'',w'), where w'' is strictly larger than w (see Figure 1 (b)). At the end of a single scan a sequence of candidates is mined.

The idea underlying the detection of candidate events for a given couple of windows (w, w') is that parameters in P responsible of a transition should also affect remaining parameters. In other terms, a dependency among parameters in P should be observed. Therefore, the following method is applied in order to determine the set of candidate events. For each parameter a_i the two multiple linear regression models computed on the remaining parameters in P are built

$$a_i = \beta'_0 + \beta'_1 a_1 + \ldots + \beta'_{i-1} a_{i-1} + \beta'_{i+1} a_{i+1} + \ldots + \beta'_m a_m,$$
$$a_i = \beta''_0 + \beta''_1 a_1 + \ldots + \beta''_{i-1} a_{i-1} + \beta''_{i+1} a_{i+1} + \ldots + \beta''_m a_m,$$

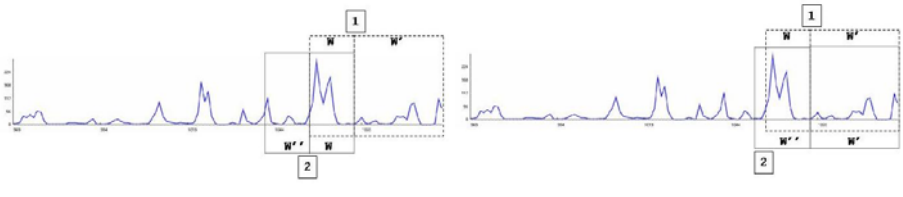
by considering the distinct data samples in w and w' respectively. The couple of regression models for which information loss in predicting a parameter a_i is the lowest are selected. The subset $Ea = \{a_j \in P - \{a_i\}\}$ such that the difference between β'_j and β''_j is greater than an automatically determined threshold σ_j is selected and associated with the time window $w : [t_F, t_L]$ of the detected event.

The set Ea is filtered in order to remove those parameters for which no interval of values can be generated which discriminates samples in w from samples in w'. In particular, for each $a_j \in Ea$ the interval $[inf_j, sup_j]$ is computed by taking the minimum (inf_j) and maximum (sup_j) value of a_j in w. If $[inf_j, sup_j]$ is weakly consistent with respect to values taken by a_j during the time window w' then a_j is kept, otherwise it is filtered out. Weak consistency is computed as the weighted average of the zero-one loss function, where weights decreases proportionally with the time points in w'. Finally, the filtered set of m' parameters will be associated with a set of intervals $\{[inf_1, sup_1], \ldots, [inf_r, sup_r], \ldots, [inf_{m'}, sup_{m'}]\}$, which correspond to the quantitative description IEa of an event.

The corresponding qualitative definition SEa of the event is a set of symbolic values $\{sv_1, \ldots, sv_r, \ldots, sv_{m'}\}$, one for each selected parameter $ea_r \in Ea$, which are determined through the same technique of temporal abstraction used for the high-level description of the states.

Among the set of generated candidate events, we select the most statistically interesting ones as those events which are *most supported*: we are interested in the most supported events since they correspond to the sequence of events $\{e_1, e_2, \ldots\}$ that solves the scientific problem formulated in Section 3.

An event e_u is called *most supported* if meets the following two conditions: i) there exists a set of candidates $\{e_1, e_2, \ldots, e_t\}$ which contains the same information of e_u, that is: $\forall e_q, q = 1, \ldots, t$, $e_q \neq e_u$, the set of parameters associated to e_q includes the set of parameters associated to e_u, the time interval associated to e_q includes the time interval associated to e_u and, finally, the set of symbolic values and intervals of values associated to parameters coincide; ii) no event e_v exists whose information is contained in a set of candidates $\{e_1, e_2, \ldots, e_{t'}\}$ with



(a) (b)

Fig. 1. Mining the candidate events: if the event is found for the pair (w,w'), then the next candidate is sought for (w'',w) (a), otherwise it is sought for the pair (w'',w') (b)

$|\{e_1, e_2, \ldots, e_{t'}\}| > |\{e_1, e_2, \ldots, e_t\}|$. The *support* of the event e_u is computed as follows: let $\{e_1, e_2, \ldots, e_z\}$ be the set of candidates such that the time interval associated to each of them contains that one of e_u, and $\{e_1, e_2, \ldots, e_t\}$ be the set of candidates as described at the point i), then the support of e_u is $supp(e_u) = (t+1)/z$.

6 Application to Sleep Disorders

Sleep disorders are issues of great importance and widely investigated in medicine because some serious diseases are accompanied by typical sleep disturbances (e.g., daylight sleepiness), and this attracts the interest of the community of Artificial Intelligence (e.g., [4]). The possible influence of physiological alterations on sleep disorders motivates rather our interest in this scenario. In particular, we apply the proposed approach to acquire knowledge about cardiovascular and breathing disorders (events) which may determine the change from a physiological stage (state) to another one during sleep. It is worth of pointing out that our aim is not to characterize what is not given or known (prediction) but to characterize what is observed (description): in other words, our application aims determining what can lead to a change of state of the sleep and this is done through a retrospective analysis, namely when the sleep process is completed.

Dataset Description. Experiments concern the polysomnography (measurements of physiological parameters during sleep) of only one patient observed from 21.30 p.m. to 6.30 a.m.. The dataset has been sampled at 1 second and is publicly accessible at PhysioBank site[2] [3]. Physiological parameters are *eeg* (electroencephalogram), *leog*, *reog* (electrooculograms), *emg* (electromyogram), *ecg* (electrocardiogram), *airflow*, (nasal respiration), *thorex* (thoracic excursion), *abdoex* (abdominal excursion), *pr* (heart rate) and *saO2* (arterial oxygen saturation): *emg, ecg, airflow, thorex, abdoex, pr, saO2* can describe possible cardiovascular and breathing disorders, while *eeg, leog, reog* can denote the stage of the patient during sleep.

Results. Experiments were conducted following the two-stepped procedure described in the sections 4 and 5: because of space limitations we detail only the experiments on the first 5040 measurements (from 21.30 p.m. to 23.30 a.m.).

In the first step, by varying the threshold *minSD* it is possible identifying states with different durations and this allows to analyze data at several time granularities. Indeed, from Table 1 we known that by setting *minSD* to 60, 120 and 300 physiological states with duration of respectively one, two and five minutes can be identified: for instance, in the experiment 2, when ω is 10% and *minSD* is 120, a sequence of 28 states is found out, or, in other words, the physiology of the patient evolves through 28 stages with duration of at least 120 seconds. In the second step events were detected for each couple of consecutive states (S_j,S_{j+1}) of each experiment: some of events are described in the following.

[2] http://www.physionet.org/physiobank/

Table 1. Results of the first step: ω denotes the maximum variability (in percentage) of the parameters, *minSD* is the minimal duration of the states, |S| indicates the total number of found states

Experiment	minSD (s)	ω (%)	|S|
1	60	10	31
2	120	10	28
3	300	10	7
4	60	20	31
5	120	20	28
6	300	20	7
7	60	50	31
8	120	50	28
9	300	50	7

Table 2. Signatures of some states found out in the experiment 5 (Table 1). The facts C_j are represented in *Datalog* language which is adopted in [10].

States	[ts_j,te_j]	Facts	High-Level Description
S_{22} (experiment 5)	[ts_{3895}, te_{4060}]	f_1: eeg in [5.882,34.314], leog in [3.922,87.255], reog in [-4.902,46.078] ... $\wedge$ f_2: reog in [14.706,83.333], leog in [-47.059,87.255], eeg in [-6.863,33.333]	eeg: STEADY, leog:INCREASE, reog: INCREASE
S_{23} (experiment 5)	[ts_{4061}, te_{4224}]	... $\wedge$ f_6: leog in [-5.882,-1.961], reog in [-6.863,-0.98], eeg in [-1.961,0.0] ... $\wedge$ f_{11}: reog in [-21.569,-2.941], eeg in [-3.922,-1.961], leog in [-21.569,0.0]	eeg: STEADY, leog:STEADY, reog: STEADY

An interesting result concerns the transition from S_{22} to S_{23} (Table 2) of the experiment 5 (Table 1). During the first state the patient's physiology shows high variance of *leog*, *reog* and peak values of *eeg*, while in the next one the values of *eeg* become milder and *leog*, *reog* have low variance. Moreover, the trend of *leog*, *reog* become stable in S_{23} w.r.t. to the increase showed in S_{22}. According to our analysis, this change can be ascribed to a succession of physiological events occurring during S_{22} (see Table 3): e_1 with duration 20 seconds ([ts_{3998}, te_{4018}]) followed by e_2 with the same duration ([ts_{4019}, te_{4039}]) and finally e_3 ([ts_{4040}, te_{4060}]). Breathing and cardiovascular disorders associated to e_1 are different from those of e_2: in e_1 we know that during [ts_{3998}, te_{4018}], *abdoex* varies in [-1.412,0.722] and has strong increasing tendency, and $airflow$ has values in [-2.322,3.482] and strong decreasing trend. Instead, during [ts_{4019}, te_{4039}] (i.e., e_2), *abdoex* is stable in [-1.663,1.443], *saO2* varies in [94.013,95.012] and has decreasing trend, and *emg* has strongly decreasing tendency and ranges in [-0.973,0.243]. In e_3 the trend of *emg* rather becomes strongly increasing although its range is slightly wider. Thus, we interpret that e_1,e_2,e_3 occur during S_{22} and may trigger S_{23}. Table 3 reports also the support of the discovered events: for instance, e_1 occurs in a set of three events out of a total set of three candidates (3/3) detected in [t_{3998},t_{4018}].

Discussion. In this paragraph, we try to evaluate the results through a qualitative and quantitative analysis, although we are aware that a deeper discussion can be performed in conjunction with clinicians.

Table 3. Events detected for the transitions $S_{22} \rightarrow S_{23}$. The columns report the specification of the events identified in the column # Event.

Transition	# Event	Ea	$\langle \ldots, [inf_k, sup_k], \ldots \rangle$	$\langle \ldots, sv_r, \ldots \rangle$	$[t_F, t_L]$	support
$S_{22} \rightarrow S_{23}$	e_3	emg	[-0.973,0.486]	STRONG INCREASE	$[t_{4040}, t_{4060}]$	3/3
	e_2	saO2	[94.013,95.012]	DECREASE	$[t_{4019}, t_{4039}]$	4/4
		emg	[-0.973,0.243]	STRONG DECREASE		
		abdoex	[-1.663,1.443]	STEADY		
	e_1	airflow	[-2.322,3.482]	STRONG DECREASE	$[t_{3998}, t_{4018}]$	3/3
		abdoex	[-1.412,0.722]	STRONG INCREASE		

A first observation allows us to qualitatively confirm the result above illustrated. Indeed, the researchers of Sleep Heart Health Study [3] pointed out that from t_{4050} to t_{4140} the stage of patient sleep goes from "Awake" shortly to "Stage-1" and then to 'Stage-2", which would correspond to the transition from S_{22} to S_{23}.

Quantitatively, results have been evaluated by estimating the accuracy in terms of sensitivity-specificity of the most supported events. Given the complex notion of event presented in this work, we defined a notion of *True Positive* (necessary for the calculation of sensitivity-specificity) which is formulated on the properties of the examined scenario, just as literature suggests [2].

In this work, an event e is called True Positive if meets the following conditions: i)$\exists$ $\{ea_1,\ldots,ea_k, \ldots, ea_{m''}\} \subseteq \{ea_1,\ldots,ea_r,\ldots,ea_{m'}\}$ $\ni$' each ea_k= emg $\vee$ ecg $\vee$ airflow $\vee$ thorex $\vee$ abdoex $\vee$ pr $\vee$ saO2 , where $|\{ea_1,\ldots,ea_k,\ldots,ea_{m''}\}| > |\{ea_1,\ldots,ea_r,\ldots,ea_{m'}\}|/2$ and ii) e is a most supported event. Informally speaking, an event is True Positive whenever it is *most supported* and most its parameters represent breathing or cardiac cycle (emg, ecg, airflow, thorex, abdoex, pr, saO2).

Here we report only the sensitivity of the events detected for the transition $S_{22} \rightarrow S_{23}$ while varying the minimal duration. By analyzing Table 4 it emerges the difficulty to recognize the better value of accuracy: indeed, possible events responsible of changes can hold for few seconds (25 s) or few minutes (125 s). This strengthens the importance and our interest to investigate the aspect of temporal variability when interpreting the pathology of a patient.

Table 4. Sensitivity of events detected for the transition $S_{22} \rightarrow S_{23}$ of the experiment 2

sensitivity (%)	minimal duration of events(s)
67	25
71	50
68	75
59	100
70	125
71	150

7 Conclusions

In this paper we investigated the problem of interpreting the physiology of a patient over time with a data-driven approach which does not consider background

medical knowledge. At this aim we proposed a two-stepped method whose peculiarity is that it discovers the possible causes which can determine the transition from a relevant physiological stage to another one. Basic assumption is that mild changes, which occur in a stage and do not in the next one, can be responsible of the transition from a stage to the other one. In this work mild changes are attributed only to variations of physiological parameters and no one external factor is taken into account.

We explored the application of the method to the scenario of sleep diseases in order to discover what events, in terms of breathing and cardiovascular disorders, can determine changes of the stage of the patient sleep. Results evaluation has been performed through an usual statistical technique. However, we are hopeful that in future domain experts or clinicians can help us to discuss the results and evaluate the usefulness of our approach.

Acknowledgments

This work is supported in partial fulfillment of the project "Ateneo 2009: Modelli e metodi computazionali per la scoperta di conoscenza in dati biomedici".

References

1. Adlassnig, K.P.: Fuzzy Systems in Medicine. In: Proc. of the International Conference in Fuzzy Logic and Technology, pp. 2–5 (2001)
2. Fawcett, T., Provost, F.: Activity Monitoring: Noticing Interesting Changes in Behavior. In: KDD 1999, pp. 53–62 (1999)
3. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23), e215–e220 (2000)
4. Guimares, G., Peter, J.H., Penzel, T., Ultsch, A.: A method for automated temporal knowledge acquisition applied to sleep-related breathing disorders. Artificial Intelligence in Medicine 23(2), 211–237 (2001)
5. Guyet, T., Garbay, C., Dojat, M.: Human/Computer Interaction to Learn Scenarios from ICU Multivariate Time Series. In: AIME, pp. 424–428 (2005)
6. Haimowitz, I.J., Le, P.P., Kohane, I.S.: Clinical monitoring using regression-based trend templates. Artificial Intelligence in Medicine 7(6), 473–496 (1995)
7. Kahn, M.G.: Modeling Time in Medical Decision-support Programs. Medical Decision Making 11(4), 249–264 (1991)
8. Lavrac, N., Zupan, B.: Data Mining in Medicine. In: The Data Mining and Knowledge Discovery Handbook, pp. 1107–1138 (2005)
9. Loglisci, C., Berardi, M.: Segmentation of Evolving Complex Data and Generation of Models. In: ICDM Workshops, pp. 269–273 (2006)
10. Malerba, D.: Learning Recursive Theories in the Normal ILP Setting. Fundam. Inf. 57(1), 39–77 (2003)
11. Shahar, Y.: A Framework for Knowledge-Based Temporal Abstraction. Artif. Intell. 90(1-2), 79–133 (1997)
12. Uckun, S.: Intelligent systems in patient monitoring and therapy management. A survey of research projects. Int. J. Clin. Monit. Comput. 11(4), 241–253 (1994)

Severity Evaluation Support for Burns Unit Patients Based on Temporal Episodic Knowledge Retrieval*

Jose M. Juarez[1], Manuel Campos[1], Jose Palma[1], F. Palacios[2], and Roque Marin[1]

[1] Computer Science Faculty – Universidad de Murcia – Spain
{jmjuarez,mancuelcampos,jtpalma,roquemm}@um.es
[2] Intensive Care Unit – Universitary Hospital of Getafe – Spain
franciscodepaula@gmail.es

Abstract. Severity scores are a sort of medical algorithm commonly used in medicine. In practise, physicians only use a few of them, usually internationally accepted ones involving very simple calculations. However, their daily use in critical care services gives rise to two potential problems. First, they do not always cover the particularities of the local population or a specific pathology may not be considered in the score. Second, these services (e.g. intensive care units or Burns Units) are strongly dependent on the evolution of the patients and, so the temporal component plays an essential role that should always be in mind. On the other hand, the knowledge required is at least partially present in the physician team of the medical unit due to the experience gained in treating individual patients, that is, in the form of episodic knowledge. Therefore, the use of techniques based on analogy reasoning, such as Case-Based Reasoning, would seem a suitable approach for dealing with part of this problem.

In this work, we present an episodic knowledge retrieval system to support the physician in evaluating the severity patients from the temporal evolution point of view. To this end, we present different techniques for temporal retrieval based on previous works on temporal similarity. We also demonstrate the suitability of this system by applying it to a specific medical problem arising in a Burns Unit.

1 Introduction

Critical care scores are helpful tools for intensivists to obtain an objective indicator of the severity of the patient [2]. Severity scores are a sort of medical algorithm to standardise decisions, strongly recommended by evidence-based medicine approaches. In practice, Intensive Care Unit (ICU) physicians use only a few such scores, often those involving very simple calculations and internationally accepted (e.g. SAPS-II, APACHE or SOFA) [8].Despite the advantages of these medical tools, their application in daily ICU practice frequently gives rise to difficulties. First, ICU domains are strongly dependent on the evolution of the patients and, therefore, the temporal component plays an essential role that should always be present in the scores. In this sense, current efforts in the field, such as [12] with regards to the SOFA score, point to the need for data-driven approaches by temporal patterns. Secondly, severity scores are the

* This study was partially financed by the Spanish MEC through projects TIN2006-15460-C04-01, PET2006_0406,PET2007_0033, and the SENECA 08853/PI/08.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 36–45, 2009.

result of international trials, and they do not always fit the particularities of the local population or the regional variations of a given pathology. For instance, the relevance of the SAPSII parameter is poor for severe burn patients with infections, since it is a general severity score and therefore not specific enough for this kind of population. On the other hand, clinical trials have demonstrated that the temporal evolution during the initial days is esential for the survival of the severe burned patients [2].

Unlike these scores, the required knowledge is partially present in the physician team of an ICU service in the form of episodic knowledge, that is, members' experience of the medical care of each particular patient that they have treated. Therefore, we consider that Artificial Intelligence techniques based on analogy reasoning, such as Case-Based Reasoning (CBR),is a suitable approach to develop tools to objectivise part of the tacit knowledge inherent in each particular ICU service. The key idea of CBR is to tackle new problems by referring to similar problems that have already been solved in the past [3], that is, CBR is based on individual experience in the form of cases. Generalization beyond such experience is largely founded on principles of analogical reasoning in which the (cognitive) concept of similarity plays an essential role.Therefore, the two main elements that should be considered in any CBR system are: the cases (the episodic knowledge) and the similarity measures (to quantify the analogy between cases).

In this sense, the CBR approach has been demonstrated to be a useful and effective methodology for developing decision support systems in the medical domain [9].

Another element that should be considered in many fields of medicine is the temporal dimension. Time is omnipresent in almost all medical scenarios (e.g. ICUs and Burns Units) but also in medical informatic approaches (diagnosis methods, time granularity, guidelines, ...). Therefore, the representation of time in cases and measures of similarity must be carefully treated.

Related proposals describe in the literature deal with the development of CBR systems in medicine, and have obtained relative success in medical service, such as [4] in oncology.However further efforts must be done in two main directions (integration into the information system and time management).

According to [9], despite the importance of the temporal dimension, the study of the impact of time on the CBR systems is still an open problem. Some advances in the CBR community in this field have mainly focused on the study of the temporal data from biosignals (mainly ECG and O_2 in blood) and processing these time series by information reduction. The proposal described in [10] is a physiological analysis system based on heart and pulmonary signal patterns using similar techniques. From a more theoretical perspective, in [11] a CBR system is proposed for the temporal abstraction of dializer biosignals in the form of time series. However, in many domains, information is usually presented as heterogeneous (numerical and nominal) and cases are composed of a set of events of different nature over time. For instance, the electronic health record (EHR) of patients describe the symptoms, tests and treatments of a patient during their stay in hospital. Heterogeneous event sequences cannot use time series similarity techniques and new proposals are required to compare this kind of cases.

In this work we tackle the problem of developing medical decision support based on CBR, integrated in the medical information process and managing the temporal dimension. Far from proposing an ICU diagnosis system, we present an episodic knowledge

system (T-CARE) to be used by the physician to help in the severity evaluation of patients from the temporal evolution point of view. To this end, we present different techniques for temporal retrieval based on previous works on temporal similarity [1,5], the implemented systems and their validation for the particular problem of patient survival in the Burns Unit after 5 days. The structure of this work is as follows. Section 2 introduces the representation of time and the episodic knowledge.Knowledge retrieval process by similarity measurements is analysed in Section 3. In Section 4, the T-CARE system is presented and related experiments in specific medical care of a Burns Unit are described. Finally, conclusions and future works are described in Section 5.

2 Episodic Knowledge: Temporal Cases

Cases describe the knowledge acquired after solving specific problems. They can be considered the atomic elements of the knowledge bases in a CBR system. The generalized idea is that a case is essentially composed of the 3-tuple *problem, solution, outcome*. However, the comments of experts made during meetings and our previous experience in the development of knowledge-based systems in medical domains suggest that the cases structure can be extended. On the one hand, a case may be subjective. Different physicians treat the same patient and each can define a case that describes the temporal evolution of a patient. Hence, it is obvious that the clinical case described by a veteran and the same case as seen by a first-year resident could vary. On the other hand, the description of a case is also sensitive to the purpose for wich it is intended for. That is, cases describing the same clinical health record may well pay special attention to different aspects of the state of the patient. For example, one case might focus on a study of ischaemic cardiopathy and a second case (of the same patient) might focus on the economic impact of long-term ICU patients.

Therefore, we define the concept of episodic knowledge, using description logics, as follows:

$$EpisodicKnowledge \doteq \exists HAS.Observer \sqcap \exists HAS.Descript \sqcap \exists HAS.Target$$
$$Descript \doteq \exists HAS.Context \sqcup \exists HAS.TCase \sqcup \exists HAS.Solution \sqcup \exists HAS.Outcome$$
$$TCase \doteq \{Event\}$$
$$Event \doteq \exists HAS.Domain \sqcap \exists HAS.Concept \sqcap$$
$$\exists HAS.Manifestation \sqcap \exists HAS.Attribute \sqcap \exists HAS.Value \sqcup HAS.Time$$
$$Time \doteq TimePoint \sqcup Interval$$

where the description of the episodic knowledge could be composed by its context (e.g. antecedents), the temporal information (TimePoint being a time-stamp), the solution, and the outcome (the success of the patient's aspects analysed). This modelling allows different kinds of temporal information to be described for the same patient. For instance:

$$event_1 =< ICU, BloodPressure, Arterial, Sistolic, 145/85mmHg, 05/Feb >$$
$$event_2 =< ICU, Propofol, Given, Dosis, 3\%, < 03:30, 21:15 >>$$

Temporal cases are traditionally represented by a set of temporal features, defining time series and temporal event sequences. In the particular situation where these features are not homogeneous (i.e. combination of qualitative and quantitative information), it is difficult for systems to retrieve similar cases.

Considering the temporal nature of events, we identify two different scenarios: sequences of time points (hereinafter event sequences) and those composed by intervals (hereinafter interval sequences). These scenarios have a direct match on the EHR. For instance, the set of tests carried out on a patient could be considered an event sequence, whilst the therapeutic administration (usually intravenous in the ICU) describes an interval sequence.

Formally, we can define an event sequence as follows:

$$\forall e \in E, \quad e = (d, c, m, a, v, t), t \in TimePoint \tag{1}$$

$$ES = \{es | es =< e_1, \ldots, e_n > \wedge e_i \in E \wedge \forall e_i, e_j \ (i < j) \ t_i \leq t_j\}. \tag{2}$$

where E is the universal set of events(e); d, c, m, a, v, and t the domain, context, manifestation, attribute, value and time of a temporal event e; ES is the universal set of event sequences; and $es \in ES$ is an ordered set defining a sequence of temporal events.

As examples of this kind of sequence we present the following cases (see Figure 1-A), which are part of the medical history of two patients from an Intensive Care Unit (ICU).

In the same way, the interval sequence scenario is defined as follows:

$$\forall i \in I, \quad i = (d, c, m, a, v, t), t = (t^-, t^+), \ t \in Interval \tag{3}$$

$$IS = \{is | is =< i_1, \ldots, i_n > \wedge i_j \in I \wedge \forall e_j, e_k \ (j < k) \ t_j^- \leq t_k^+\}. \tag{4}$$

For instance, Figure 2-A depicts part of the therapy administration of two patients from a Burns Unit.

3 Temporal Case Retrieval

In measuring the similarity between temporal cases, CBR is traditionally based on the definition of similarity functions. In this work, we adopt the binary and normalized similarity function (σ), that is: $\sigma : Case \times Case \rightarrow [0, 1]$.

In the aforementioned scenarios, temporal similarity measures (such as [6]) cannot directly apply the efficient time series techniques, but require new approaches to deal with these heterogeneous sequences.

To this end, recent proposals have focused on direct matching between pair of features within sequences, mainly based on classical distances (such as the Euclidean distance).Alternatively, temporal constraint networks (TCN) have proved to be useful tools for temporal representation and reasoning, and can easily be extended to managing imprecision and uncertainty [7]. In this work, we consider previous works in event sequence similarity [5] and in interval sequence as the result of the collaboration with Combi et al [1].

3.1 Event Sequence Similarity

When two temporal scenarios defined by event sequences are to be compared, two similarity functions are considered: one based on Euclidean distance and the other based on TCN. The temporal Euclidean function is a classical approach where $\sigma_{euc}(es_a, es_b) = 1/n * \sum_{j=1}^{n} w_j/max\_d * \sqrt{(e_j^a - e_j^b)^2}$. However, we also consider the TCN-based function proposed in [5]. In short, the approach consists of merging two event sequences in a single TCN, able to represent imprecision. The uncertainty produced by this merging step can be measured and considered as an indicator of the diffference (inversely proportional to similarity) between the event sequences (see Figure 1). This temporal

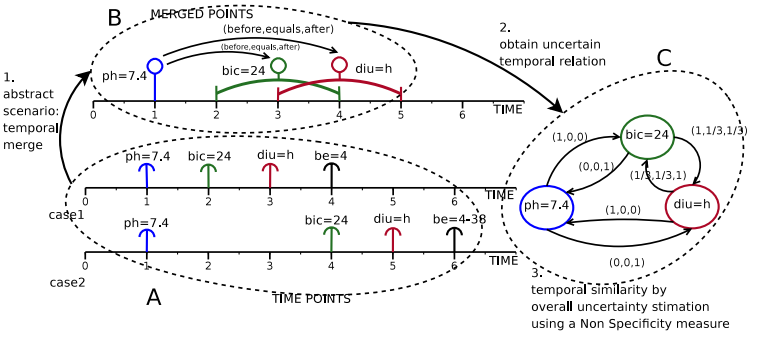


Fig. 1. Example of S_{ptcn}: from two real cases at Burns Unit defined by two event sequences. ph= skin ph level, bic=bicarbonate level, diu=diuresis volume, be=base excess.

scenario is modelled by the PTCN (Possibilistic Temporal Constraint Network) formalism, described in [5] and based on Possibility Theory. A PTCN is obtained from the set of temporal evidences of the input cases, considering each pair of matched events of the cases as a single merged point of the PTCN (Figure1-B). Each temporal relation between two merged points of the PTCN is a 3-tuple $(\pi_<, \pi_=, \pi_>)$ that describes the possibility degree of the first merged point to be before ($<$), at the same time ($=$) or after ($>$) the second merged point.

The overall uncertainty is measured using Non-Specificity (U) [7], a sort of Shanon-Entropy measure for possibility distributions (Figure1-C). The similarity measure is calculated from the U function as follows:

$$\sigma_{PTCN}(c, c') = 1/\sum_{r_{ij} \in R} U(r) = 1/\sum_{r_{ij} \in R} (-\sum_{k=1}^{3} \pi_k^{r_{ij}} - \pi_{k+1}^{r_{ij}} log_2 k). \quad (5)$$

where r_{ij} is a relation of the network and $r_{ij} = (\pi_1, \pi_2, \pi_3)$, that is, a possibility distribution describing the possibility degree of i to be before, at the same time, or after j.

3.2 Interval Sequence Similarity

The aforementioned similarity function does not consider the semantic of intervals, discarding some important aspects of the temporal information available. The extension of this similarity measure from time point algebra (the 3 relations $<$, $=$ and $>$) to interval algebra (13 relations between intervals) implies the development of time costly algorithms and, therefore, a specific similarity function is required. To evaluate temporal similarity between interval sequences, we consider the proposal described in [1] for workflow cases that can be easily adapted to any kind of interval scenario. The idea consist of: (i) comparing corresponding interval events of both sequences (Figure 2-1); (ii) comparing qualitative/quantitative temporal relations between corresponding interval events (Figure 2-B); and (iii) considering the presence/absence of some interval events (Figure 2-2).

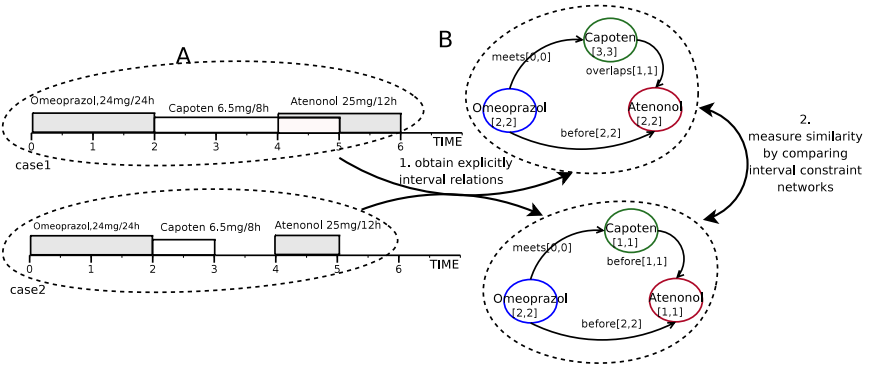


Fig. 2. Example of S_{ABIN}: from two real ICU cases of treatment defined by two interval sequences

The given interval sequences are represented through TCNs (Figure 2-B), and the similarity is evaluated by considering the distance between intervals representing corresponding interval events, and the distance for the relations between corresponding interval events.

The similarity function is a linear combination of functions to compare individually interval events and to consider the relations with respect to other interval events as follows:

$$\sigma_{ABIN}(c, c') = 1 - (w_1 \frac{\sum_{\forall N,N'} d_n(n, n')}{|N \cup N'|} + w_2 \frac{\sum_{\forall R,R'} d_r(r, r')}{|R \cup R'|} + w_3 abs(N, N')) \quad (6)$$

where N, N', R and R' are the set of nodes and relations obtained from cases c and c' respectively. Whilst d_n, d_r and abs are the normalised functions to compare nodes, relations and the absence of nodes, respectively.

4 T-CARE System: Implementation and Experiments

The Temporal Care Retrieval System (T-CARE) retrieves similar cases of patients for medical decision support by searching in a case library for patients with temporal

evolution. In T-CARE, the following tasks can be identified: (1) the acquisition of temporal cases from the EHR of a HIS and storage in the Temporal Case Library (TCL); and (2) the retrieval of similar temporal cases from the TCL. Figure 3 shows the main components of the T-CARE architecture.

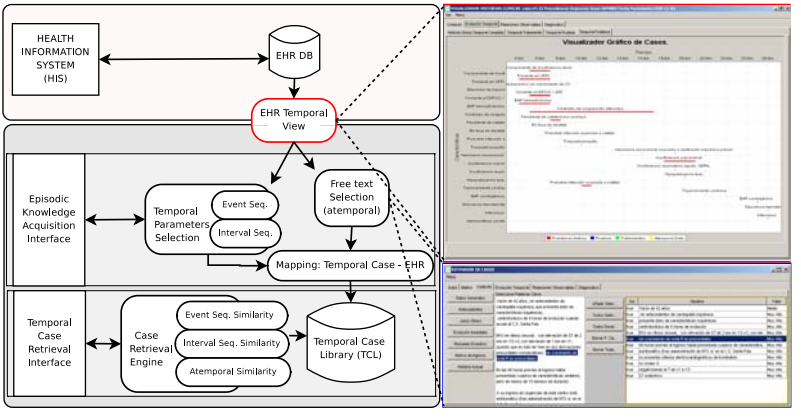


Fig. 3. $T-CARE$ architecture: Temporal CAse REtrieval System

4.1 Episodic Knowledge Acquisition

The case acquisition process consists of the semiautomatic description of the elements (clinical tests, treatments, diagnoses, ... etc.) that make up the case. In medicine, this process is quite common since a clinical session focused on a particular medical problem presents only the relevant information of the EHR of a patient in the HIS. Therefore, in T-CARE the case acquisition process is guided by the episodic knowledge description (see Section 2), where the expert must describe the goal of the case ($Target$), the clinician who describes the case ($Observer$) and the description of the case (the atemporal and temporal elements). Figure 3-right depicts the temporal case acquisition tool, implemented following these steps. This tool uses a temporal viewer, allowing the expert to obtain a summarized view of the temporal evolution of the EHR of a patient. The physician can select those elements of the EHR which are relevant to the case, allowing the physician to associate a *relevance degree* using linguistic labels (from *very relevant* to *irrelevant*) to each selected element. The temporal case acquisition automatically matches the EHR elements and the components of the episodic knowledge.

4.2 Retrieval Engine

The first phase of the T-CARE retrieval step consists of discriminating all episodic knowledge instances by considering the observer and target components and, therefore, reducing the space of search. Secondly, temporal similarity measures are used to compare the input case with the retrieved cases of the TCL. The different similarity measures used deal with the different facets of the temporal dimension (time point

events using function σ_{PTCN} and intervals using function σ_{ABIN}). Thanks to the temporal case acquisition tool, each element of the case (context elements and temporal and atemporal attributes of the problem) has a degree of relevance using a linguistic label that can be assigned to a normalized numeric value. These relevance degrees, therefore, acts as weights in order to quantify the general similarity of the different aspects of cases. In T-CARE, the components of the episodic knowledge can be classified as: atemporal information (context and atemporal attributes of the case) and temporal information (temporal attributes of the case).

Let $c \in C$ be an input case and $c' \in TCL$ be a case of the TCL. Moreover, let suppose that the atemporal attributes $A = \{a_1, ..., a_x\}$, the temporal events $es =< e_1, ..., e_y >$ and the intervals $is =< i_1, ..., i_z >$ are part of case c. A Let use also suppose that A', es' and is' be part of case c', where w^a, w^e, $w^i \in [0, 1]$ are the corresponding relevance degrees of the atemporal, temporal events and intervals of the case c'. Thus, in T-CARE the atemporal similarity measure between the cases c and c' is defined as follows:

$$\sigma_{atemp}(c, c') = 1 - \frac{\sum_{i=1}^{n} w_i^a d(a_i, a_i')}{n} \tag{7}$$

being d the normalised Euclidean distance for quantitative attributes or a dichotomic function between qualitative attributes.

Finally, the overall similarity measure between two cases is defined as follows:

$$S(c, c') = \frac{w^a \sigma_{atemp}(A, A') + w^e \sigma_{PTCN}^{N}(es, es') + w^i \sigma_{ABIN}(is, is')}{w^a + w^e + w^i}. \tag{8}$$

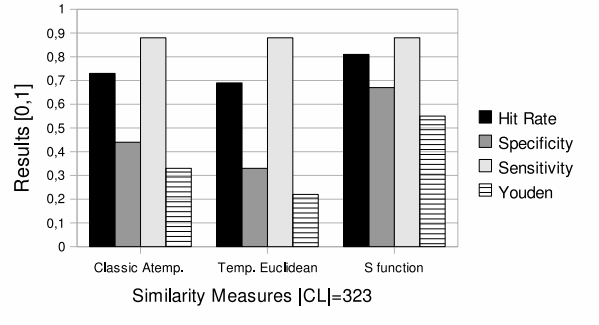


Fig. 4. Study of different similarity measures in Burns Unit

4.3 Experiments

Patients from the Burns Unit of the ICU are critical long-term patients. Clinical trials shows that demographic data and the temporal evolution of certain evidences during the first 5 days are essential for assigning a patient *survival profile*. Some of these evidences are: age, gender, depth of burn injuries, or the 5 days evolution of diuresis, level of bicarbonate, skin ph, and acidosis level. The population studied was taken from the EHR of patients between 1992 and 2002 in a Burns Unit. The episodic knowledge is an

initial set of 375 temporal cases, selected by the physicians using the temporal case acquisition tool. This population is composed by 257 males/ 118 females, has an average age of 46.2 years and has an average stay in the Burns Unit of 25.8 days. In this population, the SAPSII severity score was discarded, due to the high rate of false possitive of survival (over 80%). Therefore, the use of T-CARE seems a useful complement to support physicians to evaluate severity, considering the temporal evolution of the first 5 days of the stay.

The aim of this experiment, therefore, is to configure T-CARE in order to state the patient's survival, considering the evolution of the evidences during the first 5 days and the demographic data of individual cases. The experiment was carried out considering 52 input cases of the initial set (selected randomly), while 323 temporal cases make up the TCL where survival is known. Once the TCL is defined, the system requires the calibration of the similarity measure. To this end, the attribute weights of the global similarity measure must be assigned, stating the relevance of each parameter with respect to patient survival. According to this, we consider two strategies to state these weights: (1) obtaining the medical knowledge provided directly by the physicians; and (2) using arbitrary techiques to obtain information, that is, considering the discrimination capability of each parameter to state the survival of the TCL patients. According to the second strategy, we obtained a classification of parameters by using domain-independent techniques. In particular, we selected the Mutual Information measure (MI), based on Shanon Entropy (H) [7].

In order to compare the capability of similiarity measures to correctly classify the survival of patients, this experiment was tested using three different similarity functions. Firstly, we considered a classical CBR approach, that is, ignoring the temporal component. Secondly, considering the temporal dimension by using the temporal Euclidean distance σ_{euc}. Finally, we used the S function described in Expression 8. To evaluate the experiments, we considered the following factors: hit rate (accuracy), specificity, sensitivity and Youden index (test quality indicator common in medicine, $Specificity + Sensitivity - 1$). The results of the experiments (summarised in Figure 4) show the advantages of the different similarity measures described. Experiments using function S provided the best results considering the hit rate (over 80%). In our experiments, the sensitivity keeps the same value. In this sense, we can conclude that there is a differentiated set of survival cases detected by any similarity method. According to the physician team, specificity is an essential factor for these survival tests. Figure 4 shows the behaviour of specificity values. Note that a simple management of the temporal dimension does not always improve this value (e.g. Temporal Euclidean) and it requires of more sophisticated temporal techniques. Therefore, it is worth mentioning the substantial increment in the specificity value (0.67) using S, since S function combines time point and interval reasoning.

5 Conclusions

The aim of this work is to provide support for evaluating severity in Burns Unit patients by retrieving temporal similar cases solved in the past. To this end, we represent the knowledge required by the *episodic knowledge* definition, an *ad hoc* extension of

temporal cases for medical domains. A second contribution of this work is T-CARE, a temporal case retrieval system that searches similar episodic knowledge instances by using of temporal similarity measures proposed in previous works [1,5]. Finally, in order to demonstrate the suitability of the proposal and the implemented system, we carry out some experiments to solve a real life medical problem in the Burns Unit. Previous studies on intensive care domains also deal with the temporal dimension [10,11], but mainly focused on processing biosignals based on time series techniques. Unlike those, our work manages the temporal event and interval sequences obtained from the EHR, where time series techniques cannot be used and new similarity functions are required. The T-CARE experiments carried out with different kinds of similarity measure reveals the advantages of combining event sequence and interval sequence similarity measures since the system is able to manage different aspects of the temporal information available. Future works will be focused on improving different aspects of T-CARE: time cost reduction by indexing methods, and the accuracy of the solution, specificity and sensitivity by combining other techniques such as model-based reasoning.

References

1. Combi, C., Gozzi, M., Juarez, J.M., Marin, R., Oliboni, B.: Temporal similarity measures for querying clinical workflows. Artificial Intelligence in Medicine 46(1), 37–54 (2009) (in press)
2. Galeiras, R., Lorente, J.A., Pertega, S., Vallejo, A., Tomicic, V., de la Cal, M.A., Pita, S., Cerda, E., Esteban, A.: A model for predicting mortality among critically ill burn victims. Burns (in press) (corrected proof 2008)
3. Huellermeier, E.: Case-Based Approximate Reasoning. Springer, New York (2007)
4. Hung, S.Y., Chen, C.Y.: Mammographic case base applied for supporting image diagnosis of breast lesion. Expert Systems with Applications 30(1), 93–108 (2006)
5. Juarez, J.M., Guil, F., Palma, J., Marin, R.: Temporal similarity by measuring possibilistic uncertainty in cbr. Fuzzy Sets and Systems (160), 214–230 (2008)
6. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. Knowledge and Information Systems 3(3), 263–286 (2000)
7. Klir, G., Folger, T.: Fuzzy Sets, Unvertainty, and Information. Prentice-Hall, USA (1992)
8. Le-Gall, J.R., Lemeshow, S., Saulnier, F.: A new simplified acute physiology score (SAPS II) based on a european/north american multicenter study. JAMA (270), 2957–2963 (1993)
9. Montani, S., Portinale, L.: Accounting for the temporal dimension in case-based retrieval: A framework for medical applications. Computational Intelligence 22(3), 208–223 (2006)
10. Nilsson, M., Funk, P., Olsson, E., von Scheele, b., Xiong, N.: Clinical decision support for diagnosing stress-related disorders by applying psychophysiological medical knowledge to an instance-based learning system. Artificial Intelligence in Medicine 36(2), 159–176 (2006)
11. Portinale, L., Montani, S., Bottrighi, A., Leonardi, G., Juarez, J.: A case-based architecture for temporal abstraction configuration and processing. In: IEEE ICTAI 2006, pp. 667–674 (2006)
12. Toma, T., Abu-Hanna, A., Bosman, R.J.: Discovery and integration of univariate patterns from daily individual organ-failure scores for intensive care mortality prediction. Artif. Intell. Med. 43(1), 47–60 (2008)

Using Temporal Constraints to Integrate Signal Analysis and Domain Knowledge in Medical Event Detection

Feng Gao[1], Yaji Sripada[1], Jim Hunter[1], and François Portet[2]

[1] Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK
{fgao,yaji.sripada,j.hunter}@abdn.ac.uk
[2]Grenoble Institute of Technology, Laboratoire d'Informatique de Grenoble, France
francois.portet@imag.fr

Abstract. The events which occur in an Intensive Care Unit (ICU) are many and varied. Very often, events which are important to an understanding of what has happened to the patient are not recorded in the electronic patient record. This paper describes an approach to the automatic detection of such unrecorded 'target' events which brings together signal analysis to generate temporal patterns, and temporal constraint networks to integrate these patterns with other associated events which are manually or automatically recorded. This approach has been tested on real data recorded in a Neonatal ICU with positive results.

1 Introduction

The Neonatal Intensive Care Unit (NICU) provides support for premature and very ill babies. Physiological variables such as blood pressure and heart rate are usually recorded automatically every second on a continuous basis. In addition, discrete events, such as the administration of medication and equipment settings, are also recorded. Although some of these events (e.g. the results of laboratory tests) may be automatically gathered, most of them are manually logged by nurses and doctors. This manual recording may be compromised in three ways: (i) *omission* – the event may not be entered at all (either because it wasn't considered important enough or because it was forgotten); (ii) *imprecision* – care providers in a NICU often work under considerable time pressure – events may be entered some time after they occur and it may be uncertain as to when they actually took place; (iii) *error* – the event did not occur at all and was entered by mistake.

Given the complexity of activities in the NICU it is not surprising that a considerable amount of research has been undertaken into the automatic provision of decision support – intelligent alarming, computerised guidelines, etc. Our own approach to decision support within the BabyTalk project [1] consists in the generation of retrospective summaries in natural language (English) of several hours of recorded data.

One problem that confronts all approaches to decision support is the potential unreliability of manually entered data as discussed above. Consider two actions related to respiratory support, the most common and important therapy which is provided to

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 46–55, 2009.

premature and sick babies. Patients may be artificially ventilated through an endotracheal tube (ETT) in the throat; the process of introducing this tube is known as *intubation*. Once in place, the ETT is prone to becoming blocked with mucus; these secretions can be removed by inserting a smaller diameter tube into the ETT and applying a negative pressure – *ET suction*. Intubation is sufficiently infrequent and important that it is almost certain to be reported, but the temporal precision may not be as good as we would like. ET suction is a relatively more common activity and may not always be recorded. However, detecting the occurrence of an ET suction may be useful in explaining an improvement in oxygen saturation or in reminding a nurse to analyse the secretions obtained.

Earlier studies in medical event detection have mainly focused on exploiting only the event signature (pattern) in continuous physiological signals. For instance, Fried and Gather [2] looked at outliers, level change and trend in the heart rate (HR) and invasive blood pressures; Chambrin *et al.* [3] worked on the oxygen saturation (SaO2) signal to distinguish between probe disconnection, transient hypoxia and desaturation events while Portet *et al.* [4]. looked at the detection of bradycardia. Quinn *et al.* [5] proposed a switching Kalman filtering approach to detect medical events in NICU signals and to flag unknown suspicious patterns. The main premise of these studies is that some medical events can be detected solely from their signatures in continuous physiological signals.

However, even if some events directly related to the physiological signals *can* be detected using these methods, the detection of clinical interventions is more difficult. For example, intubation involves considerable manipulation of the baby, often leading to falls in HR and in SaO2 and to increased variability of blood pressure. Unfortunately (from the point of view of event detection and identification), similar patterns are associated with other interventions during which the baby is handled (such as nappy changing or ET suction). There is actually a many-to-many relationship between medical actions and their physiological responses (the signal patterns), and event detection in these cases is therefore not possible by signal analysis alone. However, because we know that medical events do not take place on their own, we might expect to use known patterns of occurrence of other *associated events* ($\mathbf{E_{AS}}$) to detect *indirectly* the event we are interested in - the *target event* ($\mathbf{E_T}$). In other words, for a given target event (e.g. intubation or ET suction) we *model* the associated events that normally precede, accompany, or follow it. Such events include both (i) human actions (e.g. an X-ray is usually taken after a baby is intubated) and (ii) the baby's responses to these actions (e.g. bradycardia and other physiologic perturbations). This approach is very similar to that used in the *Déjà Vu* system for recognizing specific medical situations in the ICU [6]. See section 2 for a more detailed comparison with *Déjà Vu*.

In summary, our assumption is that an analysis of the physiological signals can provide temporally accurate information on when one of a number of *possible* events happened and that the ambiguity can be resolved by considering the other events which *are* recorded. We are essentially coming at event detection from two sources having different precisions simultaneously. The physiological data provides precise timing (and some elements of identification) and the other events complete the identification.

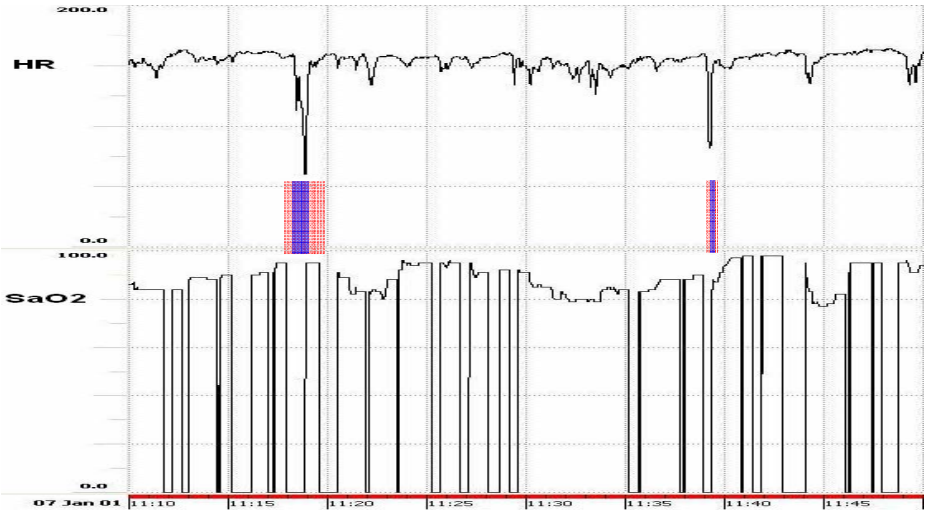


Fig. 1. Physiological signals around two ET suction events. The wide bright bars indicate two detected target events while the narrow dark bars show the actual events.

Input events

…
[11:01:56 -- 11:01:56] ADJUST VENTILATION
[11:06:19 -- 11:06:30] INCUBATOR OPEN
∗[11:11:23 -- 11:11:23] ADJUST VENTILATION
[11:13:15 -- 11:13:17] INCUBATOR OPEN
…
[11:16:09 -- 11:16:43] INCUBATOR OPEN
∗[11:18:00 -- 11:20:10] INCUBATOR OPEN
∗[11:18:10 -- 11:19:25] Discomfort
[11:20:14 -- 11:21:05] INCUBATOR OPEN
∗[11:21:54 -- 11:21:54] ADJUST VENTILATION
…
[11:26:48 -- 11:26:48] ADJUST VENTILATION
#[11:31:08 -- 11:35:09] Desaturation
[11:33:07 -- 11:33:07] ADJUST VENTILATION
…
#[11:36:12 -- 11:36:12] ADJUST VENTILATION
#[11:38:42 -- 11:39:45] Discomfort
#[11:39:08 -- 11:40:23] INCUBATOR OPEN
#[11:39:48 -- 11:39:48] ADJUST VENTILATION
[11:40:26 -- 11:40:26] ADJUST VENTILATION

Detected target events

ET Suction 1 (∗)
[11:18:22 -- 11:19:13]
Detection 1

earliest starting:	11:18:01
latest starting:	11:19:24
earliest ending:	11:18:11
latest ending:	11:20:09

ET Suction 2 (#)
[11:39:19 -- 11:39:37]
Detection 2

earliest starting:	11:39:09
latest starting:	11:39:43
earliest ending:	11:39:14
latest ending:	11:39:48

Fig. 2. Selected input and detected target events for the period shown in Figure 1

By way of illustration, Figures 1 and 2 show the input data around the time that suction events took place and the target events that were detected by our system. The Discomfort and Desaturation events were derived from an analysis of the HR and SaO2 signals. Detected events are represented with our 5-tuple format (detailed in Section 2). The input events have crisp time stamps and hence are only shown with [start -- end] times for simplicity. Two ET suction events are known to be present and their corresponding detections are listed. These two detections and their related input events are identified with * and #.

2 Approach

A very natural approach to the representation of the target and associated events is the temporal constraint network (TCN) [7]. All events in our model *(model events*) become variables while the temporal relations between them become constraints.

Temporal constraint networks have often been used for scheduling, planning and diagnostic problems [8]. Several classes of temporal constraint problems with varying degrees of computational complexity have been studied [9]. The simplest of these is called the simple temporal problem (STP) which is computationally well understood; in the STP formalisation, each arc represents a simple temporal constraint.

Reasoning about time has traditionally been studied in terms of either intervals or points. We find that acquiring knowledge from domain experts is easier with an interval based representation. For example, it is natural to describe a low blood pressure event as "*a time period during which baby has blood pressure lower than a threshold*". We therefore define an *event* as consisting of a temporal interval (possibly of zero duration) over which some property holds; we refer to this property as the *type* of the events. In order to deal with possible uncertainty in the timing of events, we represent the starting (resp. end) time point of an event with a pair of numbers, the earliest and latest times that the event could have started (resp. ended). This means that an event *e* is represented as a 5-tuple *e* = <*event type*, *earliest starting time*, *latest starting time*, *earliest ending time*, *latest ending time*>.

Although interval algebra is very expressive, reasoning is more complex than with point algebra. In particular, when we need to reason over quantitative temporal relations, the point based Temporal Constraint Satisfaction Problem (TCSP) is a more convenient choice. As a consequence, when formalizing our domain knowledge as a TCN, we choose to decompose interval based events into pairs of starting and ending points (each with earliest and latest values). Our 5-tuple event therefore maps to a pair of 3-tuple points.

Consider our model for ET suction shown in Figure 3. Nodes represent the start or end of the given event type (either inferred from a signal or a discrete event from the database). Arcs represent temporal ordering (given by the direction of the arrow) and the lower and upper bounds of the temporal distance between two nodes (numbers in square brackets). All times are in seconds and range from 0 to $+\infty$. For example, the fact that a *Desaturation* occurs between 60 and 1800 seconds *Before* an *ET Suction* is represented by the arc labelled [60, 1800].

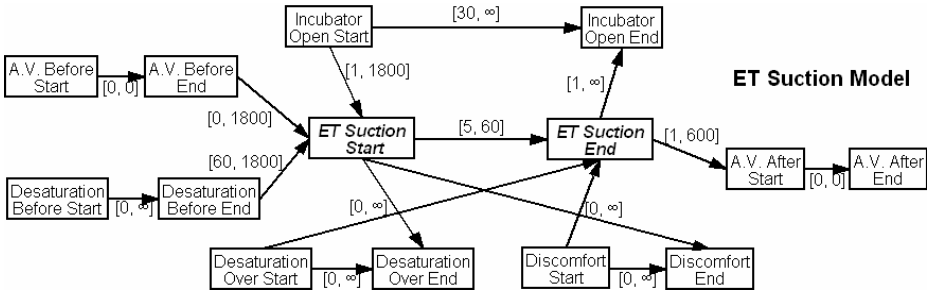


Fig. 3. Model of ET suction

The knowledge used to construct this model was acquired by interviewing experts; here are some sample constraints as expressed by those experts:

- The incubator needs to be open for at least 30 seconds for one ET suction to be performed but it may be open for longer for other actions to be undertaken.
- ET suction may start as soon as the nurse opens the incubator (i.e. 1 second later); if it is going to happen, it will do so within 30 minutes. It may take as little as 5 seconds but must not last longer than a minute.
- The nurse will adjust the ventilator after successfully completing the ET suction. She may do so immediately (1 second) even before she closes the incubator, but in any case will do so within 10 minutes.
- Very often a desaturation is taken as an indication that suction is necessary (i.e. the ETT is blocked). A nurse will take at least a minute to respond, but will do so within 30 minutes.
- There will usually be discomfort (and sometimes a desaturation) while the ET suction is taking place.
- …

The model is used for the detection of a target event by attempting to match each associated event in the model to an event instance of the same type in the input data. The temporal relations in the model are thereby constrained by the actual times of the matching data event instances; these are expressed with respect to a reference *X0* corresponding to the origin for absolute times. If the temporal network is consistent, we output the constraints on the timing of the target event.

The model is clearly an idealisation of all the events which *might* be associated with the target event. In practice some of these associated events may not be present, or at least not reported. We have therefore to be prepared to relax our model by dropping some events from it, thus creating a *partial model*, and trying again to match it with the data – the fewer the number of associated events, the weaker the model, and the lower our confidence in the existence of the target event.

In the *Déjà Vu* system [6], the time course of a clinical process is compared to a predetermined set of possible scenarios for this process. The recognition of a scenario allowed the system to predict events and to recommend interventions if necessary. Both scenarios and data were represented by a TCN. Unfortunately *Déjà Vu* was

never tested on real data. A further (and fundamental) difference between *Déjà Vu* and our system is that the former assumed that all relevant events would be found in the input data; if any event is missing then the scenario is not present. On the other hand, we are concerned with trying to reconstruct missing or under-specified events in the input data stream itself.

3 Implementation

Signal Analysis: Because the model represents intervals over which some property holds, continuous data have to be abstracted so as to generate events such as *desaturation* (episode of low oxygen in the blood) which can be derived from the oxygen saturation signal. Another crucial event is the response of the baby to the *discomfort* arising from the handling which accompanies certain actions. At present, this is derived from increased variability in the heart rate. Both signals are pre-processed with moving-median and moving-mean filters. An episode of desaturation is defined as period of time during which oxygen saturation is lower than 85%. To detect discomfort, the mean, variance, maximum and minimum of the filtered heart rate within a 30 seconds sliding window are calculated. A volatility score is derived from the variance (normalised against the long term variance) and the difference between the maximum and minimum values. Points with a score above a threshold are aggregated into intervals where we believe the baby to have been handled.

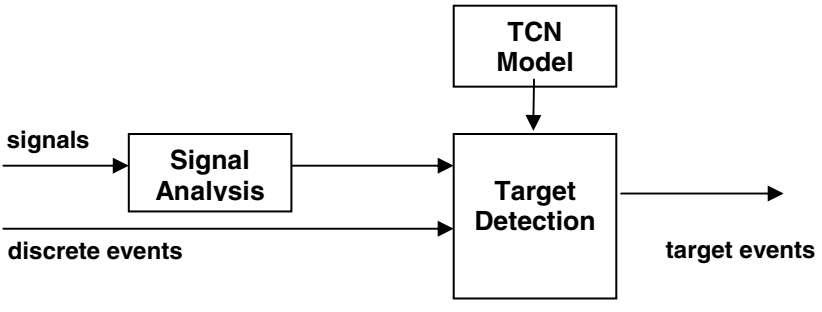


Fig. 4. Architecture of the implementation

Full Model: A model consists of events and constraints. The events are: (i) the target event ($\mathbf{E_T}$) (ii) the associated events ($\mathbf{E_{AS}}$); (iii) the reference event (*X0*). There are two types of constraint: (i) constraints between the start of a model event and the end of the same event (i.e. a constraint on the duration of the event); (ii) constraints between the start/end of one model event and the start/end of another.

Partial Models: A partial model is a modification of the full model in which one or more associated events (and the constraints involving them) have been dropped.

Input Data: We are interested in summarising what has happened over a specific time, normally several hours (e.g. a nursing shift). The data events ($\mathbf{E_D}$) therefore

consist of all of the interval abstractions generated by the signal analysis, together with all of the recorded discrete events. Note that we expect a one-to-many relationship between an associated model event and data events of the same type.

Target detector: The inputs to the target detector consist of: (i) a model (**M**) - full or partial; (ii) the data events.

The core of the detection algorithm is:

- **For each** associated model event $\mathbf{E_{AS}}$ of a instance **I** of the TCN model **M**
  - **Select** a data event $\mathbf{E_D}$ for which $\mathbf{E_D.type = E_{AS}.type}$
  - **Update** the constraints **on** $\mathbf{E_{AS}}$ using $\mathbf{E_D}$ start and end times
- **Apply** Floyd-Warshall algorithm to **I** to check consistency and modify the temporal constraints involving the target event $\mathbf{E_T}$
- **If I** is consistent **then** return the start and end times for $\mathbf{E_T}$

This algorithm has to be applied to all combinations of associated events and data events; for example, if the model contains one incubator open event, one ventilator setting and one X-ray, and the data set contains 10 incubator open events, 5 ventilator settings and 2 X-rays, there are 100 possible combinations. A considerable amount of attention has been paid to make this process as efficient as possible.

For a given model, the output is a set (possibly empty) of target event detections which have been proposed. As new target events are presented for inclusion in the set of detected target events, they are checked for temporal intersection with existing members of the set. If this is the case, the new detection is rejected. Intersection has the usual meaning but given that our target events have uncertain start and end times, we have to decide which times to use; in this case we use the earliest start and latest finish – i.e. we are checking for *possible* intersection rather than *certain* intersection.

4 Evaluation

Input Data: Our data set was collected as part of the Neonate project [10]. It consists of: (i) continuously monitored physiological signals, such as heart rate, oxygen saturation, etc. at 1Hz.; (ii) discrete events entered by a dedicated research nurse - these could be medical care actions by staff (i.e. opening the incubator, changing equipment settings, intubation, suction, etc.) or observations of the baby's status (e.g. crying, skin colour is blue). This database of discrete events constitutes a considerable richer and more accurate data set than we would expect to be collected routinely in a NICU. In particular, it provides us with a 'gold standard' for target events which we can use to test the accuracy of our detections.

In comparing our detections with the gold standard, there is an issue with the temporal granularity of intubation and ET suction. Inserting the tube is counted as an 'intubation' in Neonate; in fact several insertions (and withdrawals) may take place in a single attempt to get the tube into place. To take account of this, individual insertions which are separated by less than 10 minutes are grouped together into a single interval which spans them; we call this an 'intubation episode'. This aggregation is applied

both to intubation instances as recorded by the research nurse and to intubation instances postulated by the target detector. The same approach to aggregation is also taken for individual ET suctions.

To test each model, 83 hours of data were selected containing 7 intubation episodes (with an average duration of 6.4 minutes) and 24 suction episodes (with an average duration of 2.4 minutes). They were collected from 20 newborn infants (11 males and 9 females) with gestational ages from 23 to 37 weeks and birth weights from 500g to 3245g. Physiological signals were recorded continuously, while discrete observations were recorded by the research nurse in periods ranging in length from 45 minutes to 2 hours. Signals and observations from a subset of these observing periods were selected as input.

The data periods used for testing were selected to include periods with positive examples, periods with no episodes, and periods containing episodes which might easily be confused with the target. For example, ET suction and intubation can be confused with each other, because they are both respiratory care actions; this means: (i) the same anatomical locations (mouth and throat) are involved; (ii) the procedures are similar (e.g. adjust ventilation, open incubator, manipulate the mouth and/or throat, etc); so that (iii) the possible physiological responses are similar (i.e. bradycardia, desaturation, etc). Positive ET suction and intubation episodes therefore provide potentially confusing examples for each other. In addition, the test data included X-ray examinations. These are often (but by no means always) associated with an intubation – but never with an ET suction. We included a number of cases where X-ray examination was *not* associated with intubation.

Method: To evaluate the performance of the method the number of true positives (*tp*), false positives (*fp*) and false negatives (*fn*) are computed. For each episode, if a detected $\mathbf{E_T}$ intersects a gold standard target event (using the earliest start and latest end of the events), the detected event is taken as being a true positive. If $\mathbf{E_T}$ does not intersect any of the gold standard target events, it is counted as a false positive. All remaining unexplained gold standard episodes are then counted as false negatives.

Performance is measured with three criteria: the sensitivity $Se = tp/(tp+fn)$; the Positive Predictivity $Pp = tp/(tp+fp)$; and the F-measure $FM = 2*Se*Pp/(Se+Pp)$.

Table 1. Results of detecting episodes of ET suction and intubation

	tp	fp	fn	Se (%)	PP (%)	FM (%)
ET suction	13	5	11	54.2	72.2	61.9
Intubate	5	1	2	71.4	83.3	76.9

Results: In looking at the results presented in Table 1 it is important to note that the system does not just label events which are presented to it, but also discovers those events. To appreciate the significance of this, one needs to realise that the target intervals represent in total about 2% of the total time examined. The possibility of generating false positives is therefore considerable. The results for intubation are especially impressive when one notes that the data set included 28 chest X-ray events which were *not* preceded by intubations; none of these situations were reported as false positives.

Looking in detail at the false positives and negatives for ET suction, all 5 of the false positives were detected when other respiratory-related actions (intubation or manual ventilation) were taking place; we have not yet attempted to detect manual ventilation, but the two intubations *were* detected. Of the 11 false negatives; 6 were missed because the heart rate signals were missing or had too low a variation; we will be looking to see whether variations in other signals can improve this. Of the remainder, 3 were missed because constraints could not be satisfied, and we will attempt to refine our model.

It can be argued that false negatives are less important when the detected outputs are used for summarisation than if they were used in a recommender system. In the latter, we may well have a rule which says: "if there was *no* suction then recommend …"; for a false negative this rule would be fired incorrectly. However, with summarisation, it is usually the case that only important positive events are included in the summary – i.e. the absence of events is very rarely mentioned, and not all positive events are referred to.

5 Conclusions

We consider that our attempt to detect 'hidden' target events which occur infrequently has been reasonably successful. As far as we are aware, this is the first time such techniques have been evaluated on real ICU data, with all its inherent noise and complexity. Our future work will be directed towards refining our approach and applying it to other types of target event. Also of immediate interest is how successful it will be when we rely on routinely entered discrete events, rather than on those entered by a specially employed observer.

One area where we hope to be able to make considerable progress is in the detection of the application of clinical guidelines/protocols. Our approach to modelling is very close to that which would be required to recognise, for example, the administration of a drug to reduce elevated blood pressure and the subsequent success (or otherwise) of that treatment. In this case the 'target' would be the fact that the guideline was being adhered to. The representation of events in our model is identical to that used in the Asbru approach to the modelling of clinical guidelines [11].

Because our work is meant to generate summaries over an extended period of time (nursing shift summaries cover 12 hours), we have not been concerned with the constraints imposed by continuous operation in real time. Detection of $\mathbf{E_T}$ is not achievable in true real time when the relevant model contains associated events which come after $\mathbf{E_T}$; clearly these must be available before the model can be applied. However, the method can be made "on-line" (in soft real-time) if we are prepared to wait until the latest time by which all associated events could have occurred.

Acknowledgements

We gratefully acknowledge the assistance of other members of the BabyTalk team in Aberdeen: Ehud Reiter and Albert Gatt, and in Edinburgh: Yvonne Freer, Cindy Sykes and Neil McIntosh. BabyTalk is supported by the UK EPSRC grants EP/D049520 and

EP/D05057. Feng Gao is also supported by ORSAS. Special thanks to Luca Anselma, from the Dipartimento di Informatica, Universita` di Torino, for letting us use his temporal constraint reasoning package [12].

References

[1] Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., Sykes, C.: Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. Artificial Intelligence 173, 789–816 (2009)
[2] Fried, R., Gather, U.: Online Pattern Recognition in Intensive Care Medicine. In: Proceedings of AMIA Annual Symposium, pp. 184–188 (2001)
[3] Chambrin, M.C., Charbonnier, S., Sharshar, S., Becq, G., Badji, L.: Automatic Characterization of Events on SpO2 signal: Comparison of Two Methods. In: Proceedings of the 26th Annual International Conference of the IEEE EMBS, pp. 3474–3477 (2004)
[4] Portet, F., Gao, F., Hunter, J., Sripada, Y.: Evaluation of On-line Bradycardia Boundary Detectors from Neonatal Clinical Data. In: 29th Annual International Conference of the IEEE EMBS, pp. 3288–3291 (2007)
[5] Quinn, J.A., Williams, C.K.I., McIntosh, N.: Factorial Switching Kalman Filters Applied to Condition Monitoring in Neonatal Intensive Care. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (to appear)
[6] Dojat, M., Ramaux, N., Fontaine, D.: Scenario Recognition for Temporal Reasoning in Medical Domains. Artificial Intelligence in Medicine 14, 139–155 (1998)
[7] Dechter, R., Meiri, I., Pearl, J.: Temporal constraint networks. Artificial Intelligence 49, 61–95 (1991)
[8] Palma, J., Juarez, J.M., Campos, M., Marin, R.: Fuzzy Theory Approach for Temporal Model-based Diagnosis: An Application to Medical Domains. Artificial Intelligence in Medicine 38, 197–218 (2006)
[9] Schwalb, E., Vila, L.: Temporal Constraints: A Survey. Constraints 3, 129–149 (1998)
[10] Hunter, J., Ewing, G., Freer, Y., Logie, R., McCue, P., McIntosh, N.: Neonate: Decision Support in the Neonatal Intensive Care Unit - A Preliminary Report. In: 9th European Con-ference on Artificial Intelligence in Medicine, pp. 41–45 (2003)
[11] Shahar, S., Miksch, S., Johnson, P.: The Asgaard Project: A Task-Specific Framework for the Application and Critiquing of Time-Oriented Clinical Guidelines'. Artificial Intelligence in Medicine 14, 29–51 (1998)
[12] Anselma, L., Terenziani, P., Montani, S., Bottrighi, A.: Towards a Comprehensive Treatment of Repetitions, Periodicity and Temporal Constraints in Clinical Guidelines. Artificial Intelligence in Medicine 38, 171–195 (2006)

Temporal Data Mining of HIV Registries: Results from a 25 Years Follow-Up

Paloma Chausa[1,2], César Cáceres[1,2], Lucia Sacchi[3], Agathe León[4], Felipe García[4], Riccardo Bellazzi[3], and Enrique J. Gómez[2,1]

[1] Networking Research Center on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain
[2] Bioengineering and Telemedicine Group, Polytechnic University of Madrid, Spain
{pchausa,ccaceres,egomez}@gbt.tfo.upm.es
[3] Department of Computer Engineering and Systems Sciences, University of Pavia, Italy
{lucia.sacchi,riccardo.bellazzi}@unipv.it
[4] Infectious Diseases Unit, Hospital Clínic of Barcelona, Spain
{aleon,fgarcia}@clinic.ub.es

Abstract. The Human Immunodeficiency Virus (HIV) causes a pandemic infection in humans, with millions of people infected every year. Although the Highly Active Antiretroviral Therapy reduced the number of AIDS cases since 1996 by significantly increasing the disease-free survival time, the therapy failure rate is still high due to HIV treatment complexity. To better understand the changes in the outcomes of HIV patients we have applied temporal data mining techniques to the analysis of the data collected since 1981 by the Infectious Diseases Unit of the Hospital Clínic in Barcelona, Spain. We run a precedence temporal rule extraction algorithm on three different temporal periods, looking for two types of treatment failures: viral failure and toxic failure, corresponding to events of clinical interest to assess the treatment outcomes. The analysis allowed to extract different typical patterns related to each period and to meaningfully interpret the previous and current behaviour of HIV therapy.

Keywords: Temporal Data Mining, HIV Data Repository, Temporal Abstractions, Rule Discovery.

1 Introduction

Although the introduction of the Highly Active Antiretroviral Therapy (HAART) in 1996 has reduced the progression of HIV patients to the AIDS terminal phase, a number of uncertainties remains in the medical management of the infection. The patient control involves the intake of a high number of pills with guidelines often difficult to be implemented. Treatment regimens may be affected by HIV drug resistance as well as by possible interactions with other drugs. In addition, up to half of patients on antiretroviral therapy may experience adverse effects of the medications. Even if most side-effects decrease over time, some can be life-threatening; as a consequence, careful patient monitoring is crucial. These factors come together to decrease the quality of life in HIV infected people. Health related quality of life is lower compared to that

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 56–60, 2009.

of the general population [1], even with respect to people affected with serious illnesses like cancer or depression [2].

The recent improvements of antiretroviral therapy management are largely due to clinical trials and to increasing clinical experience. Clinical data repositories contain records from millions of patients: these data may reflect unexpected responses to therapy, previously unknown relationships between disease states and deviations from standard clinical practice that may be of interest to clinicians and researchers.

Data mining methods have been already applied to HIV data to predict drug resistance [3,4], identify relevant associations between HIV clinical variables [5] and to HAART delivery to HIV/AIDS patients [6]. However, very few of these approaches dealt with the important problem of identifying frequent temporal patterns that may occur in the data before treatment failures. To this end, we applied Temporal Data Mining (TDM) strategies to HIV data [7,8]. In particular, this work describes a temporal data mining approach to extract knowledge from a large database that contains clinical information of patients infected with HIV.

2 Materials and Methods

The knowledge extraction process used in this work follows a number of steps. First, the clinical variables included in the study are selected and prepared. Second, a set of interesting temporal patterns is defined in terms of knowledge-based temporal abstractions [9]. Finally, by resorting to an apriori-like temporal data mining method [10], the "complex" association rules that involve such interesting patterns are found.

Problem statement and data pre-processing

The algorithm is focused on the results obtained from the follow-up laboratory tests performed on HIV patients. These tests are used to analyze infection-related variables, such as viral level or CD4 count, as well as other clinical variables that characterize the patient's state. Among these there are markers of liver (GGT and bilirubin) and pancreas (lipase and amylase) activity, glucose level and fats. In stable patients follow-up tests are made every three months; in the case of complications, however, tests can be performed more frequently according to the specialists' opinion. This results in time series of variable length characterized by an irregular sampling time. In addition, some data were missing and some instances of data were present but were incorrect. One of our initial hypotheses was therefore that the sampling frequency of the clinical variables is high enough to allow us to detect significant variations in the data.

In our study we are interested in detecting increasing trends, which usually reflect a pathological behavior of the variables. Moreover, we will consider as significant only the variations in which the values, besides increasing, also cause a state change (e.g., from normal to high values). To this end, once the variables of interest are selected, each variable value is classified into one state (i.e. high, normal and low) using reference ranges established by HIV specialists (e.g., Bilirubin $> 1.2 \rightarrow$ HIGH).

Extraction of temporal association rules

To detect increasing patterns in the considered clinical variables, we exploited the Temporal Abstraction (TA) technique [9]. This framework allows to shift from a

quantitative point-based representation of time series to a qualitative interval-based one. As in the model proposed in [11], we represent the original quantitative time series data as a series of *events* and their abstract representation by *episodes*, i.e. intervals with a starting and ending time over which a feature of interest holds.

After the clinical variables are preprocessed to obtain a representation through *trend* TAs, our mining procedure goes on with the extraction of temporal precedence[1] association rules from the dataset, characterized by multivariate complex patterns in the antecedent [10]. As a consequence of the huge amount of results coming from the rule mining step, a filtering strategy is necessary to select the rules with higher medical significance and reliability. Herein the rules were evaluated on the basis of their *confidence* and of the number of patients satisfying them (*p-support*). The confidence was defined following [10]: let us assume to have a rule a temporal precedence rule $A \rightarrow_P C$. Defining TS as the duration of the observation period over which the rule is derived, RTS as the period over which the rule occurs, NAT as the number of times the antecedent occurs during TS and NARTS as the number of times the antecedent occurs during RTS, the confidence is computed as NARTS/NAT.

3 Experimental Results

The data set used in this study contains records belonging to 8000 different patients and nearly 2 million of time points. 15 clinical variables were selected: *glucose, urea, uric acid, creatinine, ALAT, ASAT, GGT, bilirubin, total cholesterol, LDL cholesterol, HDL cholesterol, triglycerides, amylase, lipase* and *alkaline phosphatase*.

The aim of this work is to detect temporal relationships between *increasing* episodes of the clinical variables preceding either a viral or a toxic failure of treatment response (e.g. Glucose Increasing AND Bilirubin Increasing PRECEDES Toxic Failure). A *viral failure* occurs when the therapy regimen does not achieve the required viral activity suppression. Despite the effectiveness of the regimen, sometimes it is necessary to change the treatment due to the side effects of the drugs (*toxic failure*). In this work, a change of therapy is labeled as a viral failure if there is a viral load test made three months before the therapy change and the level of the virus is higher than 50 copies per milliliter. Otherwise, if the viral load test has not been performed or the virus level is below the detection level, we assume that the reason for the change has been the toxicity associated with HIV drugs.

Our approach is focused on the comparison between patterns in the three principal periods of the AIDS epidemic: from the beginning of the epidemic to the introduction of the HAART therapy (1980-1997), a second period in which there were effective but toxic drugs (1998-2002) and the present time, with effective drugs with lower toxicity (2003-2008). As shown in Table 1, among all the extracted rules, clinicians selected rules with high confidence and rules that, despite a low confidence value are particularly useful to interpret the evolution of the patterns along the epidemic.

[1] A precedence relationship is herein defined by the PRECEDES temporal operator. Given two episodes, A and C, with time intervals [a1, a2] and [c1, c2], we say that A PRECEDES C if $a1 \leq c1$ and $a2 \leq c2$.

Table 1. Rules predicting toxic (a) and viral (b) failure: all of them have the form Variable 1 Increasing AND Variable 2 Increasing PRECEDES Toxic (or Viral) Failure

A- First Period

(a)			(b)		
Variables	**Confidence**	**P-Support**	**Variables**	**Confidence**	**P-Support**
Bilirubin, Amylase	0.385	29	Lipase	0.112	190
Uric Acid, GGT	0.339	20	Amylase	0.120	104
Glucose, Bilirubin	0.336	40			

B- Second Period

(a)			(b)		
Variables	**Confidence**	**P-Support**	**Variables**	**Confidence**	**P-Support**
ASAT, Triglycerides	0.122	28	ALAT, Bilirubin, A. Phosphatase	0.353	23
ASAT, T.cholesterol	0.117	25	T.cholesterol, Lipase, A. Phosphatase	0.344	20
T.cholesterol, A. Phosphatase	0.114	37	GGT, T.cholesterol, A. Phosphatase	0.315	36

C- Third Period

(a)			(b)		
Variables	**Confidence**	**P-Support**	**Variables**	**Confidence**	**P-Support**
Triglycerides	0.058	278	Lipase, A. Phosphatase	0.095	26
T.cholesterol	0.053	213	T.cholesterol, Lipase	0.089	32

4 Discussion and Conclusions

The results presented in the previous table give us the possibility of making some interesting observations about the events characterizing the three considered periods.

The rules extracted for the event of toxic failure in the first period (1980-1997) mainly involve hepatobiliary disorders, which are identified by increases of bilirubin and GGT. This information is coherent with the fact that the available treatments were characterized by a high hepatotoxicity. Moreover, in that period the most important risk factor for HIV transmission was drug intake via parenteral means, which is strictly related to a high risk of co-infection with viruses like C hepatitis and consequently to hepatic alterations. As regards viral failure, we can notice that the extracted rules are related to pancreatic alterations (increases in amylase and lipase) and also to alkaline phosphatase increase and to markers of hepatic alterations (rules not shown). This observation can be related to two facts: first, patients were exposed to a treatment with a higher toxicity and, second, their clinical situation was very critical, due to the lack of effective therapies. Finally, as we can observe from the information given by the *p-support*, in this period there are more rules related to viral failure than to toxic failure.

The rules extracted for the viral failure in the second period (1998-2002) reflect pancreatic alterations, increase of phosphatase, and hepatic alterations. This could be

due to the fact that patients were maintaining trends developed in the previous period and caused by previous treatments which made very difficult to control the viral load.

Considering the results on the third period (2003-2008), we can observe that rules related to toxic failure are much more frequent than rules related to viral failure. This shows that in this period a change in the therapy was more often related to the antiretroviral toxicity than to a wrong viral load control. In this period, the variables that frequently determine viral failure are mainly related to lipidic and alkaline phosphatase alterations, highlighting an exposition of the patients to a therapy characterized by a higher lipidic toxicity (e.g. protease inhibitors) and to a more prolonged exposition to nonnucleoside reverse transcriptase inhibitors (NNRTIs).

To conclude, the patterns extracted in the three principal periods of the AIDS epidemic are strictly related to the previous and current clinical behaviour of HIV therapy. As a next step, we also plan to introduce new information, such as the prescribed treatments and resistance mutations of the virus. With these extensions the method will provide clinicians and researchers with a deeper analysis on HIV/AIDS patients outcomes and treatment failures.

References

1. Miners, A.H., Sabin, C.A., Mocroft, A., et al.: Health-Related Quality of Life in Individuals Infected with HIV in the Era of HAART. HIV clinical trials 2, 484–492 (2001)
2. Hays, R.D., Cunningham, W.E., Sherbourne, C.D., et al.: Health-Related Quality of Life in Patients with Human Immunodeficiency Virus Infection in the United States: Results from the HIV Cost and Services Utilization Study. The American Journal of Medicine 108, 714–722 (2000)
3. Draghici, S., Potter, R.B.: Predicting HIV Drug Resistance with Neural Networks. Bioinformatics 19, 98–107 (2003)
4. Srisawat, A., Kijsirikul, B.: Combining Classifiers for HIV-1 Drug Resistance Prediction. Protein Pept. Lett. 15, 435–442 (2008)
5. Ramirez, J.C., Cook, D.J., Peterson, L.L., et al.: Temporal Pattern Discovery in Course-of-Disease Data. IEEE Engineering in Medicine and Biology Magazine 19, 63–71 (2000)
6. Ying, H., Lin, F., MacArthur, R.D., et al.: A Fuzzy Discrete Event System Approach to Determining Optimal HIV/AIDS Treatment Regimens. IEEE Transactions on Information Technology in Biomedicine 10, 663–676 (2006)
7. Post, A.R., Harrison Jr., J.H.: Temporal Data Mining. Clin. Lab. Med. 28, 83–100 (2008)
8. Raj, R., O'Connor, M.J., Das, A.K.: An Ontology-Driven Method for Hierarchical Mining of Temporal Patterns: Application to HIV Drug Resistance
9. Shahar, Y.: A framework for knowledge-based temporal abstraction. Artificial Intelligence 90, 79–133 (1997)
10. Sacchi, L., Larizza, C., Combi, C., et al.: Data Mining with Temporal Abstractions: Learning Rules from Time Series. Data Mining and Knowledge Discovery 15, 217–247 (2007)
11. Bellazzi, R., Larizza, C., Magni, P., et al.: Temporal Data Mining for the Quality Assessment of Hemodialysis Services. Artificial Intelligence in Medicine 34, 25–39 (2005)

Modeling Clinical Guidelines through Petri Nets

Marco Beccuti, Alessio Bottrighi, Giuliana Franceschinis, Stefania Montani, and Paolo Terenziani

DI, Univ. Piemonte Orientale "A. Avogadro", Via Bellini 25/g, Alessandria, Italy
{beccuti,alessio,giuliana,stefania,terenz}@mfn.unipmn.it

Abstract. Clinical guidelines (GLs) play an important role to standardize and organize clinical processes according to evidence-based medicine. Several computer-based GL representation languages have been defined, usually focusing on expressiveness and/or on user-friendliness. In many cases, the interpretation of some constructs in such languages is quite unclear. Only recently researchers have started to provide a formal semantics for some of such languages, thus providing an unambiguous specification for implementers, and a formal ground in which different approaches can be compared, and verification techniques can be applied. Petri Nets are a natural candidate formalism to cope with GL semantics, since they are explicitly geared towards the representation of processes, and are paired with powerful verification mechanisms. We show how Petri Nets can cope with the semantics of GLs in a clear way, taking the system GLARE formalism as a case study.

Keywords: clinical guidelines, Petri net, Well-formed net.

1 Introduction

The adoption of clinical guidelines (GLs), by supporting physicians in their decision making and diagnosing activities, may provide crucial advantages, both in individual-based health care, and in the overall service offered by a health care organization. Thus, several systems and projects have been developed in recent years, to realize computer-assisted GL management (see e.g., the collections [1,2]), and each system has been grounded on the definition of a proper GL representation language. The main goals of such languages are usually expressiveness and/or user-friendliness. However, in many cases, the interpretation of some constructs in these languages remains quite unclear, and/or is hidden in the code of the execution engine. As a consequence, today a wide agreement has been reached within the scientific community about the importance of pairing each GL representation language with a rigorous and formal description of its meaning, i.e., with a formal semantics [2]. While GL representation formalisms are used as the user-friendly interfaces to physicians, their formal semantics, due to their intrinsic technical complexity, are usually hidden to users. Nevertheless, they still play very important roles in the GL specification context. As a matter of fact, a semantic model allows one to provide a clear interpretation of the representation language, and guarantees that any operation performed on a GL

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 61–70, 2009.

has a precisely defined and unambiguous effect. Moreover, it also gives birth to a formal common ground on which different approaches can be compared [3], assessing what each representation can and cannot capture[1]. Additionally, the frameworks which can be used to provide a semantic interpretation of GLs are often coupled with verification techniques, which can be employed for discovering logical inconsistencies in a GL, or for proving particular properties it exhibits.

Despite its importance, the issue of copying with GL semantics has been faced only recently within the medical informatics community (see e.g. [3,5] and section 3). Moreover, existing works address the problem of representing GL primitives, but do not take into account the GL execution environment. On the other hand, in order to realistically capture the semantics of GL execution, we believe that the GL cannot be intended as an isolated process, but as one of a set of interacting processes, which also describe the behavior of additional agents (e.g. physicians, databases, labs), involved in patient care.

In this paper, we identify such processes, and describe the characteristics of their interaction. Moreover, we model the GL and GL-related processes semantics adopting the theory of Petri Nets (PNs). PNs [6] and their extensions are a family of formalisms which are well suited for modeling Discrete Event Dynamic Systems, and are explicitly geared towards the representation of interacting processes. Therefore, they are a natural candidate to cope with GL semantics in a natural and easy-to-understand way. Moreover PNs are suited to support optimal resource allocation as well as formal verification.

In the following we will consider as a reference the GLARE approach to GL representation and execution [7]. However, it is worth stressing that the methodology we propose is mostly application-independent.

2 Representing Guidelines as Petri Nets

PNs are bipartite directed graphs with two types of nodes: **place** and **transition**. The places, graphically represented as circles, correspond to the state variables of the system, while the transitions, graphically represented as boxes, correspond to the events that can induce a state change. The arcs connecting places to transitions and vice versa express the relation between states and event occurrence. Places can contain tokens drawn as black dots within places. The state of a PN, called "marking", is defined by the number of tokens in each place. The evolution of the system is given by the firing of an enabled transition[2], which removes a fixed number of tokens from its input places and adds a fixed number of tokens into its output places (according to the cardinality of its input/output arcs).

[1] [4] has made a first step towards such a comparison, but it was limited to a review of syntactic features of the representations, without considering execution semantics.

[2] A transition is enabled iff each input place contains a number of tokens greater or equal than a given threshold, and each inhibitor place contains a number of tokens strictly smaller than a given threshold. Such thresholds are defined by the cardinality of its input/inhibitor arcs [6].

In particular, in our work we use the Well-formed Net (WN) formalism, which extends the PN formalism with "colour" [8]. Its main feature is the possibility of having distinguished tokens, which can be graphically represented as dots of different colours: the colour attached to a token carries some kind of information (see the Clinical Database example below). This formalism provides two advantages: a more compact and readable representation of the system, and the possibility of using efficient solution techniques [8]. WN submodel can be composed by a composition operator [9]. The composition operator is based on the known concept of "matching labels": transitions and places are labelled and pairs of transitions (or places) with matching labels are superposed. In this paper, the labels are encoded in the transition/place name as *"Name|label"*.

In literature the majority of the GL representation languages share a set of abstract primitives[3]. These primitives can be divided in action primitives (i.e. **atomic** and **composite** actions), and control flow relation. The atomic action types are *work* actions, *query* actions and *decisions*. Work actions are atomic actions which must be executed at a given point of the GL. Query actions represent explicit or implicit requests of information, that can be obtained from a database. Decision actions embody the criteria which can be used to select among alternative paths in a GL. The composite actions are defined in terms of their components (i.e. atomic and/or composite actions) via the *part-of relation*. The control flow relations (i.e. *sequence*, *repetition*, *parallelism*) establish which actions might be executed next and in what order. It is natural to use compositional approach, so that the GLs is modelled as set of WN submodels that will be composed by the composition operator.

The WN models corresponding to atomic actions are shown in Fig. 1. In particular Fig. 1A shows how the decision action is modeled. The transition $BeginD_i|BD_i$ represents the starting of a decision process, which ends when the firing of exactly one transition $End_i|ED_i$ occurs. Observe that the transitions $End_i|ED_i$ are enabled concurrently and represent the alternative feasible paths that can be taken. The place $InputD_i|D_i$ and the place $OutputD_i|D_{i+1}$ represent the input and output of the decision process. Fig. 1B shows how the work action is modeled. Here, there is only the transition $ActionA_i|AA_i$ representing the execution of the work action. Similarly to decisions, the place $InputA_i|A_i$ and the place $OutputA_i|A_{i+1}$ represent the input and output of the work action. Fig. 1C shows how the query action is modeled. In this model, there are two transitions: $BeginQ_i|RD_i$ and $EndQ_i|AK_i$, which represent the start and the end of the data request process, respectively. Then, the places $InputQ_i|Q_i$ and $OutputQ_i|Q_{i+1}$ represent the input and output of the query action. In order to obtain the overall GL model we translate every action in the corresponding WN model and combine all these models according to the control relations specified in the GL. For instance in the case of a sequence of actions, the composition is done by superposition between the output place of the first action and the input place of the next one.

[3] Despite the generality of common concept in GL formalism, in this paper we will consider as reference the GLARE formalism.

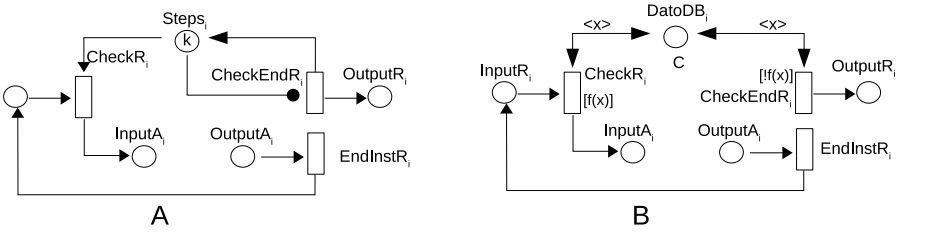


Fig. 1. The three WN models that represent the atomic actions: (A) the decision action, (B) the work action and (C) the query action

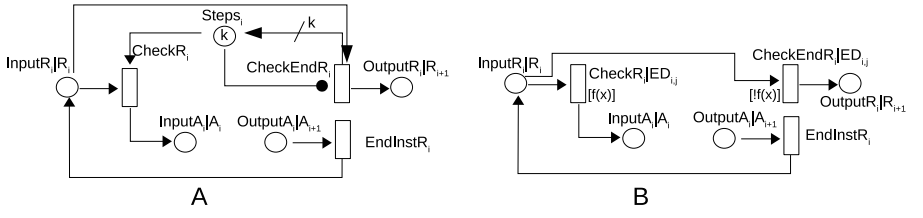


Fig. 2. The WN models representing the ways of specifying repetitions: (A) with fixed number of repetitions, (B) with exit condition

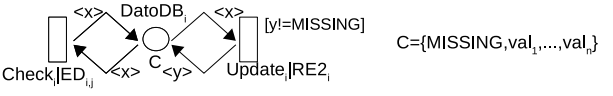


Fig. 3. The WN model describing a single datum in the clinical database

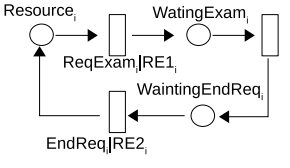


Fig. 4. The WN model describing a outside environment

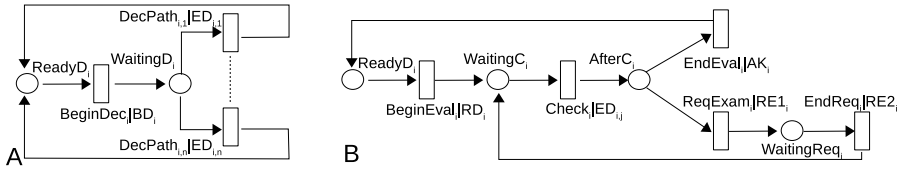


Fig. 5. The WN models describing the Physician tasks: (A) the decision process, (B) the data evaluation process

On the other hand, repetitions are managed according to two different semantics: (A) the action has to be performed a *fixed number of times*; (B) the action has to be performed until a given *exit condition* (defined on the patient data) becomes true. The WN models corresponding to these two different types of repetition are showed in Fig. 2. In the two WN models the places $InputR_i|R_i$ and $OuputR_i|R_{i+1}$ represent the input and the output of the repetition process, while the places $InputA_i|A_i$ and $OutputA_i|A_{i+1}$ represent the input and output of the actions which should be repeated.

In the former model (A), the initial marking of the place $Steps_i$ contains k tokens (graphically a k inside the place circle), that represent how many times the action has to be repeated. If there is at least a token in $Step_i$, the transition $CheckR_i$ is enabled to fire and a new instance of the action will be executed; otherwise the $CheckEndR_i$ can fire ending the repetition process and inserting k tokens in the place $Steps_i$ (this is graphically represented by a label k associated with the arc connecting $CheckEndR_i$ to $Steps_i$). Observe that the inhibitor arc (depicted as circle-headed arc) connecting $Steps_i$ to $CheckEndR_i$ assures that $CheckEndR_i$ can fire only when no tokens are in $Steps_i$.

In the latter model (B), the firing of transition $CheckR_i|ED_{i,j}$ and $CheckEndR_i|ED_{i,j}$ depend on the values of patient data (evaluated by function f(x)). It is worth noting that the two transitions cannot be enabled at the same time, because the guard of transition $CheckR_i$ is $f(x)$ and the guard of other transition is its negation [8].

In order to model and simulate GL execution on a real, specific patient, the representation of the GL *per se* is not enough: patient's characteristics need to be specified. We characterize a patient by relying on her data, which are typically maintained in the clinical database. Thus, GL execution requires the representation of the clinical database as well, interpreted as a "service" from which data can be queried, and in which updated data values can be inserted.

Updated data values are sometimes obtained from additional sources (e.g. from the hospital laboratory service). We have generically modeled such sources and services by means of a further submodel, called *outside* world.

Last but not least, GL execution is performed by a physician; therefore, the physician's behaviour needs to be modeled as well. In particular, we have identified two main tasks that the physician is expected to cover when applying a GL to a specific patient. Obviously, she is required to make decisions, i.e. she has to select exactly one diagnosis or therapy, among a set of alternative ones. In order to be as accurate and realistic as possible, we have also modeled a second task, which is the evaluation of data recency and reliability. If a data value, extracted from the database, is judged as unreliable, or not up-to-date (i.e. too old), the physician has to signal the problem, thus triggering the generation of newer data value from the outside world.

The *Clinical Database*, *Outside Environment* and *Physician* submodels representation is addressed below.

Clinical Database Net. The Clinical Database Net is represented by a set of WNs (i.e. one for each modeled datum in the database). Each WN is composed

by a unique coloured place $DatoDB_i$ and two transitions $Update_i|UP_i$ and $Check_i|ED_{i,j}$ as shown in Fig. 3. The domain associated with the coloured place (e.g. the colour class "C" in Fig. 3) represents the possible (discrete) values that the datum can assume (e.g. if the datum represents the patient temperature then the possible values in its domain could be *very high*, *high*, *normal* and *low*), so that this place can contain only tokens with colours belonging to this domain. Among the possible values associated with the tokens in $DatoDB_i$, there is always a special value, called $MISSING$, representing that no information about the datum is stored in the database. The transition $Update_i|UP_i$ has as input and output the place $DatoDB_i$ and models the update process of the datum. The use of different labels $\langle x \rangle$ and $\langle y \rangle$ on the transition input/output arcs models the fact that the new stored datum value will be selected among all the possible values in the domain. Moreover the transition guard assures that the new value can not be $MISSING$. Instead the transition $Check_i|ED_{i,j}$ models the retrieval process. The same label $\langle x \rangle$ on its input/output arcs models that the datum will not change due to the retrieval process itself.

Outside Environment Net. This WN model describes how the outside environment performs the update process of a datum, which can be required by the physician if she thinks that the current value is unreliable, or if the datum is missing. In fact the transition $ReqExam_i|RE1_i$ models the start of the update process, while the transition $EndReq_i|RE2_i$ models its end corresponding with the database update. The transition $Exam_i$ represents the execution of the required datum generation activity. Observe that the place $Resource_i$ represents the possibility of performing the required datum generation activity (e.g. the laboratory is or is not busy).

Physician Net. The Physician Net is represented by a set of WN models corresponding to the decision and to the data evaluation processes. The decision process describes how the physician takes a decision about alternative possible paths, while the data evaluation process describes how the physician decides whether a datum is reliable. In Fig. 5A the physician decision process is modeled. Note that this net is very similar to the net in Fig. 1A modeling the decision in the GL. The transition $BeginDec_i|BD_i$ represents the start of the decision process, which ends when one transition $DecPath_{i,j}|ED_{i,j}$ fires. The firing of a transition $DecPath_{i,j}|ED_{i,j}$ corresponds to the physician choice of one path; after that she becomes ready for another decision. In Fig. 5B the physician data evaluation process is modeled. The evaluation process starts with the firing of the transition $BeginEval_i|RD_i$ and ends when the transition $EndEval_i|AK_i$ fires (i.e. the physician decides that such datum is reliable). The transition $Check_i|ED_{i,j}$ represents the datum retrieval, while the free choice between the transitions $EndEval_i|AK_i$ and $ReqExam_i|RE1_1$ models the physician choice about the datum reliability. If the datum is judged as not reliable (or is missing) then the transition $ReqExam_i|RE1_i$ fires and the update process starts.

Net composition. The overall net, modeling the execution of a GL on a specific patient, is obtained by the composition of the previous WN submodels by

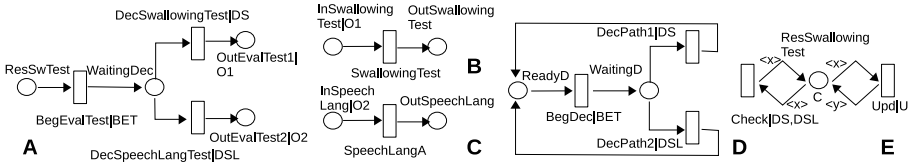


Fig. 6. The submodels involved in the composition phase

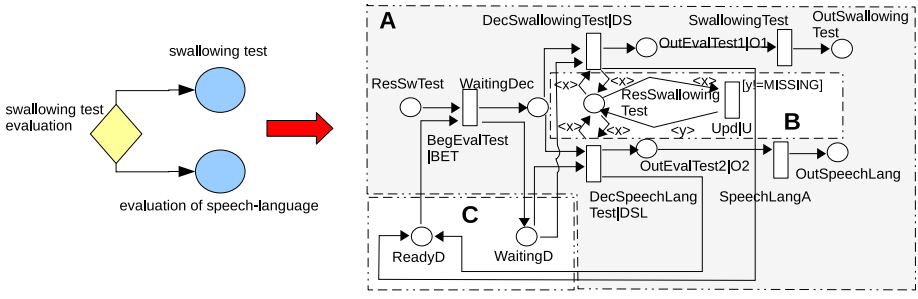


Fig. 7. The GLARE model representing the example and the corresponding composed WN model

superposition over transitions with the same label. In order to clarify how this composition is performed we provide an example, taken from the representation of an ischemic stroke GL, developed at Azienda Ospedaliera S. Giovanni Battista in Turin, Italy.

Example: we model the following decision action: the result of a swallowing test evaluation. If the physician considers the test result as negative, action "swallowing test" is performed, otherwise action "evaluation of speech-language" is executed.

In Fig. 6 all the submodels that are involved in the composition phase are shown; models A, B, C are related to the GL net, while D and E are related to the Physician and Clinical Database net.

In order to obtain the global net, shown in Fig. 7, two composition steps are necessary. The first step composes the decision action (Fig. 6A) with the work actions *swallowing test* and *evaluation of speech-language* (Figs. 6B and 6C), so that such composed net (Fig. 7 dashed box A) represents the GL model. The composition is performed by the superposition over the places $OutEvalTest1|O1$ and $InSwallowingTest|O1$, and $OutEvalTest2|O2$ and $InSpeechLang|O2$. After that, the second composition step merges the GL model (Fig. 7 dashed box A), the Physician decision model (Fig. 6D) and the Clinical Database model (Fig. 6E) by superposition over the transitions: $BegEvalTest|BET$ and $BegDec|BET$, $DecSwallowingTest|DS$, $DecPath1|DS$ and $Check|DS, DSL$, and $DecSpeechLangTest|DSL$, $DecPath2|DSL$ and $Check|DS, DSL$.

In our current implementation, the translation of GL (expressed in the GLARE formalism) into the WN submodels is performed in two steps: first the GL stored

in a XML file is translated into a set of WN submodels according to the above rules; then the submodel are composed in a unique WN model by means of the Algebra tool belonging to the GreatSPN suite [10].

3 Related Work

Although today there is a wide agreement about the importance of providing a clear semantic model for GLs, this issue has been faced only recently within the medical informatics community, and in several quite different ways. In most cases, the semantics of GLs have been only implicitly provided via an execution engine, which allows an interpretation of GLs by executing them on specific patients. Considering explicit representations, a formal operational semantics has been provided for PROforma [3] via the definition of an abstract execution engine and of rules describing how the different GL operations change the state of such an engine. On the other hand, in SAGE a mapping to standard terminologies and models (such as the virtual medical record) is advocated [11]. While the Asbru protocol representation language allows the semantics of GLs to be defined through Asbru formal semantics [5], a logical semantics to GLs has been provided in [12]. There, a graphical notation to express GLs is introduced, which can be automatically translated to the logic-based formalism provided by the SOCS computational logic framework.

We believe that our choice of relying on PNs allows us to describe GL semantics in a more natural fashion with respect to other semantic formalisms, since the mapping from GLs and GL-related processes interactions to PNs is rather straightforward. As a consequence, the output of the formalization process is easier to understand also for physicians, with respect to e.g. temporal logics.

A couple of other groups have already shown interest towards the adoption of PNs for GL representation. Quaglini's group [13], in particular, has built the system GUIDE on top of enterprise workflow standards and tools. In GUIDE, each acquired GL is translated into the Workflow Process Definition Language (WPDL), whose code is then used to build a PN. A model of the healthcare organization is also exploited to represent knowledge about available resources. A proper (commercial) software package then takes the PN and the organization model in input, and simulates the implementation of the GL in the clinical setting, in order to suggest the optimal resources allocation before the overall system is installed. Simulation enables to calculate e.g. at which time certain resources have high or low loads, what are the system bottlenecks, what are the costs of the different patients in the different stages of the GL execution, etc. Also Peleg's group has worked on the topic, studying the possibility of representing GLs as well as other complex biological processes [14] by means of PNs. They map instances of the GLIF ontology to the reference model of the Workflow Management Coalition using Protege ontology mapping rules. As in [13], then they further map this model on PNs for verification of structural properties (of biological systems) and for studying the system behaviour by simulation (for both biological systems and GLs). Unlike [13] they disregard GL resources, concentrating only on the control flow among activities.

With respect to these approaches, we perform a direct translation from GL primitives to PNs, without resorting to intermediate layers (namely, to WPDL). Moreover, and more interestingly, we do not model just the GL process, but also the behaviour of other processes involved in its execution, namely physicians, databases, etc., providing a more comprehensive view of the implementation of a GL in clinical practice. Actually, we believe that the interactions among the set of processes involved in its execution have to be properly captured, since GL semantics depend on the context in which the GL itself is meant to be applied. A more comprehensive description of the GL and of its execution environment also allows to obtain more meaningful performance indications, and to optimize resource allocation, two tasks towards which PNs are naturally very well suited.

Finally in [15] the authors propose a similar approach based on a directly translation of GL expressed in PROforma in Coloured Petri Net (CPN). We have to highlight that WN formalism used in our approach is a particular kind of CPN, that thanks to a very structured syntax for the denition of the place and transition color domains and of the arc functions and the transition guards gives the possibility to define several efficient analysis methods exploiting the intrinsic symmetries of the model. This efficient analysis methods will be very helpful when we will model a real healthcare organization instead of the executing a single GL on a single patient.

4 Conclusions

In this paper, we have afforded the problem of providing a formal semantic interpretation of GLs. In particular, having observed that GL execution is a complex phenomenon that cannot be modeled just by representing the GL *per se*, we have introduced a more comprehensive way of capturing the GL dynamics and of its execution environment, based on the idea of representing a set of processes, whose interaction models in a more realistic way the GL execution itself. PNs have thus appeared as a natural candidate to represent such environment.

Of course, a PN-based approach also allows for performance analysis and resource allocation optimization. This facility can become even more helpful by shifting the perspective from the one of executing a single GL on a single patient, to the one of dealing a real healthcare setting, in which different agents (physicians, nurses, labs) cooperate, and several, different GLs have to be executed, in order to care a set of patients. We plan to follow this direction as a future work, thus extending the approach in [13].

Moreover, PNs can be employed to support formal GL verification (i.e. for discovering logical inconsistencies in the GL, or for proving particular properties it exhibits, see chapter 4 in [2]. The use of PN in model checking (instead of other, logic-based formalisms) would provide a more easily interpretable output to end users. Additionally, PN can be easily interfaced with SPOT [16], a model checking library which relies on Transition-based Generalized Büchi Automata, allowing more compact translations of LTL formulas with respect to traditional approaches (e.g. SPIN), and which exploits global symmetries of the system,

thus speeding up computation. In the future, we plan to investigate PN-based GL verification as well, and to complete the integration (and testing) of our approach within the GLARE system.

References

1. Fridsma, D.B. (Guest ed.). Special issue on workflow management and clinical guidelines. Journal of the American Medical Informatics Association 1(22), 1–80 (2001)
2. ten Teije, A., Miksch, S., Lucas, P. (eds.): Computer-based Medical Guidelines and Protocols: A Primer and Current Trends. IOS Press, Amsterdam (2008)
3. Sutton, D.R., Fox, J.: The syntax and semantics of the PROforma guideline modeling language. Journal of the American Medical Informatics Association 10, 433–443 (2003)
4. Peleg, M., et al.: Comparing models of decision and action for guideline-based decision support: a case-study approach. Journal of the American Medical Informatics Association 10, 52–68 (2003)
5. Balser, M., Duelli, C., Reif, W.: Formal semantics of asbru - an overview. In: Proc. IDPT (2002)
6. Peterson, J.L.: Petri Net Theory and the Modeling of Systems. Prentice Hall PTR, Upper Saddle River (1981)
7. Terenziani, P., Montani, S., Bottrighi, A., Molino, G., Torchio, M.: Applying artificial intelligence to clinical guidelines: the glare approach. In: TenTeije, A., Miksch, S., Lucas, P. (eds.) Computer-based medical guidelines and protocols: A primer and current trends. IOS Press, Amsterdam (2008)
8. Chiola, G., Dutheillet, C., Franceschinis, G., Haddad, S.: Stochastic Well-formed Coloured nets for symmetric modelling applications. IEEE Transactions on Computers 42(11), 1343–1360 (1993)
9. Bernardi, S., Donatelli, S., Horvath, A.: Implementing compositionality for stochastic petri nets. International Journal on Software Tools for Technology Transfer 3(4) (2001)
10. Baarir, S., Beccuti, M., Cerotti, D., De Pierro, M., Donatelli, S., Franceschinis, G.: The GreatSPN Tool: Recent Enhancements. ACM Performance Evaluation Review Spec. Issue on Tools for Perf. Eval. 36(4), 4–9 (2009)
11. Parker, C.G., Rocha, R.A., Campbell, J.R., Tu, S.W., Huff, S.M.: Detailed clinical models for sharable, executable guidelines. In: Proc. Medinfo., pp. 45–148 (2004)
12. Alberti, M., Ciampolini, A., Chesani, F., Gavanelli, M., Mello, P., Montali, M., Storari, S., Torroni, P.: Protocol specification and verification using computational logic. In: Proc. WOA (2005)
13. Quaglini, S., Stefanelli, M., Lanzola, G., Caporusso, V., Panzarasa, S.: Flexible guideline-based patient careflow systems. Artificial Intelligence in Medicine 22, 65–80 (2001)
14. Peleg, M., Rubin, D., Altman, R.B.: Using petri nets tools to study propertuies and dynamics of bilogical systems. Journal of the American Medical Informatics Association 12, 181–199 (2005)
15. Grando, M.A., Glasspool, D.W., Fox, J.: Petri Nets as a formalism for comparing expressiveness of workflow-based Clinical Guideline Languages. In: Proc. PROHealth 2008. LNCS. Springer, Heidelberg (2008)
16. Duret-Lutz, A., Poitrenaud, D.: SPOT: an Extensible Model Checking Library Using Transition-Based Generalized Büchi Automata

Optimization of Online Patient Scheduling with Urgencies and Preferences

I.B. Vermeulen[1], S.M. Bohte[1], P.A.N. Bosman[1], S.G. Elkhuizen[2], P.J.M. Bakker[2], and J.A. La Poutré[1]

[1] Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands
I.B.Vermeulen@cwi.nl
[2] Academic Medical Centre, University of Amsterdam, The Netherlands

Abstract. We consider the online problem of scheduling patients with urgencies and preferences on hospital resources with limited capacity. To solve this complex scheduling problem effectively we have to address the following sub problems: determining the allocation of capacity to patient groups, setting dynamic rules for exceptions to the allocation, ordering timeslots based on scheduling efficiency, and incorporating patient preferences over appointment times in the scheduling process. We present a scheduling approach with optimized parameter values that solves these issues simultaneously. In our experiments, we show how our approach outperforms standard scheduling benchmarks for a wide range of scenarios, and how we can efficiently trade-off scheduling performance and fulfilling patient preferences.

1 Introduction

Due to increase of demand, improving efficiency in hospitals is becoming increasingly important. Besides the number of patients, the service that patients expect from a hospital is also increasing. Patients want more personalized care, which includes involvement in selecting appointment-times. In addition to high medical quality and resource efficiency, a hospital can compete with other hospitals by providing more patient-oriented services.

Improving efficiency in a hospital can be complex. Due to the distributed nature of a hospital, departments have local objectives and scheduling policies. The problem of scheduling a mix of patients with varying properties has to be solved locally, while hospital-wide performance depends on how schedulers interact with each other. We focus on scheduling patients to central diagnostic resources, which is often a bottleneck in patient pathways. Access time to these resources has a large influence on overall performance, as it will influence many other departments. The capacity of diagnostic resources is limited, and expensive to extend. To make efficient use of the resource, appointment-based systems are used, although in current practice the actual scheduling is often done by hand.

The basis of our scheduling problem is that different patient groups require different access times. Some patients need an appointment within a few days,

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 71–80, 2009.

others within a few weeks. We focus not only on efficient scheduling, we also want to have opportunities for patients to select their preferred timeslot. This can be achieved by dynamically controlling a trade-off between scheduling most efficiently and fulfilling patient preferences.

In this paper, we present an approach where the combination of scheduling performance and fulfilling patient preferences is optimized. Our scheduling solution consists of four main parts that we optimize simultaneously. First, resource capacity is allocated to patient groups, which allows us to have different access times per group. Second, we set dynamic rules for when an exception to the allocation can be made, this improves the overall scheduling efficiency. Third, we determine a scheduling heuristic for ordering timeslots based on efficiency. The fourth part is the use of this ordering of timeslots in trading-off scheduling efficiency and fulfilling patient preferences. A number of parameters control each part. We optimize the parameter values to find a complete scheduling solution automatically for each specific problem case we consider.

We show in our simulation experiments how our approach outperforms typical scheduling benchmarks, for a wide range of scenarios. Furthermore, we show how we can efficiently trade-off scheduling performance and fulfilling patient preferences for different patient preference models. Setting this trade-off allows hospital departments to remain in control of the scheduling, which is important for acceptance of our system in practice.

Most approaches to efficiency improvement in the hospital are from the operations research and operations management field [1]. Typical problems are strategic planning, operating room planning, capacity planning, staff scheduling, see e.g. [2]. They mostly focus on static problems, and typically do not consider online decision making, with exceptions such as [3]. They do not consider how to optimize scheduling performance in combination with patient preferences. Solutions for dynamic optimization problems usually come from the field of computational intelligence such as evolutionary algorithms [4,5].

The theoretical background of resource problems can be found in the field of queuing theory [6]. Related is the question of pooling or separating capacity and dynamic rules such as overflow rules [7]. The difference is that in our problem a timeslot must be determined upon arrival, which in a queuing system is only achieved with observable workload and first come first served (FCFS) scheduling. Our scheduling solution is not bound to FCFS but can select future timeslots per arriving patient.

In Section 2 we will discuss the problem and our approach for scheduling patients with urgencies. In Section 3 we discuss how we extend the scheduling problem and our approach to include patient preferences.

2 Scheduling with Urgencies

2.1 Problem Definition

The problem we research is how to schedule each arriving patient, such that patients are scheduled on time. For most patients, a diagnostic test must be

performed before the next consult with the physician. Most consulting hours are scheduled on a weekly basis, and the next consult is often in a week or two weeks' time. For patients with more urgent conditions, the test results must be known within a few days. However, for the most urgent patients no appointment is scheduled, and separate capacity is available in the emergency department or reserved on the resource calendar, see [8]. We focus only on patients for which an appointment must be scheduled.

In our model, patients have different urgencies, with urgency defined as the time between a patient's arrival time and required due-date. A patient is scheduled 'on time' if his appointment is before the due-date. Patients with different urgencies are scheduled to the same limited resource capacity. We assume an appointment must be made as soon as the need for the appointment is known, which we also call 'patient arrival'. This allows the hospital to provide the service of immediately communicating the appointment-time to the patient.

For non-urgent patients we set a minimum access time (MAT): the number of days between arrival and the first allowed appointment date. This allows patients to arrange their return visit to the hospital. This means that we can only schedule urgent patients to any timeslots left over on days before MAT.

Part of the problem's stochastic nature is caused by the closure of the resource in the weekend. Urgent patients will often have to be scheduled before the weekend, as the following Monday will be too late given their due-date. This causes an unequal demand over the week: at the end of the week the demand from urgent patients is larger. In our model, without losing the complexity of online scheduling with urgencies, we assume that all resource capacity can be used interchangeably and use unit-time duration for all appointments.

We formulate our model with the following. Patients arrive according to a Poisson process with arrival rate λ. Each patient p belongs to a patient group $g_p \in G$ according to a patient-group distribution D_G. The urgency of a patient is given by its group $u_p = U(g_p)$, with u_p the number of days between the arrival day and due-date. Minimum access time for non-urgent patients is given by MAT in days. Resource capacity is C, the number of timeslots on each working day.

The performance measure is based on the service levels of patient groups. Service level SL_g is the fraction of patients in group g scheduled on time (before or on their due-date). For aggregating scheduling performance over groups we use the minimum service level (MSL), $MSL = min(SL_0, ..., SL_{|G|})$, which aims at a high performance (close to 1) for each group.

2.2 Approach

We present a parameterized approach to the scheduling problem outlined above. To enable a different access time per patient group, resource capacity is allocated to groups, and patients are scheduled only to timeslots allocated to their group. In this way, the service level per group SL_g is controlled by allocating capacity $a_{g,d}$ (for group g on weekday d). The relation between service level and capacity depends on group size, urgency level, and stochastic arrival. Finding an optimal

allocation is the first step in this scheduling problem (and in many hospitals it is the only step).

In our approach we use a more flexible variation of this static capacity allocation: nested capacity allocation, patients can be scheduled to timeslots allocated to equal and lower urgency levels. Nested capacity is more flexible than strictly separated capacity as timeslots allocated to lower urgencies can be used by more patients. This reduces variability in demand, and improves resource efficiency. The optimal allocation of nested capacity can be different from the optimal allocation of static capacity.

In our approach, capacity usage is made even more efficient with conditional exceptions to the nested capacity allocation: capacity allocated to higher urgencies is also available if its utilization is below a certain threshold $t_{g,d}$. Such dynamic rules have been shown to improve performance [9]. It reduces the chance of timeslots allocated to higher urgencies being wasted.

Besides the number of timeslots allocated, the positioning of timeslots within a day also influences performance. Timeslots positioned at the end of the day have a higher chance of being used by a patient arriving during that day. It is most beneficial to position the timeslots for urgent patient at the end of the day. In Section 3 we will discuss a different positioning for patient preferences.

Capacity allocation (nested with overflow) determines for each patient which timeslots are available for scheduling. To actually schedule we have to select a timeslot from those available. We want a scheduling method that selects a timeslot such that performance is maximized over time (in Section 3, we discuss how patient preferences are involved in this selection process).

Our scheduling approach is an improvement over standard scheduling method First Come First Served (FCFS): where patients are scheduled to the earliest available timeslot, which maximizes resource utilization. However, with FCFS all timeslots up to a certain point in time are fully utilized, resulting in fewer chances for coping with a peak in demand for more urgent patients. In our approach we use a heuristic which is based on a combination of the FCFS ordering of timeslots and an ordering that counters the negative effects of FCFS, balanced utilization (BU): timeslots are ordered based on increasing utilization level per day (before the due-date). Scheduling patients based on BU, results in any available timeslots being spread out evenly over days, which increases the chances of them being beneficial for overflow from other groups. To combine the two orderings, available timeslots ts are ordered via a weighted sum of two normalized values ($w_{g,d} = 0$ equals an FCFS ordering, $w_{g,d} = 1$ equals a BU ordering):

$$\mathrm{FCFSBU}(ts) = (1 - w_{g,d})\mathrm{FCFS}(ts) + (w_{g,d})\mathrm{BU}(ts)$$

$$\mathrm{FCFS}(ts) = \frac{\text{rank of } ts \text{ in FCFS ordering}}{\text{total number of timeslots}}$$

$$\mathrm{BU}(ts) = \frac{(\text{utilization of day of } ts) - (\text{lowest utilization})}{(\text{highest utilization}) - (\text{lowest utilization})},$$

where we consider timeslots and utilization of days before the due-date. We choose the optimal value of $w_{g,d}$ per group and weekday. Per arriving patient,

we recalculate this ordering of increasing FCFSBU values over all available timeslots and schedule the patient to the first timeslot. If there are no available timeslots before the due-date, the patient is scheduled to the earliest available timeslot after his due-date (FCFS).

Recall that in our approach, we have the following three parameters per patient group per weekday: $a_{g,d}$ (number of timeslots allocated to patient group g for the weekday d), $t_{g,d}$ (the utilization threshold for overflow on capacity allocated to group g on weekday d), $w_{g,d}$ (the weight used in FCFSBU for scheduling patients of group g arriving on weekday d).

3 Patient Preferences

We extend the above scheduling problem with urgencies to additionally include patient preferences. Each non-urgent patient has a preference model P_p over timeslots, P_p states whether a timeslot is preferred for patient p. We focus on boolean-type preference model: a patient is scheduled either to a preferred timeslot or to a non-preferred timeslot. The alternative of quantifying preferences, for instance with utilities, is hard because it is difficult for patients to put values on preferences. Moreover, it is hard to compare preferences-values between patients.

With taking patient preferences into account, the overall objective O is now a weighted combination of scheduling performance (MSL), see Section 2, and patient preferences fulfillment (PP), the fraction of non-urgent patients that are scheduled to a preferred timeslot: $O = (\beta) * MSL + (1 - \beta) * PP$. By setting β a hospital department can set a preferred combination of objectives. In our experiments we show the resulting trade-off by varying the value of β.

To maximize O, that is to include patient preferences, we extend our approach the following way. Instead of scheduling the patient to the first timeslot given the FCFSBU ordering, we let the patient select from a set of timeslots that have at most an FCFSBU value of lowest FCFSBU value plus an fixed value m_g. With parameter m_g we can control how much selection-freedom a patient has. Low m_g values will limit the choice for patients and result in more efficient scheduling, while high values allow more fulfilled patient preferences.

A patient will select a preferred timeslot if it is in the set of offered timeslots. We simulate a patient's choice as uniformly random if he must select between multiple preferred timeslots, or between multiple non-preferred timeslots.

Some patients could prefer a timeslot at the end of the day, which is incompatible with the way we position timeslots within the day (urgent timeslots at the end of the day, see Section 2.2). We therefore alter the method for positioning timeslots within the day to the following: the k_d number of latest timeslots on weekday d are reserved for non-urgent patients the rest of the timeslots is positioned as in Section 2.2. Setting the value of k_d in our approach makes a trade-off between scheduling performance ($k_d = 0$) and patient preferences ($k_d > 0$).

In our experiments, we use three patient preference models P_p based on discussions with human schedulers in the hospital, described in the following paragraphs.

Work/non-work. A fraction of patients (NONWORK) is available during the day and prefers an appointment on the middle of the day, avoiding morning and afternoon traffic rush-hour while traveling to the hospital. The remaining fraction (1−NONWORK) prefers an early or late appointment to minimize the effect on their working days. Given the resource openings hours between 8am and 5pm, early is defined as before 9am, midday as between 10am and 3pm, and late is defined as after 4pm. Note that in this model there are timeslots that are not preferred by any patient. We show experimental results for different values of NONWORK.

Preferred-day. In the preferred-day model, patients have one or more preferred weekday(s). All timeslots on a preferred weekday are preferred timeslots. The days are uniformly random drawn. We show experimental results for model instances where patients each have one or two preferred weekdays.

Patient-calendar. In the patient-calendar model, which can be viewed as a combination of the two previous models, we model black-spots in a patients calendar. We divide a week in ten parts, 5 weekdays × 2 day-half's (morning/afternoon). On a number of those ten day-parts the patient will be unavailable (uniform randomly drawn). We show experimental results with varying number of black-spot day parts per patient.

4 Optimization

Our approach is parameterized and we use a search method to find the best parameter values given a scenario. We found that the problem surface was relatively smooth, and that there was an area of solutions which performed not significantly worse than the best found solution. Although we had to optimize over 50 parameters, it was still possible to find a good set of parameters values in reasonable time (< 24 hours) for a specific scenario. (Note that in practice the parameter values should be updated only as often as a few times per year.)

In the presented results below, we used an Estimation of Distribution Algorithm (EDA), see [4], with a population size of 150 and 15000 evaluations. This is a population based search method, where the distribution of each parameter value in a selection of the population is updated each generation, and used to generate individuals in the next generation. We used pair-wise comparison during selection, with a different random seed in each generation.

In our experiments, we show results of how our optimized approach (FCFS-BUdynamic) as described above, compares to the performance of three typical benchmarks each having their parameter values optimized using the EDA:

- FCFSstatic: scheduling patients First Come First Serve (FCFS) strictly to capacity allocated to their group. Capacity allocation is optimized.
- FCFSnested: scheduling patients FCFS to capacity of equal or lower urgency. Capacity allocation is optimized.
- FCFSdynamic: scheduling patient FCFS to capacity of equal or lower urgency with dynamic overflow. Capacity allocation and overflow thresholds are optimized.

5 Experiments

We have conducted many experiments to test different properties of our approach, due to space limitations we only report our main findings. Although with our EDA we automatically obtain an optimized schedule approach, we can study the found solutions and their properties. We can make the following practical conclusions based on observations in our found solutions:

- More urgent timeslots are reserved at the end of the week.
- More urgent timeslots are reserved on Thursday than on Friday.
- Overflow thresholds for urgent groups are relatively constant over the week.
- Overflow thresholds for non-urgent groups are lower on Wednesday.
- For urgent groups scheduling FCFS is more efficient than scheduling BU.
- For non-urgent groups scheduling FCFS and scheduling BU is relatively balanced, except on Fridays where it is more important to schedule FCFS.
- At the end of the week it is more important that urgent timeslots are placed at the end of the day (variable $k_d = 0$).

In our experiments, we use four patient groups, $|G| = 4$, two urgent (urgencies $U_1 = 2$ days, $U_2 = 3$ days) and two non urgent groups (urgencies $U_3 = 5$ days, $U_4 = 10$ days), with relative groups sizes: $D_G : \{D_1 = 0.14, D_2 = 0.14, D_3 = 0.28, D_4 = 0.43\}$. Having more than four different urgencies within a two-week period has little practical meaning: if groups are too similar in due-date they can be considered the same group. Minimal access time (MAT) for non-urgent patients is two days. Resource capacity C is 60 timeslots per day on weekdays, closed on the weekend. Each patient needs an appointment of one timeslot. We experiment over a number of scenarios in which we change the arrival rate, the relative group sizes (D_G), and group urgencies (U_G).

First, we present results on schedule performance without patient preferences. In Figure 1 we show the average performance (simulation length is 50,000 patients, averaged over 250 simulation runs) of the four approaches for different ρ's, ρ is ratio between the average number of arriving patients and the number of available timeslots (service rate). For all ρ's we see our approach clearly outperforms the benchmarks. The difference between using static capacity and our dynamic solution can be very large. Importantly, due to stochastic patient arrival, above a certain ρ performance will not be stable but decrease over time (a queuing effect where access time builds up). Our experiments indicate (not shown here) that performance is no longer stable with a ρ of 0.99 or larger.

To show our results are robust for different settings, we compare our approach with the three benchmarks in nine different scenarios. The scenarios differ in relative group sizes and group urgencies: we increase or decrease the due-dates for all groups; we vary the group sizes to have more or less urgent patients relative to non-urgent patients. In Table 1, we show the average performance of the four approaches with $\rho = 0.98$, for nine different scenarios. Our approach FCFSBU-dynamic has the best performance in all scenarios, although the difference is not significant in one scenario. The relative ordering of the approaches is almost the same in all scenarios, FCFSdynamic is not always significantly better than

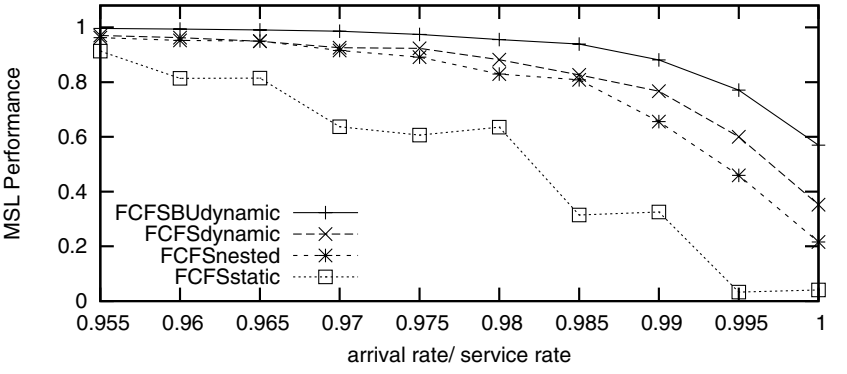


Fig. 1. Main Results

Table 1. MSL performance for nine scenarios

urgency	**urg. group size**	FCFSBUdynamic	FCFSdynamic	FCFSnested	FCFSstatic
normal	normal	**0.96**	0.87	0.86	0.57
normal	smaller	**0.95**	0.86	0.63	0.02
normal	larger	**0.96**	0.88	0.88	0.24
higher	normal	**0.74**	0.70	0.68	0.38
higher	smaller	**0.62**	0.56	0.00	0.00
higher	larger	**0.70**	**0.70**	**0.70**	0.15
lower	normal	**0.99**	0.96	0.88	0.48
lower	smaller	**0.98**	0.94	0.73	0.65
lower	larger	**0.98**	0.95	0.96	0.65

FCFSnested. This shows that our approach can be implemented to achieve the best scheduling results for various settings.

We next discuss results of optimizing the trade-off between schedule performance (MSL) and satisfying non-urgent patient preferences. Two additional variables k_d, and m_g have to be optimized, see Section 3. Given our three patient preference models we optimize solutions for different values of β (the weight in the overall objective) to get a trade-off between the two objectives.

In Figure 2a we show the trade-off between schedule performance and patient preferences, given that a non-urgent patient has a preference for either a daytime appointment or an early or late appointment (work/non-work model). We show results for three values of the non-work fractions: 0.5, 0.75, 0.9. Note that the fraction of daytime-timeslots on a day is 0.55 and the fraction of early/late-timeslots on a day is 0.22. For all three values, we see a clear trade-off between patient preferences and schedule performance. The maximum satisfaction level corresponds with a decrease in MSL of around 0.1. However, in this setting we can get close to the maximum amount of fulfilled preferences with only a very small decrease in schedule performance.

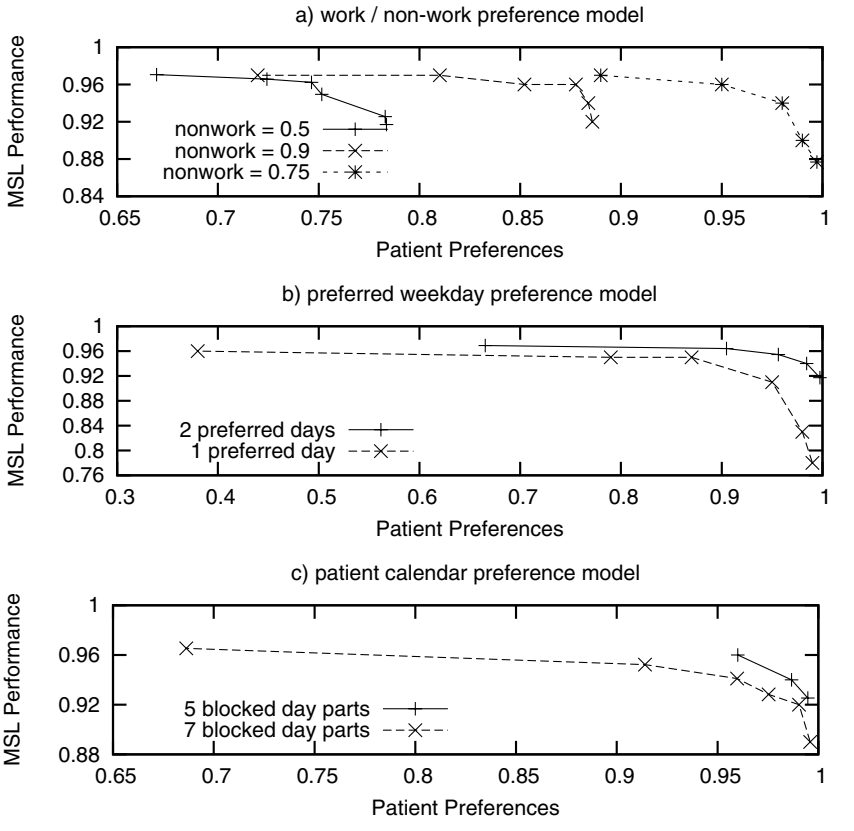


Fig. 2. Schedule Performance vs. Patient Preferences

In Figure 2b we show the trade-off for our preferred weekday model, where we consider patients having one or two preferred weekdays. If patients have only a single preferred weekday, satisfying 80% of all patients' preferences is relatively easy, but an significant decrease in schedule performance has to be expected if all patients want to get a preferred timeslot. However, this effect disappears if patients have two preferred weekdays. In Figure 2c we show the trade-off for our patient calendar model, where we vary the number of day parts a patient is unavailable. Even if patients are unavailable for 7 out of 10 day parts, 90% of the patients can get an preferred appointment, with a limited decrease in schedule performance.

6 Discussion and Conclusions

We provide an automatic optimized solution for the problem of scheduling patients with different urgencies and preferences. We show how we outperform benchmarks, independent of scenario specifics. We are able to find any preferred

trade-off between schedule performance and providing patients with the service of selecting a preferred timeslot.

We use an approach for allocating capacity, setting overflow thresholds, schedule heuristics, and offering timeslots, for which we optimize all parameter values simultaneously. We show results for multiple detailed patient preference models. Previously, in [10], some initial work for some parts of our approach was conducted, with limited experimental settings and manually set parameters.

The use of automatic optimizer such as the EDA, gives us the opportunity to find solutions for many parameters in reasonable time, for any setting. This makes our approach very generic and potentially beneficial in many different places in hospitals. In future work our approach will be extended to include non-interchangeable resources, and/or appointments with different durations.

The presented method for making a trade-off between schedule performance and freedom in selecting timeslots gives opportunity to various extensions. Based on the same trade-off we can also schedule combination-appointments over multiple departments, which we are researching in future work.

References

1. Vissers, J., Beech, R.: Health operations management: patient flow logistics in health care. Routledge, London (2005)
2. VanBerkel, P.T., Blake, J.T.: A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. Health Care Management Science 10(4), 373–385 (2007)
3. Patrick, J., Puterman, M.L.: Improving resource utilization for diagnostic services through flexible inpatient scheduling: A method for improving resource utilization. Journal of the Operational Research Society 58, 235–245 (2007)
4. Bosman, P., Grahl, J., Thierens, D.: Enhancing the performance of maximum-likelihood gaussian edas using anticipated mean shift. In: Rudolph, G., Jansen, T., Lucas, S., Poloni, C., Beume, N. (eds.) PPSN 2008. LNCS, vol. 5199, pp. 133–143. Springer, Heidelberg (2008)
5. Branke, J., Mattfeld, D.: Anticipation in dynamic optimization: The scheduling case. In: Parallel Problem Solving from Nature, pp. 253–262. Springer, Heidelberg (2000)
6. Hopp, W.J., Spearman, M.: Factory Physics: The Foundations of Manufacturing Management, 2nd edn. Irwin/McGraw-Hill, Boston (2001)
7. van Dijk, N.M.: To pool or not to pool? the benefits of combining queuing and simulation. In: Proceedings WSC 2002, Winter Simulation Conference, San Diego, pp. 1469–1472 (2002)
8. Bowers, J., Mould, G.: Managing uncertainty in orthopaedic trauma theatres. European Journal of Operational Research 154(3), 599–608 (2004)
9. Vermeulen, I., Bohte, S., Elkhuizen, S., Lameris, J., Bakker, P., La Poutré, J.: Adaptive resource allocation for efficient patient scheduling. Artificial Intelligence in Medicine 46(1), 67–80 (2009)
10. Vermeulen, I., Bohte, S., Elkhuizen, S., Bakker, P., La Poutré, J.: Decentralized online scheduling of combination-appointments in hospitals. In: Proceedings of ICAPS 2008, pp. 372–379. AAAI Press, Menlo Park (2008)

Towards the Merging of Multiple Clinical Protocols and Guidelines via Ontology-Driven Modeling

Samina Raza Abidi and Syed Sibte Raza Abidi

NICHE Research Group, Faculty of Computer Science, Dahousie University, Canada

Abstract. Decision support systems based on computerized Clinical Protocols (CP) and Clinical Practice Guidelines (CPG) fall short when dealing with patient co-morbidites, as this demands the concurrent merging of multiple CP/CPG. We present an ontology-based approach for the merging of CPG and CP at two levels–i.e. knowledge modeling level and knowledge execution level. We have developed specialized ontological modeling constructs to facilitate merging of CPG and CP. We demonstrate the merging of multiple location-specific CP and disease-specific CPG.

1 Introduction

Clinical Protocols (CP) and Clinical Practice Guidelines (CPG) are evidence-based knowledge artifacts that are designed to streamline institution-specific processes and disease-specific recommendation, respectively. There are a number of initiatives to computerize these paper based resources to utilize them for decision support and care planning in clinical setting at the point of care [1]. Notwithstanding the various successes in the computerization of these healthcare knowledge artifacts, the reality is that at execution time each CPG/CP need to be executed as an independent entity because there are no conceptual and executional provisions to link multiple CPG/CP to handle co-morbidities.

In this paper we pursue the problem of *merging* multiple CP and CPG at both the knowledge and execution levels. We take a semantic web approach that entails the use of ontologies to model the knowledge within CP and CPG [2]. Next, we attempt the merging of ontologically-modeled CP and CPG along common concepts, locations and decision-points using specialized ontology mapping constructs and merging points. In this regard, we will present two ontology-driven merging exercises: (a) The merging of location-specific CP for prostate cancer management to realize a unified prostate cancer management CP; and (b) The merging of disease-specific CPG to handle co-morbidities.

2 Approaches for Merging Multiple CPG and CP

To handle co-morbid conditions using computerized computerized CPG/CP, our approach is to systematically *merge* the computerized CPG/CP of the co-morbid diseases to generate a 'broad' evidence-based knowledge resource. Merged CPG/CP will allow to optimize the care process in terms of (a) avoiding duplication of intervention tasks,

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 81–85, 2009.

resources and diagnostic tests; (b) re-using results of common activities; (c) ensuring that different clinical activities, across active CPG/CP, are clinically compatible and their simultaneous application does not comprise patient safety; and (d) standardizing care across multiple institutions. We argue that, from a knowledge management perspective, the challenge is to develop mechanisms to 'merge' [3] multiple CPG/CP at both the knowledge modeling and knowledge execution levels. We have identified two CPG/CP merging scenarios and explore them in this paper:

1. *Merging at Knowledge Modeling Level*: In this scenario, multiple CPG/CP are merged to develop a unified 'co-morbid knowledge model' that encompasses (a) the individual knowledge of the candidate CPG/CP; and (b) the defined semantic and pragmatic relationships between the candidate CPG/CP. Here, the knowledge modeler merges the candidate CPG/CP by establishing a conceptual mapping between common concepts (such as tasks, resources, professionals, results and so on) across different CPG/CP. The merged CPG/CP model represents each CPG/CP as a combination of both unique/specialized and common/generic concepts, thus ensuring that each modeled CPG/CP maintains its unique identity and yet at the same time is part of a broader knowledge model. Knowledge level merging is particularly suitable for (a) combining a disease-specific CP for different institutions to develop a generic CP model; and (b) for combining CPG/CP of co-morbid diseases by including specialized knowledge about how to integrate them in different situations.
2. *Merging at the Knowledge Execution Level*: In this scenario, multiple CPG/CP are merged in a dynamic manner to create an *adaptable* CPG/CP that modulates with respect to the patient conditions and prospective sequence of care processes. CPG/CP merging in this case involves establishing linguistic, terminological and conceptual correspondences between the active CPG/CP models in a look-forward manner during the execution of the CPG/CP. Here, an a priori unified model is not created, rather CPG/CP merging takes place as and when needed based on predefined merging criterion and rules during the execution of the CPG/CP for a specific patient. A validation exercise, which can be both manual or rule-based, ensures that the merged CPG/CP is clinically pragmatic for the patient. Execution-level merging is typically suitable for merging CPG/CP based on common tasks across co-morbidities diseases. For CPG/CP modeled as ontologies, ontology alignment and reconciliation techniques [4] can be used to merge them.

In our work, we pursue the merging of CPG and CP by (a) representing the healthcare knowledge encapsulated within the CPG and CP as ontologies [5] [6], and (b) applying specialized ontology mapping/alignment constructs to merge multiple CPG/CP along common concepts or tasks.

3 Merging at the Knowledge Modeling Level

The idea is to merge multiple CPG/CP in terms of a unified knowledge model that identifies common elements and accounts for disease or institution-specific variations. We have developed two concepts, termed as *branching nodes* and *merging nodes* to pursue merging at the knowledge level. The *branching node* allows a CPG/CP to branch

off the unified model in case the next task/information/constraint is unique. In our ontological knowledge modeling approach, this is achieved by the modeling construct, *Class Intersection*, that models a unique instance that combines two classes. Below we show 'institution-specific' class intersections denoting an intersection between the INSTITUTION class with some other aspect to represent an instance that is unique to an Institution.

- INSTITUTION-TASK-INTERSECTION represents an intersection between classes INSTITUTION and TASK to signify a unique individual, such as a unique TaskA that is only performed at InstitutionB.
- INSTITUTION-TREATMENT-INTERSECTION represents a unique TreatmentX that is offered in only in a specific Institution.
- INSTITUTION-FOLLOWUP-INTERSECTION represents a unique FOLLOWUP offered at a specific Institution.
- INSTITUTION-CLINICIAN-INTERSECTION represents the clinician performing a specific TASK, TREATMENT or FOLLOWUP at an Institution.
- INSTITUTION-INTERVAL-INTERSECTION represents the interval duration for a specific event at a particular Institution.
- INSTITUTION-FREQUENCY-INTERSECTION represents the frequency of a specific activity at a particular Institution.

In a unified CP/CPG model, when a CP branches off on a unique path then the *merging node* serves as a point to synchronize the multiple branches to realize a unified CPG/CPG if: (a) no further activities are left in the branch; or (b) the next task is common with other branches. There are two types of merging nodes: (a) The *Merge-Wait node* waits for all the incoming branching to the node to be satisfied before the execution moves forward; and (b) The *Merge-Proceed* node simply merges the branch to the unified model and continues the execution without waiting for the completion of the other branches. In figure 1, we illustrate both the branching and merging nodes. Note that after the task 'RecieptOfBiopsyReportByUrologist' the three institutions perform unique tasks (modeled as Institution-Task Intersections) and therefore three separate institution-speciifc branches are spawned, each having unique individuals for *hasTask* and *isFollowedByConsultation* relations. Later, the task 'Consult-4' serves as a merging node to realize a unified CP [7].

4 Merging at the Knowledge Execution Level

In conceptual terms, execution level merging involves the alignment of the knowledge models representing the CPG and CP. In our case, we use separate ontologies to model both CPG [5] and CP[6], therefore merging is pursed as an ontology alignment exercise [4] based on the presence of common plans/steps that exist across multiple active CP/CPG. Merging at the execution level is complex and involves a temporal aspect to maintain a state graph that encapsulates the tasks completed and the forthcoming tasks. CPG/CP merging is, therefore, based on the commonality of the forthcoming tasks and the re-usability of results of previous tasks. The outcome of this exercise is (a) a comprehensive decision model, encompassing multiple CPG/CP; (b) optimization of resources

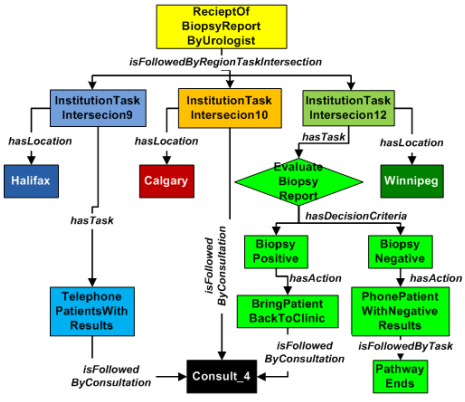


Fig. 1. Branching and Merging of Clinical Pathways

Fig. 2. Merging two concurrent CPG along a common step: (a) Common step takes place at the same time; (b) Common step takes place at different times

by reducing repetitive tests/actions, and (c) efficient execution of common tasks. The dynamic merging of multiple CPG/CP, whilst maintaining clinical pragmatics, is quite challenging because (i) recommendations that are common across multiple CPG are not necessarily administered at the same time, and (ii) certain parts of the merging CPG may later result in contradictions or adverse effects. In our work, we pursue three CPG merging scenarios.

Scenario 1: Both guidelines recommend a common step at the same time. Both CPG merge at the common step and then branch off to their respective paths when the common step is completed (shown in figure 2a).

Scenario 2: In case the common step is not executed at the same time by two CPG, then CPG merging is still possible if the CPG in front (in terms of its execution order) can wait before executing the common step–i.e. the *ability-to-wait* constraint for the common step can be satisfied. To model this merging scenario, for each ACTION-STEP we have specified the following attributes: (a) *expected-duration* to represent the average execution time for a step; and (b) *logic-to-calculate-acceptable-wait* to specify the criteria to calculate the maximum acceptable wait time before starting the step (shown in figure 2b).

Scenario 3: Two CPG can be merged if they can re-use the results of a common step. To ensure that the result is not outdated, we have specified an attribute *acceptable-duration-of-results-if-available* that will ensure that the trailing CPG is using a valid result.

5 Concluding Remarks

In this paper we have discussed our ontology based approach to model the merging of multiple CPG/CP. We evaluated the merged knowledge for representational adequacy and efficiency [2], and found that the merged ontological models adequately capture the concerned concepts. The key feature of our approach is that it supports execution semantics whilst maintaining clinical pragmatics. We argue that by investigating the merging of different CPG/CP one can (a) generalize the knowledge across different institutions; and (b) identify specialized tasks at each institution for different diseases [8].
Acknowledgement: This research project is supported by a research grant from Green Shield Foundation (Canada), aiming to investigate "Decision Support Services for Managing Chronic Diseases with Co-Morbidities".

References

1. Peleg, M., Tu, S., Bury, J., Ciccarese, P., Fox, J., Greenes, R.A., Hall, R., Johnson, P.D., Jones, N., Kuma, A.: Comparing computer-interpretable guideline models: A case-study approach. Journal of American Medical Informatics Association 10, 52–68 (2003)
2. Bodenreider, O.: Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearbook Medical Informatics, 67–79 (2008)
3. Sascha, M., Stefan, J.: Process-oriented knowledge support in a clinical research setting. In: Proceedings of 12th IEEE Symposium on Computer-Based Medical Systems. IEEE Computer Society Press, Los Alamitos (2007)
4. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Heidelberg (2007)
5. Abidi, S.S.R., Shayegani, S.: Modeling the form and function of clinical practice guidelines: An ontological model to computerize clinical practice guidelines. In: Riano, D. (ed.) K4HeLP 2008. LNCS (LNAI), vol. 5626, pp. 81–91. Springer, Heidelberg (2009)
6. Hurley, K., Abidi, S.S.R.: Ontology engineering to model clinical pathways: Towards the computerization and execution of clinical pathways. In: 20th IEEE Symposium on Computer-Based Medical Systems, Maribor, Slovenia, June 20-22. IEEE Press, Los Alamitos (2008)
7. Abidi, S., Abidi, S.S.R., Hussain, S., Butler, L.: Operationalizing prostate cancer clinical pathways: An ontological model to computerize, merge and execute institution-specific clinical pathways. In: Riano, D. (ed.) K4HeLP 2008. LNCS (LNAI), vol. 5626, pp. 1–12. Springer, Heidelberg (2009)
8. Lenz, O., Reichert, M.: It support for healthcare processes: Premises, challenges, perspectives. Data and Knowledge Engineering 61, 39–58 (2007)

Analysing Clinical Guidelines' Contents with Deontic and Rhetorical Structures

Gersende Georg[1], Hugo Hernault[2], Marc Cavazza[3], Helmut Prendinger[4], and Mitsuru Ishizuka[2]

[1] Haute Autorité de Santé, F-93218, Saint-Denis La Plaine, France
[2] Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[3] School of Computing, University of Teesside TS1 3BA Middlesbrough, United Kingdom
[4] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
g.georg@has-sante.fr, hugo@nii.ac.jp, m.o.cavazza@tees.ac.uk, helmut@nii.ac.jp, ishizuka@i.u-tokyo.ac.jp

Abstract. The computerisation of clinical guidelines can greatly benefit from the automatic analysis of their content using Natural Language Processing techniques. Because of the central role played by specific deontic structures, known as recommendations, it is possible to tune the processing step towards the recognition of such expressions, which can be used to structure key sections of the document. In this paper, we extend previous work on the automatic identification of guidelines' recommendations, by showing how Rhetorical Structure Theory can be used to characterise the actual contents of elementary recommendations. The emphasis on causality and time in RST proves a powerful complement to the recognition of deontic structures and supports the identification of relevant knowledge, in particular for the identification of conditional structures, which play an important role for the subsequent analysis of recommendations.

Keywords: natural language processing, rhetorical structure theory, guidelines.

1 Introduction

The computerisation of clinical guidelines had led to a renewed interest in their automatic processing. Whilst encoding of clinical guidelines can take place manually [1] or with the assistance of visual interfaces [2], this process can be facilitated by introducing Natural Language Processing (NLP) techniques in the process. We have shown in previous work [3] how the automatic identification of specific linguistic markers of clinical recommendations, known as deontic operators, could be used to provide a first level of structuring. The main outcome of this approach has actually been a software to analyse the structure of clinical guidelines during their development process. In this paper, we describe an extension of our previous approach in which the actual contents of recommendations can be further structured using a medically-relevant subset of rhetorical relations.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 86–90, 2009.

2 Clinical Guidelines and Their Analysis

Clinical guidelines are sophisticated documents which are often syntactically complex. One research direction consists in standardising guidelines' writing or even recurring to controlled languages [4]. Whilst their automatic processing is beyond the state-of-the-art of NLP techniques, we have shown recently that much benefit could be gained from the recognition of key expressions which would structure portions of text according to the document's logic. This is a case where local or shallow processing can be used to structure free text segments. G-DEE (for *Guidelines Document Engineering Environment*), is a text analysis environment dedicated to the study of clinical guidelines [5]. It supports multiple document processing functions including the automatic recognition of recommendations using shallow NLP techniques (such as Finite-State Automata, FSA) to recognise deontic expressions corresponding to recommendations. Since 2007, G-DEE has been integrated into the process of clinical guidelines authoring at the French National Authority for Health (Haute Autorité de Santé), which is in charge of the elaboration of all official guidelines in France. G-DEE has been used from the first draft of recommendations to the final version of the document, to provide an independent analysis of guidelines structure. Since its introduction two years ago, the number of requests for authoring by HAS project leaders with G-DEE has steadily increased (44 cumulative number of requests considering that one guideline can be analyzed from 1 to 3 by G-DEE) indicating a growing interest amongst users (its use has not been made compulsory).

3 Complementarity between Deontic and Rhetoric Structures

In addition to its use to support the guidelines' authoring process, we've shown in previous work [6] that structuring guidelines with deontic operators can help identifying important clinical actions that can be matched to underlying protocols. However, extending the automatic processing of guidelines to the actual contents of individual recommendations, i.e. processing the free text content of deontic operators' scopes, remains a challenging task from an NLP perspective. Our FSA-based recognition of deontic operators already required significant linguistic resources, despite being focused on specific linguistic descriptions. It thus seems difficult to adapt similar principles for the analysis of scopes which exhibit much greater syntactic variability and semantic coverage. Ideally, we would seek a method which reconciles broad linguistic coverage, shallow NLP, and the ability to further structure the contents of recommendations' scopes. All this points towards discourse-processing methods, and led us to explore the potential use of Rhetorical Structure Theory (RST) [7]. Although some authors have suggested that legal texts were not amenable to RST formalization [8] and we have shown the proximity between legal texts and clinical guidelines in their use of deontic structures [5], we were comforted in our hypothesis by the many previous references applying RST to medical NLP and medical language generation [9-10].

4 RST Parsing of Recommendations' Scopes

In order to uncover the rhetorical structure of clinical guidelines, we used a RST discourse parser based on Support Vector Machines (SVM). The parser uses a rich set of shallow lexical, syntactic and structural features from the text, and processes its input in two steps. Firstly, a discourse segmenter cuts the text into "elementary discourse units" (EDUs), the atomic units of discourse which are the terminal nodes of the rhetorical structure tree. Each word and its context are represented by a feature vector. In a second step, the calculated EDUs are passed to the tree-building component, which creates the full RST tree. In order to improve the computational properties of the classification problem and ensure a good separability between the label classes, we used the reduced set of 18 relations defined in [11] and used by [12] amongst the original set of 78 rhetorical relations. While we have observed a natural and quite efficient complementarity between deontic recognition and the RST analysis of recommendations' scopes, it would still be appropriate to investigate whether RST parsing could be used as a sole principle for guidelines' structuring and recommendations analysis. On the theoretical side, examples described by Gallardo [13] suggest that RST functions fail to capture key elements of recommendations. Direct RST analysis of recommendations mostly produces structures based on conditions and elaborations. When conditions are explicitly represented as part of the recommendation, RST analysis correctly identifies part of the recommendation, although it fails to provide a complete segmentation along the lines of those produced by G-DEE, with proper identification of scopes.

5 The Recommendations' Processing Pipeline

Since our deontic parser has been validated through user experiments and through real-world deployment within a guidelines production agency (HAS), we naturally thought of extending G-DEE by a further step of RST analysis, targeting the recommendations' scopes. We adopted a processing model based on the fusion of outputs from G-DEE and the RST parser. We have developed a module operating in two steps: (i) localization of a deontic operator within the RST structure; (ii) fusion of the RST structure with the front-scope and the back-scope (resulting in these scopes being structured by RST functions, see Figure 1). A pre-processing step consists of analysing the guideline using G-DEE to determine sentences that correspond to recommendations. An RST analysis is then performed on the file containing the recommendations identified by G-DEE. The G-DEE processor scans the sentence and extracts the deontic operator using the specific mark-up <*DeontOp*>. The next step consists of localizing the same deontic operator in the XML RST file, using the G-DEE processor that proceeds through a standard finite-state processing algorithm. The successful match leads us to determine whether the deontic operator is a part of the nucleus (N) or the satellite (S) by the recognition of the RST function as shown above. The G-DEE processor then scans the sentence from the RST file, and extracts the function corresponding to the front-scope (either the nucleus or the satellite previously recorded information). It then uses a dedicated FSA to mark-up the corresponding front-scope with an appropriate tag (<*Manner-Means*> in the example above). In a similar way, the function corresponding to the back-scope is recorded and the G-DEE processor tags the back-scope accordingly (<*Temporal*> in the example).

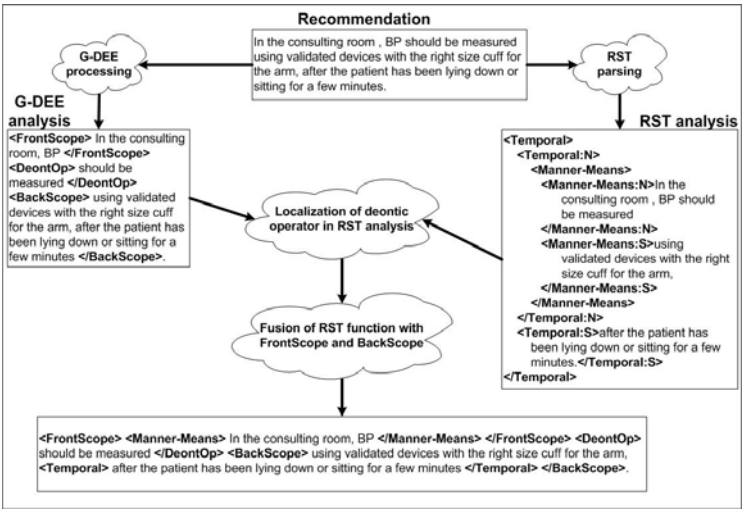


Fig. 1. Refining Recommendations' Structure by Merging Deontic and Rhetorical Mark-ups

6 Results and Discussion

We have extracted recommendations from the 2005 Hypertension Guidelines (in English, "Management of adults with essential hypertension"), obtaining a test set of 79 recommendations. Overall, RST processing with basic functions had a very significant contribution for approximately 25% of recommendations. This means that not only it did refine the recommendations' structure, but the new relations were directly meaningful. The most useful RST functions detected on these Guidelines are: *Condition* (10 occurrences), *Manner-Means* (5), *Temporal* (4), and *Enablement* (3). RST parsing also contributed to an improved structure with generic functions, through the *Elaboration* function, for 14% of recommendations: this includes isolating the grade of the recommendation or some specific target from within (generally back-) scopes. A better joint recognition of functions could achieve substantial improvements of the *Conditional* relations.

7 Conclusions

The analysis of recommendations' scopes using RST can successfully extend our previous approach, improving automatic structuring for 44% of recommendations, which increases significantly the quality of the automatic processing, even more so considering that documents tend to be analysed several times during their authoring cycle. Further, it remains compatible with our philosophy of document processing, which is to structure text segments using discourse markers, specific (e.g. deontic), or not. This type of automatic analysis tends to be well-accepted by guidelines' developers as it is designed as a human-in-the-loop approach. This is also an interesting test case for medical NLP, where the recognition of discourse structures, rather than of named entities or actions

(for instance through Information Extraction or terminological processing) can support the identification of clinically relevant information over an entire text.

Acknowledgments. Dave du Verle developed the first version of the RST parser used in these experiments. Gersende Georg and Marc Cavazza have been funded by NII for a summer visit in 2008 during which this collaboration was established.

References

1. Shiffman, R., Karras, B., Agrawal, A., Chen, R., Marenco, L., Nath, S.: GEM: A proposal for a more comprehensive guideline document model using XML. J. Am. Med. Informatics Assoc. 7, 488–498 (2000)
2. Shahar, Y., Young, O., Shalom, E., Mayaffit, A., Moskovitch, R., Hessing, A., Galperin, M.: The Digital electronic Guideline Library (DeGeL): a hybrid framework for representation and use of clinical guidelines. Stud. Health Technol. Inform. 101, 147–151 (2004)
3. Georg, G., Jaulent, M.-C.: A Document Engineering Environment for Clinical Guidelines. In: King, P.R., Simske, S.J. (eds.) Proceedings of the 2007 ACM Symposium on Document Engineering, pp. 69–78. ACM Press, New York (2007)
4. Fuchs, N.E., Kaljurand, K., Schneider, G.: Attempto Controlled English Meets the Challenges of Knowledge Representation, Reasoning, Interoperability and User Interfaces. In: FLAIRS Conference, pp. 664–669 (2006)
5. Georg, G., Jaulent, M.-C.: An Environment for Document Engineering of Clinical Guidelines. In: Proceedings AMIA Symposium, pp. 276–280 (2005)
6. Georg, G., Cavazza, M.: Integrating Document-based and Knowledge-based Models for Clinical Guideline Analysis. In: Bellazzi, R., Abu-Hanna, A., Hunter, J. (eds.) AIME 2007. LNCS (LNAI), vol. 4594, pp. 421–430. Springer, Heidelberg (2007)
7. Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory: Toward a functional theory of text organisation. Text 8(3), 243–281 (1988)
8. Taboada, M., Mann, W.C.: Applications of Rhetorical Structure Theory. Discourse Studies 8(4), 567–588 (2006)
9. Piwek, P., Hernault, H., Prendinger, H., Ishizuka, M.: T2D: Generating Dialogues Between Virtual Agents Automatically from Text. Intelligent Virtual Agents, 161–174 (2007)
10. Grasso, F.: Rhetorical coding of health promotion dialogues. In: Dojat, M., Keravnou, E.T., Barahona, P. (eds.) AIME 2003. LNCS, vol. 2780, pp. 179–188. Springer, Heidelberg (2003)
11. Carlson, L., Marcu, D., Okurowski, M.E.: Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In: Proc. of the Second SIGdial Workshop on Discourse and Dialogue. Annual Meeting of the ACL, vol. 16, pp. 1–10 (2001)
12. Soricut, R., Marcu, D.: Sentence level discourse parsing using syntactic and lexical information. In: Proc. of the 2003 Conference of the North American Chapter of the Association For Computational Linguistics on Human Language Technology, vol. 1, pp. 149–156 (2003)
13. Gallardo, S.: Pragmatic support of medical recommendations in popularized texts. Journal of Pragmatics 37(6), 813–835 (2005)

A Hybrid Approach to Clinical Guideline and to Basic Medical Knowledge Conformance

Alessio Bottrighi[1], Federico Chesani[2], Paola Mello[2], Gianpaolo Molino[3], Marco Montali[2], Stefania Montani[1], Sergio Storari[4], Paolo Terenziani[1], and Mauro Torchio[3]

[1] DI, Univ. del Piemonte Orientale, via Bellini 25/g, 15100 - Alessandria, Italy
{alessio.bottrighi,terenz,stefania}@mfn.unipm.it
[2] DEIS - Univ. di Bologna, viale Risorgimento 2, 40136 - Bologna, Italy
{federico.chesani,paola.mello,marco.montali}@unibo.it
[3] Az. Ospedaliera S. Giovanni Battista, via Bramante 88, Torino, Italy
mtorchio@molinette.piemonte.it, giampaolo.molino@virgilio.it
[4] DI - Univ. di Ferrara, via Saragat 1, 44136 - Ferrara, Italy
sergio.storari@unife.it

Abstract. Several computer-based approaches to Clinical Guidelines have been developed in the last two decades. However, only recently the community has started to cope with the fact that Clinical Guidelines are just a part of the medical knowledge that physicians have to take into account when treating patients. The *procedural* knowledge in the guidelines have to be complemented by additional *declarative* medical knowledge. In this paper, we analyse such an interaction, by studying the conformance problem, defined as evaluating the adherence of a set of performed clinical actions w.r.t. the behaviour recommended by the guideline and by the medical knowledge.

1 Introduction

Clinical Guidelines are "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" [4]. In the last decade, the research about computerized guidelines has reached a relevant role within the Artificial Intelligence in Medicine community, and many different approaches and projects have been developed to create domain-independent computer-assisted tools for managing, acquiring, representing and executing clinical guidelines [8]. Only recently, however, some approaches have started to consider that Clinical Guidelines (henceforth CG) cannot be interpreted (and executed) "in isolation". CGs are just a part of the medical knowledge that physicians have to take into account when treating patients. The *procedural knowledge* in the guidelines have to be complemented by additional *declarative medical knowledge*. In this paper we explore such an interaction from the viewpoint of the conformance problem, intended as the adherence of an observed execution to both types of knowledge.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 91–95, 2009.

2 Guidelines, Basic Medical Knowledge and Conformance

Several conditions are usually implicitly assumed by experts building a CG:

(*i*) *ideal patients*, i.e., patients that have "just the single" disease considered in the CG (thus excluding the concurrent application of more than one CG), and are "statistically relevant" (they model the typical patient affected by the given disease), not presenting rare peculiarities/side-effects;
(*ii*) *ideal physicians* executing the CG, i.e., physicians having the basic medical knowledge, allowing them to properly apply the CGs to specific patients;
(*iii*) *ideal context* of execution, so that all necessary resources are available.

Assumption (*i*) is needed, since the variety of possible patients is potentially infinite, and the CG cannot explicitly cope with all possible nuances and exceptions. Assumption (*ii*) is also necessary, since the CG focuses on the proper treatment of a specific disease, and can not code all the "basic medical knowledge". Nevertheless, such a knowledge is needed when a CG will be executed, to properly adapt the "general" prescriptions in the CG to the specific case constituted by a particular patient. Finally, experts cannot know all possible execution contexts, so that they usually assume availability of resources (assumption (*iii*)).

As a consequence of these assumptions, a CG cannot be interpreted as a protocol which has to be applied tout cour, and the actions prescribed by CGs cannot be interpreted as "must do" actions. The intended semantics of CGs cannot be analysed in isolation from the *basic medical knowledge*. Roughly speaking, given a patient X to which a CG $\mathcal{G}$ has to be applied in a context $\mathcal{C}$, $\mathcal{G}$ has to be interpreted as a set of *default prescriptions*: whenever X and $\mathcal{C}$ *fit* with $\mathcal{G}$'s prescriptions, they must be executed. However, X (or $\mathcal{C}$) may have peculiar features which are not explicitly covered by $\mathcal{G}$. In such a case, the *basic medical knowledge* must be considered to identify the correct actions.

The interplays between CG's knowledge and the Basic Medical Knowledge (BMK from now on) can be very complex, as shown by the following examples.

Example 1. Patients suffering from bacterial pneumonia caused by agents sensible to penicillin and to macrolid, allergic to penicillin, must be treated with macrolid.
BMK: Don't administer drug X to a patient if she is allergic to X.

In Ex. 1, two alternative treatments (penicillin or macrolid) are possible given the CG, but one of them is excluded, given the underlying basic medical knowledge, because of allergy to penicillin. Here the BMK "complete" the CG and help to discriminate among different alternatives. In other cases, the basic medical knowledge may apparently contradict the recommendations in the CG.

Example 2. Patient with acute myocardial infarction presenting with acute pulmonary edema; before performing coronary angiography it is mandatory to treat the acute heart failure.
BMK: The execution of any CG may be suspended, if a problem threatening the patient's life suddenly arise. Such a problem has to be treated first.

In Ex. 2 the execution of a CG is suspended, due to the arise of a problem threatening the patient's life. Notice that the "contradiction" (logical inconsistency) between CG's recommendations and BMK is only apparent. It arises just in cases one interpret CG's recommendations as *must do*, while, as a matter of facts, they may be emended by BMK.

Also, there seems to be no general rule in case of "apparent contradiction": it could be that BMK recommendations "win" over CG ones, or vice versa. In Ex. 3 a treatment is performed even if it may be dangerous for the patient. In some sense, not only some CG's prescriptions are "defeasible", since they may be overridden by BMK, but the same also holds for part of BMK.

Example 3. In a patient affected by unstable angina and advanced predialytic renal failure, coronary angiography remains mandatory, even if the contrast media administration may cause a further final deterioration of the renal functions, leading the patient to dialysis.

When considering conformance of an execution w.r.t. a CG, additional actions not foreseen by the CG could be considered as an issue. This could happen as a consequence of some particular routine (even a CG applied at the same time), like in Ex. 4.

Example 4. Calcemia and glycemia are routinely performed in all patients admitted to the internal medicine ward of Italian hospitals, regardless of the disease.

To summarize, CG semantics is very complex, and cannot be simply interpreted as a strict, normative procedures. The context of execution and the BMK complement the prescriptions in the CGs, bridging (at least in part) the gap between the "ideal" and the "real" application cases. In this hybrid situation the property of conformance, intended as the adherence of an execution to CGs and BMK recommendation, becomes more and more important, and yet difficult to be captured. Defining conformance as the simple conjunction of the conformances to each different piece of knowledge might not be the best choice. Both the CG knowledge and the BMK can be defeated, hence there is no general rule on the prevalence of one or the other. The conformance evaluation is a task that *necessarily* requires the intervention of a physician. However, such a task can be very difficult and time consuming.

3 Evaluating Conformance

We propose to combine, in an hybrid system, tools used for independently evaluating conformance to CGs and BMK. The aim is twofold: on one side, we combine both types of knowledge; on the other side, we facilitate the physician task by identifying discrepancies between actual executions and CG/BMK recommendations, and by suggesting possible explanations. The architecture is shown in Figure 1.

Procedural and workflow-like systems seems to be the most common choice for representing CGs [6], while declarative approaches might be preferable to

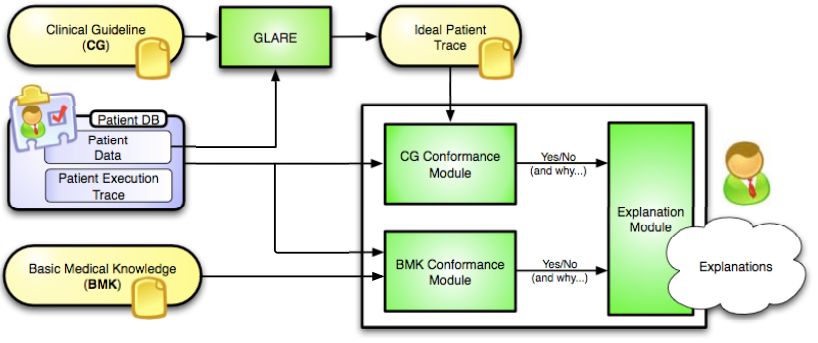


Fig. 1. Architecture of our hybrid approach for the conformance evaluation

represent BMK, due to its similarity to knowledge bases. Our system takes as input three different pieces of information: a CG described using the GLARE language [9], a BMK described by means of the $\mathcal{S}$CIFF language [1], and a set of data about the patient (the "Patient DB"). The Patient DB includes the *patient data* (a description of the patient and her medical situation); and the *actual execution* (a list of the medical actions executed on that patient).

The conformance is evaluated by means of two different modules, and their output is combined to provide possible explanations. The first module is the *CG Conformance Module*. It compares the observed execution with the "ideal" execution, to spot possible differences:(i) executed actions not envisaged by the CG, and (ii) actions foreseen by the CG but not observed in the execution. The ideal execution is obtained through the GLARE system, by applying specific patient data to the generic CG description.

The second module, the *BMK Conformance Module*, tests the conformance of the actual execution w.r.t. the BMK, expressed using the $\mathcal{S}$CIFF formalism. The $\mathcal{S}$CIFF Proof Procedure generates expectations about the events, and automatically checks if such expectations are fulfilled or not. Besides a yes/no answer, the $\mathcal{S}$CIFF Proof Procedure provides also explanations, in terms of the expectations about the *happened*/*not happened* events.

Finally, the *Explanation Module* combines the outputs of the two previous modules: a list of discrepancies (observed/not observed medical actions) is given to the physician, together with expectations that might justify the presence/absence of the observed discrepancies.

4 Related Works

Several proposals for representing CGs have been made: in [6,8] the interested reader can find surveys and comparisons of the many formalisms. Recently, there has been a growing interest in declarative paradigms for modeling process models. In [5] the declarative language CIGDEC for clinical guidelines is presented.

CIGDEC lets the user to define constraints that must hold among activities execution, but integration with procedural specifications are not considered. In [3] the authors integrate the domain knowledge within an adaptive Process Management System (PMS), by means of semantic constraints. A generic criterion for semantic correctness of processes is given. However, unlike $\mathcal{S}$CIFF semantic constraints can not take into account contexts, time, locations and data in general.

In [7] the problem of augmenting CG rules with collateral medical information was first pointed out. More recently, in Protocure and Protocure II EU projects there has been an extensive attempt to couple CGs and BMK. CGS are modeled using Asbru, while the BMK is given as a set of formulas in future-time LTL. The theorem prover KIV is used to perform quality checks [2], focussing on properties of CG "per se", and ignoring the conformance dimension.

Acknowledgements. This work has been partially supported by the FIRB project TOCAI.it (RBNE05BFRK) and by the Italian MIUR PRIN 2007 project No. 20077WWCR8.

References

1. Alberti, M., Chesani, F., Gavanelli, M., Lamma, E., Mello, P., Torroni, P.: Verifiable agent interaction in abductive logic programming: the SCIFF framework. ACM Transactions on Computational Logics 9(4) (2008)
2. Hommersom, A., Groot, P., Lucas, P.J.F., Balser, M., Schmitt, J.: Verification of medical guidelines using background knowledge in task networks. IEEE Transactions on Knowledge and Data Engineering 19(6), 832–846 (2007)
3. Ly, L.T., Rinderle, S., Dadam, P.: Integration and verification of semantic constraints in adaptive process management systems. Data Knowl. Eng. 64(1), 3–23 (2008)
4. Field, M.J., Lohr, K.N. (eds.): Clinical Practice Guidelines: Directions for a New Program, Institute of Medicine, Washington, DC. National Academy Press (1990)
5. Mulyar, N., Pesic, M., van der Aalst, W.M.P., Peleg, M.: Declarative and procedural approaches for modelling clinical guidelines: Addressing flexibility issues. In: ter Hofstede, A.H.M., Benatallah, B., Paik, H.-Y. (eds.) BPM Workshops 2007. LNCS, vol. 4928, pp. 335–346. Springer, Heidelberg (2008)
6. Peleg, M., Tu, S., Bury, J., Ciccarese, P., Jones, N., Fox, J., Greenes, R.A., Hall, R., Johnson, P.D., Jones, N., Kumar, A., Miksch, S., Quaglini, S., Seyfang, A., Shortliffe, E.H., Stefanelli, M.: Comparing computer-interpretable guideline models: A case-study approach. JAMIA 10(1) (2003)
7. Shiffman, R.N.: Representation of clinical practice guidelines in conventional and augmented decision tables. J. Am. Med. Inform. Assoc. 4(5), 382–393 (1997)
8. Ten Teije, A., Miksch, S., Lucas, P. (eds.): Computer-based Medical Guidelines and Protocols: A Primer and Current Trends. Studies in Health Technology and Informatics, vol. 139. IOS Press, Amsterdam (2008)
9. Terenziani, P., Montani, S., Bottrighi, A., Molino, G., Torchio, M.: Applying Artificial Intelligence to Clinical Guidelines: the GLARE Approach. In: Teije, et al. (eds.) [8], vol. 139 (July 2008)

Goal-Based Decisions for Dynamic Planning

Elizabeth Black[1], David W. Glasspool[2], M. Adela Grando[2], Vivek Patkar[3], and John Fox[1]

[1] Department of Engineering Science, University of Oxford
[2] School of Informatics, University of Edinburgh
[3] Medical School, University College London

Abstract. The need for clinical guidelines to be implemented at different sites, to adapt to rapidly changing environments, and to be carried out by distributed clinical teams, implies a degree of flexibility beyond that of current guideline languages. We propose an extension to the PRO*forma* language allowing hierarchical goal-based plans. Sub-plans to achieve goals are proposed at runtime so that changing circumstances may be flexibly accommodated without redefining the workflow.

1 Introduction

The effective deployment and maintenance of computer-interpretable guidelines (CIGs [5]) requires a great deal more than simply translating a paper guideline to machine-readable form [7,10]. Firstly, if a clinical guideline is to be rolled out over a number of sites, adapting it to work in each location can require significant effort [10]. Secondly, healthcare is a fundamentally distributed activity, a CIG must provide an execution model that allows the tasks to be delegated between individuals and to be carried out by different members of the clinical team. Thirdly, healthcare environments are highly dynamic, CIGs need to flexibly adapt to changing circumstances. Fourthly, healthcare environments include a great deal of uncertainty, it is typically difficult to predict exactly what the results of a procedure will be before it is carried out.

In this paper we propose a way to implement flexible goal-based plan specialisation within the task ontology of our group's own CIG language, PRO*forma* [3], taking advantage of its argumentation-based decision model to separate decision-relevant knowledge from plan specifications. Our approach allows *a)* flexible local implementation of guidelines taking into account local resources and preferences, *b)* tailoring clinical management based on the patient's response to treatment and *c)* monitoring and manipulating significant clinical goals that are normally implicit in clinical guidelines rather than simple procedural execution of guideline plans which always have a danger of failing. In section 2 we outline the concept, using a concrete example from a complex medical domain (breast cancer treatment) to illustrate the approach. We discuss related approaches and future work in section 3.

Our starting point is the executable process modelling language PRO*forma*, which has a declarative syntax and a well-defined operational semantics [8]. PRO*forma* bases its process model on a minimal ontology of task classes that

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 96–100, 2009.

can be composed into networks representing arbitrarily complex plans or procedures. There are four main task classes derived from the root class in the task ontology (called a "keystone"). *Actions* represent procedures to be executed on the external environment (e.g. administer a drug, update a database). *Enquiries* are carried out to acquire information from some person or external system. *Decisions* define choices about what to believe or what to do. *Plans* group tasks (including other plans) and connect them by simple scheduling constraints, where a task can only be performed if another task has been completed.

The four task types inherit a small set of attributes which control task enactment. *Precondition* is a boolean condition that must be satisfied for the task to start execution; *Postcondition* is a boolean condition asserted as true if the task completes successfully; *Goal* is an informal statement of the goal of the task; and the task *State* underlies the execution semantics for the language. *State* may take one of four values [8]: *Dormant*—task has not yet been started; *In progress*—task has been started and is being executed; *Completed*—task has been completely executed; and *Discarded*—task's scheduling constraints have been satisfied but its preconditions are not true. When a PRO*forma* plan is started all component tasks are in the dormant state. The execution engine repeatedly examines the properties of the tasks in order to determine what state changes and other actions should occur.

2 Goal-Based Hierarchical Planning Based on PRO*forma*

Based on the PRO*forma* task hierarchy, execution semantics and decision model, our proposal adds a new class and new attributes to the task ontology, revises the execution model and defines new scheduling constraints.

In current PRO*forma*, a task may list as "antecedents" a set of tasks that must all be *completed* before it can become *in progress*. In [4], we found this too limited to express several typical workflow patterns and therefore we introduce standard Petri Net scheduling constraints (Begin, End, Join Xor/And, Split Xor/And). We add the new attributes *roles* and *actor* to the root *keystone* class of the task ontology: *roles* is fixed at design time and restricts the possible set of *actors* (which is fixed at runtime and corresponds to the actual executor of the task) allowing delegation of responsibility for tasks. Our proposal also adds a new type of task, *goal*, which inherits from the root task *keystone*. Plans may contain (sub)plans, goals, actions, decisions and enquiries.

Each plan has the new attributes *goalsToAchieve* (the set of goals that the plan could potentially achieve) and *expectedEffects*, which replaces the notion of *postcondition* from PRO*forma* and corresponds to a set of states, not necessarily desirable, that may potentially be satisfied after execution of the plan. Our approach assumes a centralised plan library that might be authored by the relevant governing organisation; typically, each plan in this library will be a collection of goals connected with scheduling constraints, abstracting away from the details of *how* these goals must be achieved. At a local level (say at a particular health authority or hospital) we would expect another plan library with more concrete

plans, defining specific actions, decisions and enquiries to provide locally relevant methods for achieving more abstract goals.

We also assume a repository of clinical knowledge which can be used during plan execution to guide selection of appropriate methods to achieve goals. Again this would be centrally maintained and populated by clinical evidence such as synthesised trial data. This clinical knowledge is applied to the plan execution process via the medium of logical argumentation [2]. Facts and evidence in the repository are composed at run-time into arguments for and against using different plans to achieve outcomes. Arguments have the form:

$\langle PlanName, Sign, Conditions, Outcome, Level \rangle$

where $PlanName$ is the identifier of the plan, $Sign$ gives the polarity of the argument (for or $against$), $Conditions$ is a set of conditions that must be satisfied in order to instantiate the argument, $Outcome$ gives the particular outcome state that the argument refers to, and $Level$ gives a level of support (where this is appropriate to the outcome in question). For example:[1]

"If a patient has early stage breast cancer, then there is an argument for carrying out a mastectomy with axillary clearance as this leads to improved breast cancer specific survival with a likelihood of 97%", denoted as

$\langle mastAxClearance, for, \{earlyBrCa\}, imprBCSS, 97\% \rangle$

(For simplicity we assume that outcome, plan and goal identifiers are standardised at a national level.) These arguments may be synthesised using generic argument schemas [2] or may be provided ready-made in the repository.

During plan execution, when a goal g becomes *in_progress*, all available plan libraries are checked to find the set of candidate plans P that each include g in their *goalsToAchieve* and whose preconditions are met. For each plan in P, the clinical knowledge repository is consulted to construct or find all arguments whose $Outcome$ is in the set of *expectedEffects* for the plan and whose set of $Conditions$ can be satisfied to instantiate the argument.

For example, consider we have the active goal to improve Breast Cancer Specific Survival: $Achieve(imprBCSS)$. Checking the plan libraries to see which candidate plans include this goal as part of their *goalsToAchieve* attribute and whose preconditions are met gives us four candidate plans, however the *role* attributes of two of these plans require special surgical skills that are not available in this particular hospital. Hence, we are left with two candidate plans: carry out a lumpectomy axillary clearance $LumpAxClearPlan$, or carry out a mastectomy axillary clearance $MastAxClearPlan$. We consult the clinical knowledge repository to find the arguments whose $Outcome$ is in the set of *expectedEffects* for these two plans and whose set of conditions are satisfied. This gives us the following set of arguments that are displayed to the user(s).

$$\langle MastAxClearPlan, for, \{earlyBrCa\}, imprBCSS, 97\% \rangle$$
$$\langle MastAxClearPlan, against, \{earlyBrCa\}, lossShoulderMob, 20\% \rangle$$
$$\langle MastAxClearPlan, against, \{earlyBrCa\}, lossSelfImage, likely \rangle$$
$$\langle LumpAxClearPlan, for, \{earlyBrCa\}, imprBCSS, 95\% \rangle$$

[1] We make no claims about the clinical accuracy of any of the examples in this paper.

$\langle LumpAxClearPlan, against, \{earlyBrCa\}, lossShoulderMob, 20\% \rangle$
$\langle LumpAxClearPlan, against, \{earlyBrCa\}, extraRad6weeks, 95\% \rangle$

A decision mechanism using argumentation-based reasoning (e.g. [2]) can now be applied in order to automatically recommended one or more candidate plans. However, here we allow the user to choose a candidate based on their valuation of the different outcomes. They select the less extensive surgical procedure *LumpAxClearPlan* because the patient highly values preserving her self image.

Goals have the attributes *success_condition* and *abort_condition*. When a goal first becomes *in progress*, the *success_condition* is checked; If it evaluates as true, then the state of the goal moves straight to *completed*. The *success_condition* is also checked when a goal is *in progress* and a candidate plan that has been selected to achieve it has either become *completed* or *discarded*; If the condition is true, then the goal state changes to *completed*; If the condition is false, then the goal remains *in progress* and another candidate plan must be selected to achieve the goal. In this manner, we provide some flexibility to deal automatically with unexpected failure of methods. If the *abort_condition* of an *in progress* goal becomes true at any time, then the goal becomes *discarded* and any *in progress* plan that is being used to try to achieve the goal also becomes *discarded*.

Continuing the running example, the sub plan *LumpAxClearPlan* is completed and the tumour is completely removed. However, histo-pathology report suggests the cancer has spread to the lymph nodes and is ER negative. Therefore the success condition of the goal $Achieve(imprBCSS)$ is not yet satisfied and so this goal is still *in progress*. The plan repository is searched again and two alternative plans *AdjChemo1Plan* and *AdjChemo2Plan* (adjuvant chemotherapy plans 1 and 2) are generated as candidate plans (as preconditions on both these candidate plans are true and both plans have as part of their *goalsToAchieve* the goal *Achieve(imprBCSS)*). As before, arguments for and against these candidate plans are constructed from the clinical knowledge repository, allowing the users to select $AdjChemo1Plan$. However, whilst this plan is *in progress*, the patient develops distant cancer metastasis and the disease stage is no longer early Breast Cancer ($earlyBrCa$). Because *not(earlyBrCa)* is an *abort_condition* on our goal *Achieve(imprBCSS)* the goal is now *discarded* and so the remaining part of *AdjChemo1Plan* is also discarded.

3 Discussion

The Asbru language [6] has adopted a similar general approach to hierarchical refinement of plans during execution by selection of appropriate sub-plans from a library based on goals ("intentions" in Asbru) and pre– and post–conditions. Generic workflows that can be specialised at run time are also presented in [9] based on a state (referred to as a "scenario").

We believe that three aspects of the present work are particularly distinctive: Firstly, in our approach goals are first-class members of the task hierarchy, not simply attributes of tasks. This allows complete separation between the goal and possible procedures that could achieve it, improving the flexibility and clarity of

the approach in our view. Secondly, the use of *argumentation logic* to abstract decision-relevant knowledge away from plan specifications allows clinical knowledge to be maintained entirely separately from the plan library. The approach also allows the information presented to the user to be highly focussed and appropriate. This has proved beneficial in many clinical decision systems [2]. Thirdly, the PRO*forma* task hierarchy is simple and provides a sound basis for this extension. The execution semantics are essentially defined for the *keystone* class and need little adjustment to support the new *goal* class. Finally the support for roles and actors provides a basis for a move towards a more distributed setting, where multiple members of the clinical team work on multiple care plans in parallel.

Future work will take two directions. Firstly we propose to extend the argumentation model to support patients' and clinicians' *values* (see [1]), allowing finer-grained personalisation of the detailed decomposition of care plans. Secondly, the specification of *roles* provides a means of delegating responsibility for achieving a goal to another computer system or another clinician, the goal-based abstraction allowing the delegating system to ignore the details of what actions the delegate will take.

Acknowledgements. This work was supported by EPSRC grant EP/F057326/1 and by a programme grant from Cancer Research UK to J. Fox and D. Glasspool.

References

1. Atkinson, K., Bench-Capon, T.J.M.: Practical reasoning as presumptive argumentation using action based alternating transition systems. Artificial Intelligence 171(10-15), 855–874 (2007)
2. Fox, J., Glasspool, D.W., Grecu, D., Modgil, S., South, M.: Argumentation-based inference and decision-making—a medical perspective. IEEE Intelligent Systems 22, 34–41 (2007)
3. Fox, J., Johns, N., Rahmanzadeh, A.: Disseminating medical knowledge—the PRO*forma* approach. Artificial Intelligence in Medicine 14, 157–181 (1998)
4. Grando, M.A., Glasspool, D.W., Fox, J.: Petri nets as a formalism for comparing expressiveness of workflow-based clinical guideline languages. In: 2nd Int. Workshop on Process-Oriented Information Systems in Healthcare (2008)
5. Peleg, M., Tu, S., Bury, J., Ciccarese, P., Fox, J., Greenes, R.A., Hall, R., Johnson, P.D., Jones, N., Kumar, A., Miksch, S., Quaglini, S., Seyfang, A., Shortliffe, E.H., Stefanelli, M.: Comparing computer-interpretable guideline models: a case-study approach. J. American Medical Informatics Assoc. 10(1), 52–68 (2003)
6. Shahar, Y., Miksch, S., Johnson, P.: The Asgaard project: A task-specific framework for the application and critiquing of time-oriented clinical guidelines. Artificial Intelligence in Medicine 14, 29–51 (1998)
7. Shiffman, R., Greenes, R.: Improving clinical guidelines with logic and decision-table techniques. Medical Decision Making 14(3), 247–343 (1994)
8. Sutton, D.R., Fox, J.: The syntax and semantics of PRO*forma* guideline modelling language. J. Am. Med. Inform. Assoc. 10(5), 433–443 (2003)
9. Tu, S., Musen, M.A.: A flexible approach to guideline modeling. In: Proc. AMIA Symposium, pp. 420–424 (1999)
10. Waitman, L.R., Miller, R.A.: Pragmatics of implementing guidelines on the front lines. J. American Medical Informatics Assoc. 11(5), 436–438 (2004)

Genetic Algorithm Based Scheduling of Radiotherapy Treatments for Cancer Patients

Dobrila Petrovic[1], Mohammad Morshed[1], and Sanja Petrovic[2]

[1] Faculty of Engineering and Computing, Coventry University, Coventry, UK
{D.Petrovic,M.Morshed}@coventry.ac.uk
[2] School of Computer Science, University of Nottingham, Nottingham, UK
sxp@cs.nott.ac.uk

Abstract. This paper presents a multi-objective model for scheduling of radiotherapy treatments for cancer patients based on genetic algorithms (GA). The model is developed and implemented considering real life radiotherapy treatment processes at Arden Cancer Centre, Coventry, UK. Two objectives are defined: minimisation of the Average patient waiting times and minimisation of Average tardiness of the patient first treatment fractions. Two scenarios are analysed considering the availability of the doctors to approve treatment plans. The schedules generated by the GA using real data collected from the collaborating Cancer Centre have good performance. It is demonstrated that enabling doctors to approve treatment plans instantly has a great impact on Average waiting time and Average tardiness for all patient categories.

Keywords: Radiotherapy, Genetic Algorithms, Scheduling, Waiting Times.

1 Introduction

Research in scheduling theory has evolved over the years and has been the subject of much literature [1]. Scheduling patients in the health care domain has attracted considerable researchers' attention [2], [3]. However, there are relatively few papers that treat radiotherapy patient scheduling. Among the first approaches to radiotherapy patient scheduling was given in [4]. In a recent study [5], algorithms for booking treatments for radiotherapy patients were proposed. These algorithms were aimed at reducing patients waiting time for the first treatment sessions based on the target waiting times framed by JCCO (Joint Council of Clinical Oncology) [6].

Generally, the exact methods cannot be applied to generic radiotherapy treatment scheduling problems due to the complexity of constraints and the size of problems. A novel multi-objective GA has been proposed to handle a patient scheduling problem identified in Arden Cancer Centre, University Hospitals Coventry and Warwickshire, NHS Trust, UK.

The paper is organised in the following way. In Section 2, the radiotherapy treatment process under consideration and problem statements are defined. Section 3 is dedicated to the description of the developed multi-objective GA, while Section 4 presents the analysis of computational results obtained. Section 5 provides the conclusion and directions for future work.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 101–105, 2009.

2 Radiotherapy Treatment Process and Problem Statement

The radiotherapy treatment process in Arden Cancer Centre includes four different stages: planning, physics, pre-treatment and treatment. Each patient follows the treatment path set by an assigned doctor. The doctor has to be available to approve and sign the treatment plan for each patient. The patients' images and documents pass from the planning to the physics unit or directly to the pre-treatment unit depending on the complexity of the cancer. Dosimetry calculations for complex cases are carried out in the physics unit and re-checked in the pre-treatment unit. Simple cases are considered directly in the pre-treatment unit. The patients are then booked for a prescribed treatment machine for specified number of fractions [3].

The radiotherapy patient scheduling is considered as a multi-objective scheduling problem with recirculation. A patient may visit a given machine several times on his/her treatment path. Two objectives relevant for radiotherapy scheduling problems are defined: minimisation of Average waiting time and minimisation of Average tardiness of the patients. The waiting time is defined as the time that elapses from the decision to treat the patient until the first treatment fraction administration. The following main assumptions are made: the radiotherapy units follow a five day working week, doctors work on rota, details of all patients to be scheduled are known in advance, and the prescribed numbers of fractions have to be administered on consecutive days.

3 Multi-objective GA for Radiotherapy Treatment Schedules

A multi-objective GA is developed to generate a schedule for radiotherapy patients. A GA is a search algorithm inspired by natural selection and genetics [7], [8]. It uses a population of candidate solutions represented as strings which are evaluated by a fitness function. New solutions are generated using two operators: crossover, which combines two solutions and generates a new solution by replacing a part of one string, usually randomly selected, with a corresponding part of another string, and mutation, which is used to alter one or more, usually randomly selected parts of one string.

In this paper, we use an operation based string representation, which indirectly represents a schedule [9]. For example, in the case of four patients and four machines the string can have the following form [3-1-3-4-2-1-2-1-2-3-4-3-2-1-4-4], where all operations for a patient are named using the patient-id. In the given example, the first operation to be scheduled is the first operation of patient 3, then the first operation of patient 1, second operation of patient 3, and so on. The last operation for each patient is the administration of the first fraction on the prescribed treatment machine. The strings are of equal size. In a case when patient does not require the maximum number of operations, the remaining operations are still specified in the string with patient-id but related processing times are set to 0. A good feature of this schedule representation is that it always represents a feasible solution. The crossover and mutation operators applied on strings preserve the number of operations for each patient.

The fitness function is defined considering two criteria: Average waiting time and Average tardiness of the patients. It is used wherever the evaluation of string takes place. The waiting time and tardiness have different scales (waiting time is longer than tardiness), and, therefore, the corresponding values have to be normalized, before they

are summed to form a single fitness values of a string. Normalisation is carried out as a linear mapping of the interval formed by the minimum and the maximum achieved values of each of the objective functions into the interval [0, 1]. A string needs to be decoded into the corresponding schedule before evaluation. The operations are scheduled according to their sequence in the string in such a way that each operation is allocated the earliest available time on the machine the operation requires.

The initial string population is created using some simple strategies, e.g., sequence all operations of the patient 1 first, then patient 2, etc. or sequence the first operation of all patients, then second operations of all patients, etc. and finally strings are generated randomly, until the whole population of the specified size is created. In each iteration, the elitist strategy is applied, i.e., a certain number of the strings with highest values of the fitness function are directly input into the population of the subsequent iteration.

4 Analysis of the Results

The multi-objective GA described above is applied to generate schedules for radiotherapy patients in Arden Cancer Centre on daily basis. There are 13 resources available including a simulator, CT scanner, mould room and doctors in the planning unit, 1 physics unit, 1 pre-treatment unit, and 7 treatment machines including 3 high energy linear accelerators (linacs), 2 low energy linacs, 1 Deep X-Ray and 1 Beta-tron.

Real data collected from Arden Cancer Centre were used to develop a simulation model for the radiotherapy processes. The simulation model is used to generate data about each newly arriving patient [3]. Based on the historical data, it is estimated that the daily number of newly arrived patients has a Poisson distribution.

The parameters of GA were tuned and selected using tests which were run up to 100 generation on different combinations of parameters. The selected parameters of the GA were: Number of generation = 50, Population size = 100, Crossover rate = 0.80, Mutation rate = 0.03, Number of best solutions which are passed to successive generation = 5, and Daily number of patients has a Poisson distribution with expected rates 8.88, 7.76, 7.47, 6.59 and 11.6 for Monday to Friday, respectively.

The GA is used to generate schedules for "a worming up period" of 57 days during which the available times of the machines, facilities and doctors were partially booked. The GA was then used to generate a schedule for newly arrived patients on one day and the scheduling performance was evaluated considering only those patients.

Due to the stochastic nature of the GAs and uncertainty in the daily number of newly arrived patients and their categories, we used the GA to generate schedules for

Table 1. Scheduling performance of 10 runs of 10 different daily sets of patients

Test sets	1	2	3	4	5	6	7	8	9	10
Average$^{\Delta}$	12.35	13.68	15.36	12.70	15.72	14.53	13.97	15.11	11.18	19.52
Best$^{\Delta}$	9.34	11.16	13.84	11.16	14.86	13.65	13.42	14.70	10.65	18.25
Worst$^{\Delta}$	14.46	16.38	17.11	15.35	16.28	15.32	15.02	16.42	12.69	21.86
Average*	2.29	0.98	0.96	0.83	0.67	1.61	0.98	0.91	0.82	2.01
Best*	1.96	0.72	0.91	0.72	0.31	1.53	0.82	0.32	0.31	1.94
Worst*	4.01	1.94	1.04	0.92	1.27	1.87	1.03	0.94	0.95	3.54
	$^{\Delta}$Average Waiting times					*Average Tardiness of the patients				

10 different sets of daily arrived patients and repeated it 10 times for each set of patients with different initial populations. The ranges of the obtained objective function values, given in Table 1, are not wide for all 10 daily sets of patients.

Doctor's availability has been identified as one of the bottlenecks in achieving the target waiting times. Therefore, the Cancer Centre is considering various ways to speed the ways the doctors can sign treatment plans. We analysed the schedule performance in two scenarios: I - Doctors can sign treatment plans only if on rota, and II - Doctors can sign treatment plans promptly (e.g., on-line or using pagers, etc.). The experiment of generating a daily schedule was carried out for 10 different sets of daily arriving patients for both scenarios I and II. The performances of the schedules, Average waiting time and Average tardiness, for each of the 10 days are illustrated in Fig. 1 and 2, where R, P and E stand for Radical, Palliative, and Emergency patients.

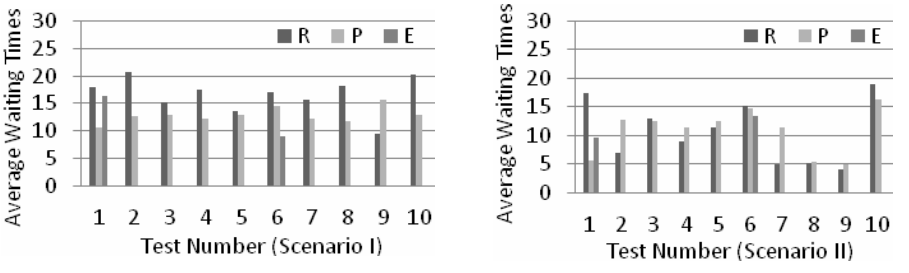


Fig. 1. Comparison of Average waiting times (in days) obtained for different patient categories

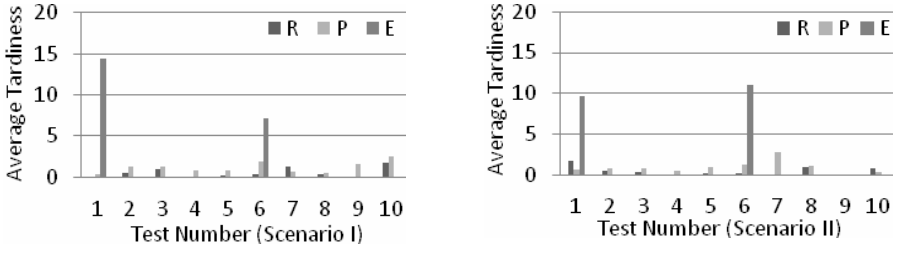


Fig. 2. Comparison of Average tardiness (in days) obtained for different patient catagories

The obtained results showed that enabling doctors to approve the treatment plans promptly had a great impact on the schedule performance. The average of Average waiting time achieved in 10 days was reduced from 16.54, 12.84, and 12.64 days to 10.63, 10.77, and 11.53 days for radical, palliative and emergency patients, (by 35%, 16%, and 8%, respectively). Similarly, the average of Average tardiness was reduced from 0.52, 1.14, and 10.79 days to 0.45, 0.91, and 10.28 days for radical, palliative and emergency patients, (by 13%, 20%, and 4%, respectively). It was noticed that the emergency patient's had the higher waiting times then recommended by JCCO, while waiting times for palliative and radical patients were within the JCCO targets.

It is interesting to notice that in both scenarios the average of Average waiting times achieved in all 10 days for radical and palliative patients were better than the current Average waiting times recorded using the real data, collected in 2007, that were 35 days for radical and 15 days for palliative patients. We should point out that the cancer

centre does not operate during holidays, and hospital does use overtime working hours, in special situations. These assumptions were not considered in the developed GA.

5 Conclusions

A multi-objective GA for scheduling of radiotherapy patients is presented. The GA is applied to a real life radiotherapy problem. Following the recommendation concerning the target patient waiting times set by JCCO, two objectives relevant to radiotherapy scheduling optimisation are defined: (1) minimisation of Average waiting times and (2) minimisation of Average tardiness. The performance of generated radiotherapy schedules are measured for two possible scenarios regarding the timing of doctor's approvals of patient's treatment plans. The obtained results proved that enabling doctors to sign treatment plans rapidly has a great impact on scheduling performance. The average waiting time and tardiness can be reduced by 35% and 20% respectively.

The future work will be carried out in different directions as follows: (1) New objectives relevant for measuring the performance of a radiotherapy schedule will be included, such as the minimisation of maximum tardiness. (2) Different experiments will be carried out, for example, to analyse the effects of reserving certain time slots on the treatment machines for the emergency patients. (3) The domain knowledge will be identified and formalized to be included into the procedures of generating initial solutions and the GA operators in order to improve efficiency of generated schedules.

Acknowledgement. The authors would like to thank the Engineering and Physical Sciences Research Council (EPSRC) for supporting this research, grant no - EP/C549511 and EP/C54952X/1, and Arden Cancer Centre, UK for their collaboration.

References

1. Pinedo, M.: Scheduling: Theory, Algorithms and Systems. Prentice-Hall, New Jersey (2002)
2. Conforti, D., Guerriero, F., Guido, R.: Optimisation Models for Radiotherapy Patient Scheduling. 4OR: Quart J. Ops. Res. 6(3), 263–278 (2008)
3. Kapamara, T., Sheibani, K., Petrovic, D., Haas, O., Reeves, C.R.: A Simulation of a Radiotherapy Treatment Systems: A Case Study of a Local Cancer Centre. In: Proc. of ORP3 Conference, Guimaraes, Portugal, pp. 29–35 (2007)
4. Larsson, S.N.: Radiotherapy Patient Scheduling Using a Desktop Personal Comp. J. Clinical Oncology 5, 98–101 (1993)
5. Petrovic, S., Leung, W., Song, X., Sundar, S.: Algorithms for Radiotherapy Treatment Booking. In: 25th Workshop of the UK Planning and Scheduling Special Interest Group, Nottingham, UK, pp. 105–112 (2006)
6. Joint Collegiate Council for Oncology (JCCO). Reducing Delays in Cancer Treatment: Some Targets. Technical report, Royal College of Physicians, London (1993)
7. Deb, K.: Multiobjective Optimisation using Evolutionary Algorithms. John Wiley & Sons, New York (2001)
8. Goldberg, D.: Genetic Algorithms in Search, Optimisation & Machine Learning. Addison Wesley, Reading (1989)
9. Gen, M., Tsujimura, Y., Kubota, E.: Solving Job-shop Scheduling Problems by Genetic Algorithm. In: Proc. IEEE International Conference on Systems, Man, and Cybernetics, Texas, vol. 2, pp. 1577–1582 (1994)

Feasibility of Case-Based Beam Generation for Robotic Radiosurgery

Alexander Schlaefer[1,2] and Sonja Dieterich[2]

[1] Medical Robotics, University of Lübeck, Lübeck, Germany
[2] Department of Radiation Oncology, Stanford University, Stanford, USA

Abstract. Robotic radiosurgery uses the kinematic flexibility of a robotic arm to target tumors and lesions from many different directions. This approach allows to focus the dose to the target region while sparing healthy surrounding tissue. However, the flexibility in the placement of treatment beams is also a challenge during treatment planning. So far, a randomized beam generation heuristic has been proven to be most robust in clinical practice. Yet, for prevalent types of cancer similarities in patient anatomy and dose prescription exist. We propose a case-based method to solve the planning problem for a new patient by adapting beam sets from successful previous treatments. Preliminary experimental results indicate that the novel method could lead to faster treatment planning.

1 Introduction

The therapeutic use of radiation plays an important role in the treatment of cancer. A sufficiently high dose of ionizing radiation will typically cause cell death and presents an effective way to kill tumor cells. We consider the most common form of radiation treatment, where beams of radiation are delivered from outside the patient and typically pass through healthy tissue before reaching the target. Hence, the treatment planning problem consists of maximizing the therapeutic effect while minimizing potential side effects.

Conventionally, side effects are limited by fractionation, i.e., delivering the dose in a large number of small fractions. This approach is based on the different ability of normal and cancerous cells to recover from radiation damage. Clearly, if the shape of the high dose region is highly conformal to the shape of the tumor with a steep dose gradient towards surrounding tissue, the side effects will be small compared to the effect in the tumor. Hence, fewer and more effective fractions can be applied. Recently, robotic beam delivery has enabled such focused dose delivery [1].

While 3D conformal and intensity modulated radiation therapy (IMRT) typically use 5 - 9 beam directions [2], more than 100 directions are commonly considered with the robotic system. Moreover, since conventional systems use a radiation source mounted on a gantry, coplanar beam configurations are often preferred in practice. In contrast, the robotic arm facilitates placement of beams from many different non-coplanar positions and with arbitrary orientation. A set of typically 100 - 300 individually weighted beams is selected for treatment, allowing for very tumor conformal distributions.

The treatment planning problem is solved by identifying a set of beams and beam weights such that the resulting dose distribution represents the optimal trade-off with

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 106–115, 2009.

Fig. 1. The robotic CyberKnife system by Accuray Inc. The system consists of a robotic arm (1), linear accelerator (2), X-ray sources (3) and cameras (4), a robotized patient couch (5), and an optical tracking system. (Image courtesy of Accuray Inc.)

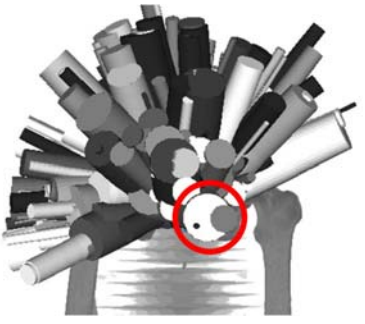


Fig. 2. An illustration of the set of treatment beams used to irradiate a prostate tumor. The orientation, diameter, and weight of the beams is shown, with darker and lighter beam color indicating lower and higher weight, respectively. The beams inside the circle start in the same node, but have different diameter and orientation.

respect to various clinical criteria. Typical criteria include constraints on the minimum and maximum dose in the tumor and in critical structures, the shape of the dose distribution, and the total treatment time. Given the large number of potential beam configurations, the resulting optimization problem is typically degenerate, i.e., there is a number of different but clinically acceptable solutions.

Still, finding a suitable clinical plan can be a time consuming iterative process. While a number of approaches to guide the search have been proposed [3,4,5,6], one of the limiting factors is the sheer size of the combinatorial problem to select the subset of beams. Most planning approaches separate the beam orientation and the beam weighting problem. For conventional radiation therapy, numerous exhaustive and heuristic search schemes to identify suitable beam orientations have been studied [2,7,8]. Besides, from manual planning and as a result of the underlying physics, simple standard beam sets have emerged for prevalent tumors [9,10,11,12,13]. These class solutions are based on similarities in the anatomy, and the actual direction of the beams is modified by the human planner to account for anatomical features of the specific patient. Essentially, class solutions employ some of the core principles of case-based reasoning: a similar planning problem can be solved by a similar beam set, and the actual beam directions are adapted to the specific problem.

For robotic radiosurgery, the problem is much more complex, and no standard solutions are known. Finding a spatial arrangement and shape for up to 300 treatment beams corresponds to identifying the actual apertures in IMRT, a problem decisively different from the above mentioned beam orientation problem. While this problem has not been regarded in a case-based fashion, the idea to consider solutions from similar clinical cases seems promising. We study an approach to retrieve candidate beams from a

database of similar cases, and to adapt these beams to be useful for the current planning problem. The method is applied to a small test set of prostate cases, and our results indicate that case-based treatment planning for robotic radiosurgery is feasible.

2 Robotic Radiosurgery

In robotic radiosurgery, megavoltage X-ray beams are delivered by a lightweight linear accelerator mounted on a 6-degrees-of-freedom robotic arm (Fig. 1). The workspace of the robot is limited to a set of points distributed on a roughly spherical surface around the patient. Each point is selected such that no collision between robot and patient or other system components occurs, and referred to as beam node. Usually, more than 100 beam nodes are considered during planning. While the node determines the starting point of a beam, its orientation is arbitrary. The collimators used for robotic radiosurgery have a circular shape and currently twelve different diameters, ranging from 5 to 60 mm in the target region, can be selected. Another degree of freedom is the beam weight, i.e., the time for which each beam is switched on and delivers dose. As a physical parameter the beam weight is bound to be non-negative. Figure 2 illustrates the set of treatment beams for a prostate tumor treatment. Note that the beams are only shown in the proximity of the patient, the actual distance from beam source to target is approximately 0.8 m.

Popular choices for optimization of treatment plans in radiation therapy include gradient decent methods, simulated annealing, and linear programming. The latter has been successfully used for robotic radiosurgery, particularly due to its completeness with respect to the considered beam set. Planning typically proceeds as follows. First, a set of candidate beams is randomly selected. Second, the optimization problem is solved. Third, the subset of beams with non-zero weight is selected for treatment. Clearly, this approach depends on the choice of the candidate set. For example, if we consider 100 beam nodes and 12 collimator diameters, and a sample of 10 random beam orientations per node and diameter, we get 12000 candidate beams. While a larger set will generally improve the resulting plan quality [14], it also causes long and sometimes prohibitive planning times.

3 Cases and Similarity

In the context of beam generation, a case could be defined by the complete clinical problem and the related set of weighted treatment beams as a solution. However, this definition would include a large set of features and establish a fixed geometric relationship between all beams and the patient's anatomy. Thus, finding two sufficiently similar cases will be rather unlikely. Moreover, the foremost goal of this study is to identify a more suitable set of candidate beams and therefore it is not necessary to find a perfect match.

To generate a candidate beam set, we consider each beam node separately. A case then consists of features describing the planning problem with respect to beam node n, and the set of beams starting in n forms the solution. As case-based reasoning is built on the assumption that similarity implies usefulness of a known case with respect to a new problem [15], we need to identify features that capture the core aspects of treatment

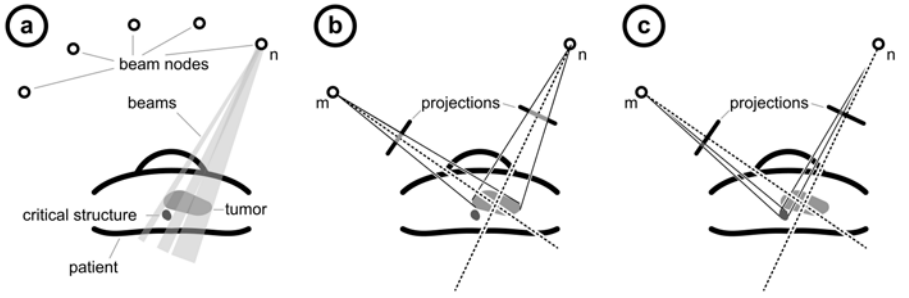


Fig. 3. An illustration of the planning problem: a) the general problem and a set of beams starting in node n, b) projections of the tumor with respect to nodes m and n, c) projections of the critical structure with respect to nodes m and n. Note that for n the two projections would overlap.

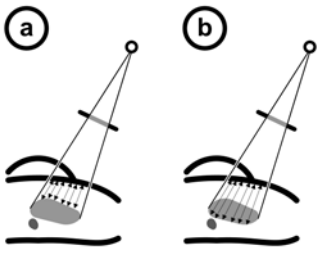


Fig. 4. Two projection matrices containing a) the minimum and b) the maximum radiological depth are computed to represent location and shape of the structures.

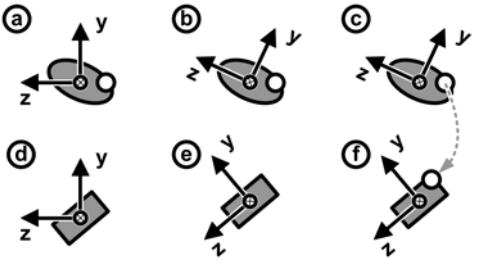


Fig. 5. Adaption is performed by aligning the y- ad z-axis of the coordinate system with the principle components of the tumor (a,b,d,e). Beams represented with respect to the coordinate system of the case can be transformed using the coordinate system of the problem (c,f).

planning. Obviously, the shape and location of the tumor and critical structures need to be considered. Other parameters include the dose bounds for each structure.

A common tool to assess the utility of beams in conventional radiation therapy is the beam's-eye-view (BEV), i.e., a projection of the considered structures onto a plane perpendicular to the beam's central axis [16,17]. We consider a similar projection with respect to a plane perpendicular to a line from the beam node to the centroid of the target. Figure 3 illustrates a beam node n with a set of beams targeting the tumor, and the projections of tumor and critical structure for two nodes m and n. Note that the projections for m and n differ in size. Moreover, for n the projections of the two structures overlap, i.e., beams passing through both structures are typically less useful.

While the projections contain information about the outline of the structures, the actual three-dimensional shape and the location are not captured. Interestingly, the effect of radiation depends mostly on the type of tissue passed. For example, the attenuation of a beam is different when traveling through lung tissue, muscle, or bone. Hence,

instead of computing the geometrical distance between a beam node and the structures, we consider their radiological depth. Figure 4 illustrates how two depth projections per structure comprise information on location and shape with respect to the beams' effect.

To represent the beams and the minimum and maximum depth projection for the various structures relative to a beam node, we introduce a cartesian coordinate system with its x-axis following the line from beam node to target centroid. Now we can define a case more formally. The features of the problem part include a cartesian coordinate system C, and for each structure the minimum and maximum dose bound as well as the minimum and maximum depth projection. The solution contains a set of beams, where each beam is defined by its diameter and its vector of orientation with respect to C.

The local similarity measure for the dose bounds is computed as

$$sim(b_p, b_c) = 1 - \frac{|b_c - b_p|}{b_c + b_p}$$

where b_c and b_p are the non-negative dose bounds for case and new problem, respectively. When comparing two depth projections, the two-dimensional matrices are aligned according to the related coordinate systems. Local similarity is calculated based on the Euclidian distance over all overlapping matrix elements. Let A_p and A_c be the sets of non-negative matrix elements for problem and case, and let A denote the set of tupels $\{a_p, a_c\}$ of corresponding, non-negative elements, then similarity is calculated as

$$sim(A_p, A_c) = \frac{1}{1 + \sqrt{\sum_i (A_{i_1} - A_{i_2})^2}} * \frac{2\,|A|}{|A_p| + |A_c|}$$

where the first term penalizes a difference in matrix elements of the overlapping region, i.e., the depth values, and the second term penalizes a difference in the shape of the two projections. In our inital experiments we established global similarity as a product of local similarities.

4 Beam Generation and Adaption

The generation of a candidate beam set for a new patient proceeds as follows. First, the projection matrices and the coordinate systems are computed for each beam node. Second, for each of the resulting problems the most similar cases are retrieved from the case base and sorted in descending order of similarity. In order to compute the similarity between the projection matrices, they need to be aligned in a meaningful way. While any image registration method could be applied, a fast algorithm is preferable to speed up the retrieval. We perform a principle component analysis (PCA) on the tumor to align the coordinate system with the tumor projection, and we consider the mirrored matrices to account for different projection angles from different beam nodes. Third, starting with the most similar case, beams are added to the candidate beam set. The required adaption needs to map beams given with respect to the retrieved case into the current beam node. As the y- and z-axis of both coordinate systems are alinged with the principle components of the tumor projection, the beam orientation can be expressed by the same vector for case and problem, see Figure 5. The actual candidate beams are generated by transforming the vector from beam node coordinates into patient coordinates.

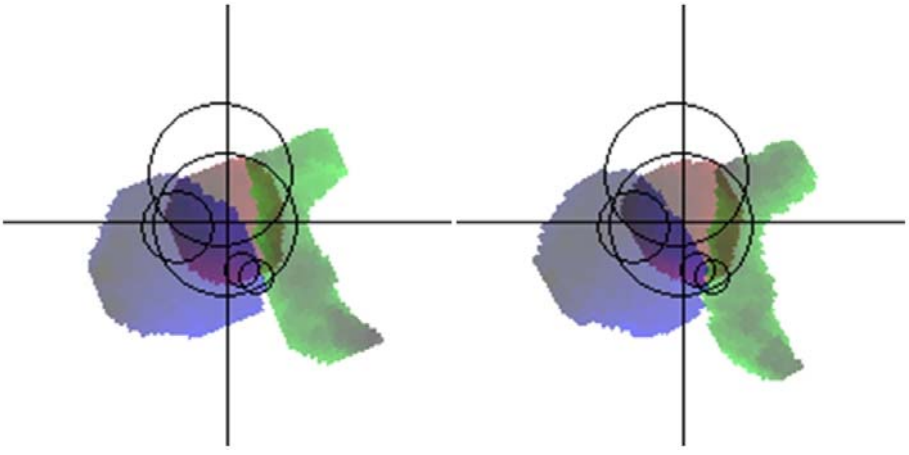


Fig. 6. Two different projections for a prostate case (red: prostate, blue: bladder, green: rectum) illustrating the idea to use an acceptable beam arrangement for one node (left) as candidate beams for a node with similar projections (right)

Adaption of the beam weights is then performed by solving the linear optimization problem.

Figure 6 illustrates the adaption for two beam nodes. The red, green, and blue areas are the projection of prostate, rectum, and bladder, respectively. Darker pixels denote a larger radiological depth. A careful inspection of both sides reveals that the structures

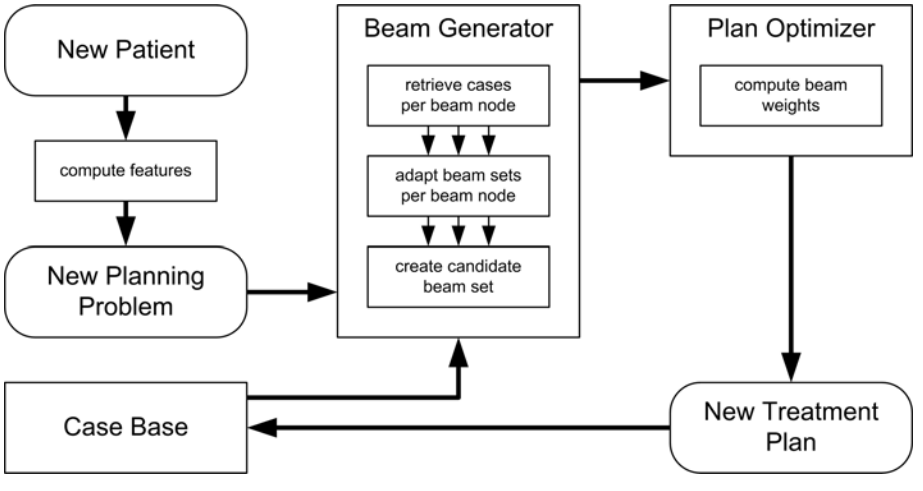


Fig. 7. The overall planning approach. Starting with a new patient, the planning problem is established, beams a generated, and the optimal set of treatment beams is fed back into the case base.

Table 1. The set of 10 prostate patient cases considered for our tests. An active beam node has at least one beam with non-zero weight, and each active beam node forms a case.

Case	Type	Tumor volume (cm^3)	Active beam nodes
20036	Prostate	26.1	75
20148	Prostate	85.1	49
20209	Prostate	99.6	50
20218	Prostate	45.6	49
20256	Prostate	121.2	53
20292	Prostate	88.5	74
20322	Prostate	71.2	51
20357	Prostate	97.3	50
20511	Prostate	61.4	67

are viewed from different directions. Yet, as the projections are aligned according to the PCA, the set of beams can be mapped from the case (left side) to the problem (right side).

The overall planning approach is summarized in Figure 7. For a new clinical case the computed tomography (CT), the contours of the structures of interest, and the position of the beam nodes with respect to the CT are known. The planning problem is established by computing the set of features for each beam node, and effectively breaking the complex problem into a set of simpler sub-problems. The beam generator compares each sub-problem against the case base and adapts the beams of the most similar cases to the respective beam node. All resulting beam sets are merged into one candidate beam set and fed into the plan optimizer. Here, the plan optimizer can be seen as a second stage adaptor that unifies and adapts the sub-solutions to form the global solution. When the weighted beams form an acceptable treatment plan, the subset of beams with non-zero weight is stored in the case base. Again, this is done per beam node, i.e., each sub-problem and the corresponding sub-solution form a case.

5 Experimental Results

In order to study the potential benefit of the proposed beam generation method, we consider a small test set of clinical prostate cases summarized in Table 1. All cases have acceptable solutions obtained with the randomized beam generation method. Note that this retrospective study was institutional review board (IRB) approved and the case numbers are not related to patient IDs.

Obviously, a beam set suitable for a large tumor will be less useful for a very small tumor, and vice versa. Therefore we chose cases 20148 and 20292 as out test set, as for both cases similarly sized prostates remain in the training set. For each of the active beam nodes of every patient case in the training set, a case was generated and stored in the case base.

In order to get comparable results, the beam generation was slightly modified. When generating the candidate beam set, beams from each beam node were added in a round-robin fashion until a preset threshold was reached. For comparison, 10 random beam

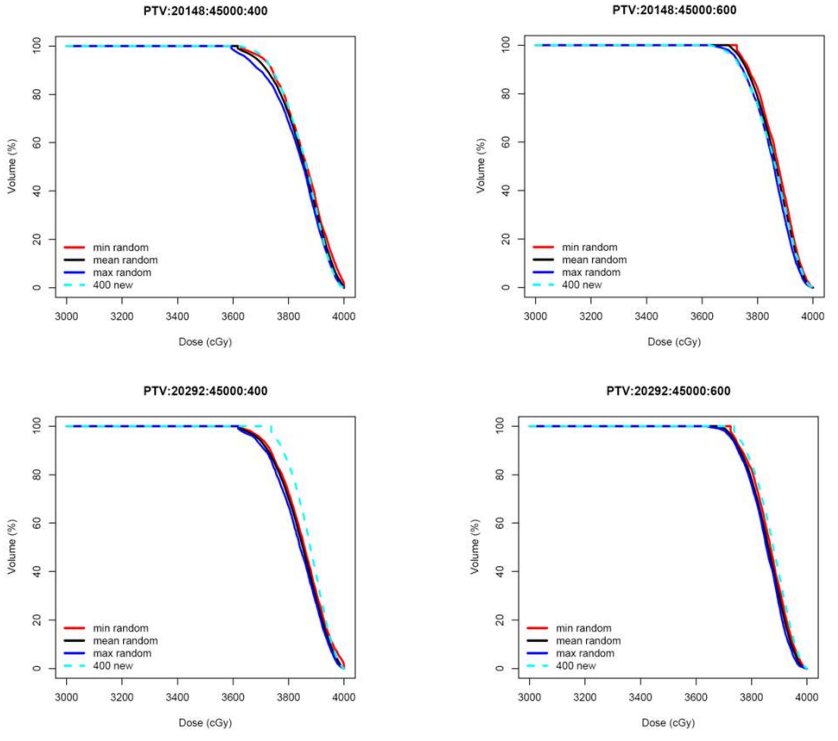


Fig. 8. Results for cases 20148 and 20292 comparing the case based approach with randomized planning. The dose-volume histogram (DVH) for 400 candidate beams generated with the new method (cyan) are compared to the best (blue), average (black), and worst (red) results for randomized candidate beam sets of 400 and 600 beam, respectively. A DVH is a cummulative histogram relating the dose d to the volume covered by at least d. Clearly, the complete prostate is covered by the minimum dose in the prostate. As lower doses may compromise the effect of the treatment, our objective is to increase the minimum dose.

sets of different size where generated for each test case. Figure 8 shows the results for case based candidate beam sets with 400 beams, and randomized candidate beam sets with 400 and 600 beams, respectively. The planning objective was to improve the dose homogeneity, i.e., to increase the minimum dose in the prostate. Clearly, randomized approaches could, by chance, find a perfect solution. However, they can also be arbitrarily bad, and the results indicate that the case-based method outperforms the randomized approach for similar sized candidate beam sets. Even for the larger randomly generated sets the results remain comparable.

The advantages of the case-based approach come at the cost of additional runtime for the case retrieval. However, for our example the overhead is relatively small. While generating 400 or 600 random beams is almost instantaneous, the CBR approach with

400 beams takes 12.0 s and 17.3 s for patient cases 20148 and 20292, respectively. The time spent for dose calculation and setup of the optimization problem is similar for 400 beams. For 600 random beams dose calculation and setup take approximately 8 s longer than for 400 CBR generated beams. The main difference is in the optimization time, though. For clinical case 20148, the optimization takes 498.2 s and 1121.8.2 s with 400 and 600 random beams, respectively. Optimization with 400 CBR generated beams takes 711.7 s. Similarly, for clinical case 20292, the optimization takes 457.3 s and 1003.2 s with 400 and 600 random beams, and 626.2 s with 400 CBR generated beams.

6 Discussion

The results demonstrate the feasibility of a case based approach to improve beam generation for robotic radiosurgery. While the retrieval adds to the runtime, the overhead is relatively small compared to the impact of a larger candidate beam set. To obtain similar plans, the case-based method would run for approximately 13 min and 11 min, while the randomized method would take approximately 19 min and 17 min, respectively.

Our current results are preliminary, as a larger case base would be needed to cover the range of typical planning problems. We plan to add further cases and to study the use of more sophisticated similarity measures. For example, methods from image registration could be used to compare the projections, and the importance of the different features could be learned from the data.

7 Conclusion

Typical areas for applications of case-based reasoning in the medical domain are decision support and clinical diagnosis [18,19,20]. In radiation therapy, case-based systems have been proposed to recommend dose prescriptions [21], and for treatment planning [22]. While the latter approach was developed in the context of much simpler beam geometries, it indicated that case-based treatment planning is possible. We have shown that this can be extended to the much more complex case of inverse planning for robotic radiosurgery. Although a larger scale evaluation considering more clinical cases and different clinical goals is necessary to confirm the robustness of the proposed method, our results are encouraging.

References

1. Schweikard, A., Bodduluri, M., Adler, J.R.: Planning for camera-guided robotic radiosurgery. IEEE transactions on robotics and automation 14(6), 951–962 (1998)
2. Stein, J., Mohan, R., Wang, X.-H., Bortfeld, T., Wu, Q., Preiser, K., Ling, C.C., Schlegel, W.: Number and orientations of beams in intensity-modulated radiation treatments. Med. Phys. 24(2), 149–160 (1997)
3. Schlaefer, A., Schweikard, A.: Stepwise multi-criteria optimization for robotic radiosurgery. Medical Physics 35(5), 2094–2103 (2008)

4. Schlaefer, A., Blanck, O., Schweikard, A.: Interactive multi-criteria inverse planning for robotic radiosurgery. In: Proceedings of the XVth International Conference on the Use of Computers in Radiation Therapy (ICCR) (2007)
5. Craft, D., Halabi, T., Bortfeld, T.: Exploration of tradeoffs in intensity-modulated radiotherapy. Phys. Med. Biol. 50, 5857–5868 (2005)
6. Rosen, I., Liu, H.H., Childress, N., Liao, Z.: Interactively exploring optimized treatment plans. Int. J. Radiation Oncology Biol. Phys. 61(2) (2005)
7. Wang, X., Zhang, X., Dong, L., Liu, H., Wu, Q., Mohan, R.: Development of methods for beam angle optimization for IMRT using an accelerated exhaustive search strategy. Int. J. Radiation Oncology Biol. Phys. 60(4), 1325–1337 (2004)
8. Li, Y., Yao, D., Yao, J., Chen, W.: A particle swarm optimization algorithm for beam angle selection in intensity-modulated radiotherapy planning. Phys. Med. Biol. 50, 3491–3514 (2005)
9. Schreibmann, E., Xing, L.: Feasibility study of beam orientation class-solutions for prostate IMRT. Med. Phys. 31(10), 2863–2870 (2004)
10. Mott, J.H., Livsey, J.E., Logue, J.P.: Development of a simultaneous boost IMRT class solution for a hypofractionated prostate cancer protocol. Br. J. Radiol. 77, 377–386 (2004)
11. Arránsa, R., Gallardob, M.I., Rosellóc, J., Sánchez-Doblado, F.: Computer optimization of class solutions designed on a beam segmentation basis. Radiother Oncol. 69(3), 315–321 (2003)
12. Khoo, V.S., Bedford, J.L., Webb, S., Dearnaley, D.P.: Class solutions for conformal external beam prostate radiotherapy. Int. J. Radiation Oncology Biol. Phys. 55(4), 1109–1120 (2003)
13. Wells, D.M., Niederer, J.: A medical expert system approach using artificial neural networks for standardized treatment planning. Int. J. Radiation Oncology Biol. Phys. 41(1), 173–182 (1998)
14. Schweikard, A., Schlaefer, A., Adler, J.R.: Resampling: An optimization method for inverse planning in robotic radiosurgery. Med. Phys. 33(11), 4005–4011 (2006)
15. Burkhard, H.D.: Similarity and distance in case based reasoning. Fundam. Inf. 47(3-4), 201–215 (2001)
16. Goitein, M., Abrams, M., Rowell, D., Pollari, H., Wiles, J.: Multi-dimensional treatment planning: II. beam's eye-view, back projection, and projection through CT sections. Int. J. Radiation Oncology Biol. Phys. 9(6), 789–797 (1983)
17. Kalet, I.J., Austin-Seymour, M.M.: The use of medical images in planning and delivery of radiation therapy. J. Am. Med. Inform. Assoc. 4(5), 327–339 (1997)
18. Holt, A., Bichindaritz, I., Schmidt, R., Perner, P.: Medical applications in case-based reasoning. Knowl. Eng. Rev. 20(3), 289–292 (2005)
19. Pantazi, S.V., Arocha, J.F., Moehr, J.R.: Case-based medical informatics. BMC Med. Inform. Decis. Mak. 4, 19 (2004)
20. Fritsche, L., Schlaefer, A., Budde, K., Schroeter, K., Neumayer, H.H.: Recognition of critical situations from time series of laboratory results by case-based reasoning. J. Am. Med. Inform. Assoc. 9(5), 520–528 (2002)
21. Song, X., Petrovic, S., Sundar, S.: A case-based reasoning approach to dose planning in radiotherapy. In: Wilson, D., Khemani, D. (eds.) Workshop Proceedings, The Seventh International Conference on Case-Based Reasoning (ICCBR 2007), pp. 348–357 (2007)
22. Berger, J.: Roentgen: radiation therapy and case-based reasoning. In: Proceedings of the Tenth Conference on Artificial Intelligence for Applications, pp. 171–177 (1994)

Conversational Case-Based Reasoning in Medical Classification and Diagnosis

David McSherry

School of Computing and Information Engineering, University of Ulster
Coleraine BT52 1SA, Northern Ireland, United Kingdom
dmg.mcsherry@ulster.ac.uk

Abstract. In case-based reasoning (CBR) approaches to classification and diagnosis, a description of the problem to be solved is often assumed to be available in advance. Conversational CBR (CCBR) is a more interactive approach in which the system is expected to play an active role in the selection of relevant tests to help minimize the number of problem features that the user needs to provide. We present a new algorithm for CCBR called iNN(k) and demonstrate its ability to achieve high levels of accuracy on a selection of datasets related to medicine and health care, while often requiring the user to provide only a small subset of the problem features required by a standard k-NN classifier. Another important benefit of iNN(k) is a goal-driven approach to feature selection that enables a CCBR system to explain the relevance of any question it asks the user in terms of its current goal.

Keywords: case-based reasoning, classification, diagnosis, accuracy, feature selection, transparency, explanation.

1 Introduction

Holt *et al.* [1] predict continued growth in medical applications of case-based reasoning (CBR) as the health sector becomes more accepting of decision support systems in clinical practice. Factors that may influence a CBR system's acceptability to users include (1) its ability to explain the reasoning process, and (2) its ability to solve problems for which the user is unable to provide a complete description. Explanation is a topic that continues to attract significant research interest in CBR [2], with recent contributions tending to focus on a CBR system's ability to explain or justify its conclusions [3-5]. In CBR systems that play an active role in guiding the selection of tests on which their conclusions are based, it is also reasonable for users to expect the system to explain the relevance of test results they are asked to provide [6].

In contrast to traditional CBR approaches to classification and diagnosis, conversational CBR (CCBR) makes no assumption that a description of the problem to be solved is available in advance. Instead, a problem description (or query) is incrementally elicited in an interactive dialogue with the aim of minimizing the number of questions the user is asked before a solution is reached [6-12]. As shown in applications such as interactive fault diagnosis and helpdesk support, guiding the

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 116–125, 2009.

selection of relevant tests is an important benefit of CCBR in situations where the user is unable to provide a complete problem description.

However, test (or feature) selection in CCBR is often based on strategies in which the absence of a specific goal or hypothesis makes it difficult to explain the relevance of questions the user is asked [12]. For the most part, CCBR research has also tended to focus on application domains in which the case base is typically *heterogeneous* (i.e., different attributes are used to describe different cases) and/or *irreducible* (i.e., each case has a unique solution) [7-8]. These are both features that are seldom found in the classification datasets that are common in medical applications of CBR. A related issue is that measures such as precision and recall are often used in the evaluation of CCBR systems rather than classification accuracy, which cannot be assessed by traditional methods for an irreducible dataset [11].

In this paper, we present a new algorithm for CCBR in classification and diagnosis called iNN(k) and evaluate its performance in terms of the accuracy and efficiency of problem-solving dialogues. Feature selection in iNN(k) is guided by the goal of confirming a *target* class and informed by a heuristic measure of a feature's discriminating power in favor of the target class. As well as helping to minimize the average length of CCBR dialogues in the approach, this has the important advantage of enabling a CCBR system to explain the relevance of a selected feature in terms of its current goal.

In Sections 2 and 3, we describe our approach to CCBR in iNN(k) and demonstrate the approach in a CCBR system called CBR-*Confirm*. In Section 4, we present empirical results that demonstrate the ability of iNN(k) to achieve high levels of accuracy on four datasets related to medicine and health care, while often requiring the user to provide only a small subset of the problem features required by a standard k-NN classifier. Our conclusions are presented in Section 5.

2 CCBR in iNN(k)

In this section, we describe the basic concepts in our approach to CCBR, including the similarity measure used to construct the iNN(k) retrieval set, the measure of discriminating power used to guide the selection of relevant features, and the criteria used to decide when to terminate a CCBR dialogue. The examples that we use to illustrate the approach are based on the contact lenses dataset, a simplified version of the real-world problem of selecting a suitable type of contact lenses for an adult spectacle wearer [13-14]. The classes to be distinguished in the dataset are no contact lenses (63%), soft contact lenses (21%), and hard contact lenses (17%). Attributes in the dataset are age, spectacle prescription, astigmatism, and tear production rate.

Case Structure. We assume the dataset (or case base) to be such that the same attributes are used to describe each case (though there may be missing values in the dataset), and each class is typically represented by several cases. Currently, we also assume that only nominal and/or discrete attributes are used to describe cases, or that prior discretization of continuous attributes has been undertaken if necessary. A case C consists of a case identifier, a problem description, and a solution. The problem description is a list of features $a = v$ of length $|A|$, where A is the set of attributes in the case base, and $v \in domain(a) \cup \{\text{unknown}\}$ for each $a \in A$, where $domain(a)$ is the

set of all known values of a in the case base. For each $a \in A$, we denote by $\pi_a(C)$ the value of a in C. The solution stored in C, which we denote by $class(C)$, is a diagnosis or other class label. We say that a case C supports a given class G if $class(C) = G$.

Query Elicitation and Representation. In iNN(k), an initially empty query is extended in a CCBR dialogue by asking the user for the values of attributes that are most useful according to the criteria described in this section. The user can answer *unknown* to any question, in which case the user's answer is recorded and the dialogue moves on to the next most useful question. A non-empty query is represented as a list of problem features $Q = \{a_1 = v_1, ..., a_n = v_n\}$, where $n \leq |A|$ and $v_i \in domain(a_i) \cup \{\text{unknown}\}$ for $1 \leq i \leq n$. We denote by A_Q the set of attributes in the current query Q. For each $a \in A_Q$, we denote by $\pi_a(Q)$ the value of a in Q.

Similarity Measure. In contrast to algorithms that rely on feature weights to address the dimensionality problem in k-NN, all attributes are equally weighted in our approach to CCBR. For any case C and non-empty query Q, we define:

$$Sim(C, Q) = \frac{\sum_{a \in A_Q} sim_a(C, Q)}{|A|} \quad (1)$$

where for each $a \in A_Q$, $sim_a(C, Q) = 1$ if $\pi_a(Q) \neq$ unknown and $\pi_a(C) = \pi_a(Q)$, and $sim_a(C, Q) = 0$ if either of these conditions is not satisfied. For an empty query Q, we define $Sim(C, Q) = 0$ for every case C.

The iNN(k) Retrieval Set. For $k \geq 1$, we refer to the set of cases retrieved by iNN(k) in each cycle of a CCBR dialogue as the iNN(k) retrieval set. Its role in classification differs from that of the retrieval set constructed by a standard k-NN classifier. As well as providing a final solution in most CCBR dialogues, the iNN(k) retrieval set is also used to monitor the progress of a CCBR dialogue and decide when to terminate the dialogue. Also in contrast to the standard k-NN retrieval set, the iNN(k) retrieval set may contain more than k cases, though never less than k cases as in some CCBR approaches [9]. A strategy commonly used in k-NN to ensure the retrieval of exactly k cases is to break ties between equally similar cases by giving priority to those that are nearest the top of the list of cases in the case base. In iNN(k), we adopt the alternative strategy of retrieving any case for which the number of more similar cases is less than k. More formally, we define the iNN(k) retrieval set for a given query Q to be:

$$r(Q, \text{iNN}(k)) = \{C \mid more\text{-}similar(C, Q) < k\} \quad (2)$$

where $more\text{-}similar(C, Q)$ is the number of cases C^* such that $Sim(C^*, Q) > Sim(C, Q)$. For example, 12 cases in the contact lenses dataset are equally similar (0.25) to the query $Q = \{$tear production rate = normal$\}$, and their solutions include all classes in the dataset. In this situation, even the iNN(k) retrieval set for $k = 1$ will include the 12 cases that are equally good candidates for retrieval. It can also be seen that the iNN(k) retrieval set for the empty query at the start of a CCBR dialogue is the set of all cases in the case base.

Discriminating Power. There is no *a priori* feature selection in iNN(k), for example as in some approaches to the dimensionality problem in k-NN. Instead, the selection

of features that are most relevant in the solution of a given problem is based on a heuristic measure of a feature's *discriminating power* that we now define. For any class G, attribute a, and $v \in domain(a)$, the discriminating power of $a = v$ in favor of G is:

$$d(a = v, G) = \frac{p(a = v \mid G) - p(a = v \mid \neg G)}{\mid domain(a) \mid} . \quad (3)$$

For example, the feature age = presbyopic appears in 6 of the 15 cases in the contact lenses dataset that support no contact lenses and in 2 of the 9 cases that support soft or hard contact lenses. As age has three nominal values in the dataset,

$$d(\text{age = presbyopic, no contact lenses}) = \frac{1}{3} \times \left(\frac{6}{15} - \frac{2}{9} \right) = 0.06 . \quad (4)$$

However, the feature in the dataset with most discriminating power (0.40) in favor of no contact lenses is tear production rate = reduced. Note that any case with a missing value for a is ignored when estimating the probabilities in (3).

Selecting a Target Class. The target class used to guide feature selection in iNN(k) is the class G^* that is supported by most cases in the iNN(k) retrieval set for the current query. If two or more classes are supported by equal numbers of cases in the iNN(k) retrieval set, then the one that is supported by most cases in the case base as a whole is selected as the target class. As the user's query is empty at the start of a CCBR dialogue in iNN(k), the target class is initially the class that is supported by most cases in the case base. However, the target class may be revised at any time as the user's query is extended.

Local and Global Feature Selection. One important parameter in our approach to CCBR is the integer $k \geq 1$ used to construct the iNN(k) retrieval set. Another is whether feature selection is based on *local* or *global* assessment of a feature's discriminating power. In *local* feature selection, only features that appear in one or more cases in the iNN(k) retrieval set that support the target class are candidates for selection. Also, a feature's (local) discriminating power (3) is based only on cases in the iNN(k) retrieval set. In *global* feature selection, features that appear in any case that supports the target class are candidates for selection, and their (global) discriminating power is based on all cases in the case base. Where necessary to distinguish between the local and global versions of iNN(k), we will refer to them as iNN(k)-L and iNN(k)-G respectively. As we show in Section 4, the optimal choice of parameters in the algorithm depends on the dataset, for example with iNN(1)-L giving the best performance on the contact lenses dataset in our experiments.

Selecting the Most Useful Question. The feature selected by iNN(k) as most useful for confirming a target class G^* in a CCBR dialogue depends on the current query Q and on whether local or global feature selection is used in the algorithm. In iNN(k)-L, the most useful feature $a^* = v^*$ is the one with most local discriminating power in favor of G^* over all features $a = v$ such that $a \in A - A_Q$, $v \in domain(a)$, and $\pi_a(C) = v$ for at least one $C \in r(Q, \text{iNN}(k))$ such that $class(C) = G^*$. In iNN(k)-G, the condition that $\pi_a(C) = v$ for at least one $C \in r(Q, \text{iNN}(k))$ such that $class(C) = G^*$ is replaced by

the weaker condition that $\pi_a(C) = v$ for at least one case C such that $class(C) = G^*$. Also, discriminating power is globally assessed rather than locally as in iNN(k)-L.

Deciding When to Terminate a CCBR Dialogue. A CCBR dialogue in iNN(k) typically continues until all cases in the iNN(k) retrieval set have the same class label G. At this point, the dialogue is terminated and G is presented as the solution to the user's problem. Less commonly, a point in the dialogue may be reached where all possible questions have been asked but there are still cases with different class labels in the iNN(k) retrieval set. In this situation, the class G that is supported by most cases in the iNN(1) retrieval set (whether $k > 1$ or not) is selected as the solution to the user's problem. If two or more classes are supported by equal numbers of cases in the iNN(1) retrieval set, then the one that is supported by most cases in the case base as a whole is selected as the solution. It is also possible for a point in the dialogue to be reached where a most useful question cannot be identified because all cases in the iNN(k) retrieval set (or case base as a whole) that support the target class have missing values for all remaining attributes. Again, the solution in this situation is the class that is supported by most cases in the iNN(1) retrieval set.

iNN(k)-L. Having described the main features of our approach to CCBR in iNN(k), we end this section with a brief summary of iNN(k)-L, the version of iNN(k) that we use to demonstrate the approach in Section 3. As shown in Fig. 1, conditions for termination of the CCBR dialogue are tested in Steps 1, 2, and 4 of the algorithm. The class G^* supported by most cases in the iNN(k) retrieval set is selected as the target class in Step 3. The feature $a^* = v^*$ with most (local) discriminatory power in favor of G^* is selected in Step 5. The user is asked for the value of a^* in Step 6. Finally, the user's answer is used to extend the current query in Step 7 and the cycle is repeated.

3 CCBR in CBR-Confirm

In a CCBR dialogue based on iNN(k), an initially empty query is extended by asking the user questions with the goal of confirming a target class. In this section, we demonstrate the approach in a CCBR system called CBR-*Confirm*. A brief discussion of the system's explanation capabilities is followed by an example dialogue based on the contact lenses dataset [13-14] in which iNN(k) is used with $k = 1$ and local feature selection.

3.1 Explaining the Relevance of a Selected Feature

Before answering any question in CBR-*Confirm*, the user can ask why it is relevant. Rather than explaining the relevance of a selected feature $a^* = v^*$ in terms of its discriminating power, CBR-*Confirm* looks one step ahead to determine its effects on the class distribution in the iNN(k) retrieval set. For example, if Q is the current query, G^* is the target class, and all cases in the iNN(k) retrieval set for $Q \cup \{a^* = v^*\}$ support G^*, then the effect of $a^* = v^*$ will be to confirm the target class. Alternatively, the selected feature may have the effect of eliminating all cases that support a competing class G from the iNN(k) retrieval set. In general, this does not mean that cases that

support G will never be readmitted to the iNN(k) retrieval set as the user's query is further extended. Nonetheless, a reasonable explanation of the selected feature's relevance is that it provides evidence against the competing class. If a selected feature has neither of these effects, CBR-*Confirm*'s explanation of its relevance is simply that it may help to confirm the target class.

Algorithm: iNN(k)-L

Input: An integer $k \geq 1$, a case base with attributes A, and an initially empty query Q

Output: A solution class G

Process: Repeat Steps 1-7 until one of the stopping criteria is satisfied:

1. If all cases in $r(Q, \text{iNN}(k))$ have the same class label G, then return G
2. If $A_Q = A$ then return the class G supported by most cases in $r(Q, \text{iNN}(1))$
3. Select the class G^* that is supported by most cases in $r(Q, \text{iNN}(k))$ as the target class
4. If all cases in $r(Q, \text{iNN}(k))$ that support G^* have missing values for all $a \in A - A_Q$, then return the class G supported by most cases in $r(Q, \text{iNN}(1))$
5. Select the feature $a^* = v^*$ with most local discriminating power $d(a^* = v^*, G^*)$ in favor of G^* over all features $a = v$ such that $a \in A - A_Q$, $v \in domain(a)$, and $\pi_a(C) = v$ for at least one $C \in r(Q, \text{iNN}(k))$ such that $class(C) = G^*$
6. Ask the user for the value of a^*
7. If the value of a^* is unknown to the user then $Q \leftarrow Q \cup \{a^* = \text{unknown}\}$ else $Q \leftarrow Q \cup \{a^* = v\}$, where v is the value of a^* reported by the user

---

Fig. 1. iNN(k) with local feature selection

3.2 Explaining a Conclusion

At the end of a CCBR dialogue, CBR-*Confirm* presents the class G it has selected as a solution to the user's problem and explains its conclusion by showing the user the most similar case that supports G, or the one that appears first in the case base if there are two or more such cases. Also with the aim of increasing transparency, features that match the user's query are highlighted in the solution case.

3.3 Example CCBR Dialogue

Fig. 2 shows a CCBR dialogue in CBR-*Confirm* based on the contact lenses dataset [13-14]. As the problem of contact lenses selection is highly simplified in the dataset, the example dialogue should not be regarded as a realistic example of decision making in the domain. At the start of the example dialogue, the majority class in the dataset (no contact lenses) is selected by CBR-*Confirm* as the target class. The user is now asked for the tear production rate because tear production rate = reduced is the feature with most discriminating power in favor of the target class. In light of the user's answer (normal), the target class changes to soft contact lenses, and astigmatism = no is identified as the feature with most (local) discriminating power in favor of the new target class.

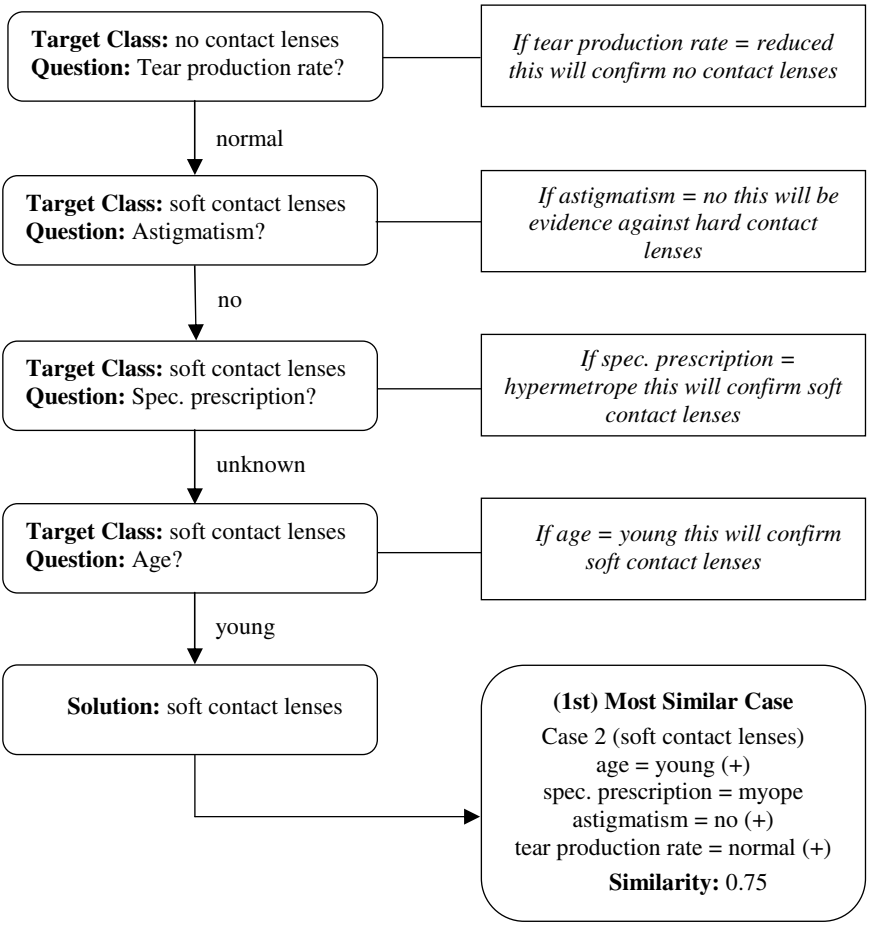


Fig. 2. Example CCBR dialogue in CBR-*Confirm* based on iNN(1)-L

The target class (soft contact lenses) remains unchanged as the dialogue continues and is finally confirmed even though the spectacle prescription is unknown to the user in the 3rd cycle. Matching features in the solution case that the user is shown (Case 2) are indicated by a plus sign (+). The example dialogue also shows the explanation of question relevance provided by CBR-*Confirm* in each cycle if requested by the user.

4 Empirical Study

In this section, we investigate the hypothesis that the performance of iNN(k) on a given dataset depends on the value of k and also on the choice between local and global feature selection in the algorithm. The performance measures of interest are classification accuracy and dialogue efficiency (i.e., average number of questions required to reach a solution). We also assess the accuracy of iNN(k) in comparison

with a standard *k*-NN classifier. The datasets used in our experiments (all of which are available from the UCI Machine Learning Repository [14]) are:

- Contact Lenses [13] (24 cases, 4 attributes, 3 classes)
- Breast Cancer (286 cases, 9 attributes, 2 classes)
- Lymphography (148 cases, 18 attributes, 4 classes)
- SPECT Heart [15] (267 cases, 22 attributes, 2 classes)

All attributes in the selected datasets are nominal or discrete, and there are missing values only in the Breast Cancer dataset. We use a leave-one-out cross validation approach to evaluate each algorithm on the selected datasets. Each case in turn is temporarily removed from the dataset to provide the description of a problem to be solved by (1) a standard *k*-NN classifier, and (2) a CCBR system based on iNN(*k*). Features in the left-out case are revealed by a simulated user in response to questions selected by the CCBR system. For any attribute with a missing value in the left-out case, the simulated user answers *unknown* when asked for its value. At the end of each dialogue, we record the number of questions required to reach a solution and whether or not the solution is correct (i.e., the same as in the left-out case). We also present the problem description from the left-out case to the *k*-NN classifier, and record whether or not the solution it provides is correct.

For each dataset, Table 1 shows the accuracy achieved by *k*-NN for $k = 1$, 3, and 5 and by iNN(*k*)-L/G for $k = 1$, 2, and 3. The best accuracy results for each dataset are shown in bold. Average lengths of iNN(*k*) dialogues (and the number of attributes in each dataset) are shown in brackets. Maximum levels of accuracy achieved by iNN(*k*) in our experiments exceeded those achieved by *k*-NN for all datasets. For example, accuracy on Breast Cancer was highest (75.2%) in iNN(2)-G and iNN(3)-G. iNN(2)-G also gave the highest levels of accuracy on Lymphography (86.5%) and SPECT Heart (84.3%), while accuracy on Contact Lenses was highest (83.3%) in iNN(1)-L and iNN(2)-L.

Average dialogue length in iNN(*k*)-L and iNN(*k*)-G can be seen to increase or remain unchanged for all four datasets as *k* increases from 1 to 3. A tendency for average dialogue length to increase can also be seen as we move from iNN(*k*)-L to iNN(*k*)-G for $k = 1$, 2, and 3.

Table 1. Accuracy of *k*-NN and iNN(*k*)-L/G on the selected datasets. The best accuracy results for each dataset are shown in bold. Average lengths of iNN(*k*) dialogues are shown in brackets.

Dataset	*k*-NN			iNN(*k*)-L			iNN(*k*)-G		
	$k = 1$	$k = 3$	$k = 5$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
Contact Lenses (4)	75.0	62.5	70.8	**83.3** (2.1)	**83.3** (2.1)	70.8 (2.4)	70.8 (2.1)	70.8 (2.3)	70.8 (2.4)
Breast Cancer (9)	72.7	73.4	73.4	70.3 (6.3)	73.1 (6.9)	74.5 (7.5)	71.7 (6.8)	**75.2** (7.6)	**75.2** (8.2)
Lymphography (18)	78.4	79.7	81.8	74.3 (5.0)	79.1 (6.3)	83.1 (7.3)	79.1 (5.5)	**86.5** (7.5)	85.8 (8.6)
SPECT Heart (22)	73.0	74.9	78.7	77.5 (8.1)	82.4 (8.8)	82.0 (9.8)	79.0 (9.7)	**84.3** (11.2)	83.5 (12.3)

The results support our hypothesis that the performance of iNN(k) on a given dataset depends on the value of k and on the choice between local and global feature selection in the algorithm. A trade-off between accuracy and dialogue efficiency can be seen, for example, in the iNN(k) results for Lymphography. For this dataset, an increase in accuracy from 74.3% to 86.5% is gained at the expense of an increase in average dialogue length from 5.0 to 7.5. Even so, the average number of features (7.5) required for 86.5% accuracy in iNN(2)-G is much less than the number of features (18) required for 81.8% accuracy in 5-NN. However, accuracy does not always increase as k increases in iNN(k). For example, it can be seen to decrease or remain unchanged in iNN(k)-G for all datasets as k increases from 2 to 3. Average dialogue length required for maximum accuracy in iNN(k) ranges from 42% to 84% of features in the four datasets, with an overall average of 58%.

5 Conclusions

In this paper, we presented a new algorithm for CCBR in classification and diagnosis called iNN(k) and demonstrated its ability to achieve high levels of accuracy on four datasets related to medicine and health care, while often requiring the user to provide only a small subset of the features in a complete problem description. For example, the best accuracy results for Lymphography (86.5%) and SPECT Heart (84.3%) in iNN(k) are based on only 42% and 51% on average of the features required by a standard k-NN classifier. Feature selection is guided in iNN(k) by the goal of confirming a target class and informed by a heuristic measure of discriminating power. As demonstrated in CBR-*Confirm*, this has the important advantage of enabling a CCBR system to explain the relevance of any question it asks the user. While we have focused on CCBR in this paper, there is nothing to prevent a non-interactive version of iNN(k) being used in situations where a problem description is provided in advance as in traditional CBR approaches to classification and diagnosis. Such a non-interactive version of iNN(k) would use the same criteria to select the most relevant features from a given problem description as used by CBR-*Confirm* when interacting with a human user. Eliminating the need for prior discretization of continuous attributes in iNN(k) is another important direction for future research.

Acknowledgments. The author is grateful to the reviewers for their helpful comments. Thanks also to Matjaz Zwitter and Milan Soklic for providing the Breast Cancer and Lymphography datasets in the UCI Machine Learning Repository.

References

1. Holt, A., Bichindaritz, I., Schmidt, R., Perner, P.: Medical Applications in Case-Based Reasoning. Knowledge Engineering Review 20, 289–292 (2005)
2. Leake, D., McSherry, D.: Introduction to the Special Issue on Explanation in Case-Based Reasoning. Artificial Intelligence Review 24, 103–108 (2005)
3. Doyle, D., Cunningham, P., Bridge, D., Rahman, Y.: Explanation Oriented Retrieval. In: Funk, P., González Calero, P.A. (eds.) ECCBR 2004. LNCS (LNAI), vol. 3155, pp. 157–168. Springer, Heidelberg (2004)

4. Evans-Romaine, K., Marling, C.: Prescribing Exercise Regimens for Cardiac and Pulmonary Disease Patients with CBR. In: ICCBR 2003 Workshop on Case-Based Reasoning in the Health Sciences (2003)
5. McSherry, D.: Explaining the Pros. and Cons. of Conclusions in CBR. In: Funk, P., González Calero, P.A. (eds.) ECCBR 2004. LNCS (LNAI), vol. 3155, pp. 317–330. Springer, Heidelberg (2004)
6. McSherry, D.: Interactive Case-Based Reasoning in Sequential Diagnosis. Applied Intelligence 14, 65–76 (2001)
7. Aha, D.W., Breslow, L.A., Muñoz-Avila, H.: Conversational Case-Based Reasoning. Applied Intelligence 14, 9–32 (2001)
8. Aha, D.W., McSherry, D., Yang, Q.: Advances in Conversational Case-Based Reasoning. Knowledge Engineering Review 20, 247–254 (2005)
9. Bogaerts, S., Leake, D.: What Evaluation Criteria are Right for CCBR? Considering Rank Quality. In: Roth-Berghofer, T.R., Göker, M.H., Güvenir, H.A. (eds.) ECCBR 2006. LNCS (LNAI), vol. 4106, pp. 385–399. Springer, Heidelberg (2006)
10. Gu, M., Aamodt, A.: Evaluating CBR Systems Using Different Data Sources: a Case Study. In: Roth-Berghofer, T.R., Göker, M.H., Güvenir, H.A. (eds.) ECCBR 2006. LNCS (LNAI), vol. 4106, pp. 121–135. Springer, Heidelberg (2006)
11. McSherry, D.: Minimizing Dialog Length in Interactive Case-Based Reasoning. In: 17th International Joint Conference on Artificial Intelligence, pp. 993–998. Morgan Kaufmann, San Francisco (2001)
12. McSherry, D., Hassan, S., Bustard, D.: Conversational Case-Based Reasoning in Self-Healing and Recovery. In: Althoff, K.-D., Bergmann, R., Minor, M., Hanft, A. (eds.) ECCBR 2008. LNCS (LNAI), vol. 5239, pp. 340–354. Springer, Heidelberg (2008)
13. Cendrowska, J.: PRISM: an Algorithm for Inducing Modular Rules. International Journal of Man-Machine Studies 27, 349–370 (1987)
14. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences (2007)
15. Kurgan, L.A., Cios, K.J., Tadeusiewicz, R., Ogiela, M., Goodenday, L.S.: Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis. Artificial Intelligence in Medicine 23, 149–169 (2001)

Histopathology Image Classification Using Bag of Features and Kernel Functions

Juan C. Caicedo, Angel Cruz, and Fabio A. Gonzalez

Bioingenium Research Group
National University of Colombia
{jccaicedoru,aacruzr,fagonzalezo}@unal.edu.co
http://www.bioingenium.unal.edu.co

Abstract. Image representation is an important issue for medical image analysis, classification and retrieval. Recently, the bag of features approach has been proposed to classify natural scenes, using an analogy in which visual features are to images as words are to text documents. This process involves feature detection and description, construction of a visual vocabulary and image representation building through visual-word occurrence analysis. This paper presents an evaluation of different representations obtained from the *bag of features* approach to classify histopathology images. The obtained image descriptors are processed using appropriate kernel functions for Support Vector Machines classifiers. This evaluation includes extensive experimentation of different strategies, and analyses the impact of each configuration in the classification result.

1 Introduction

Medical imaging applications are challenging because they require effective and efficient content representations to manage large image collections. The first stage for medical image analysis is modeling image contents by defining an appropriate representation. This is a fundamental problem for all image analysis tasks such as image classification, automatic image annotation, object recognition and image retrieval, which require discriminative representations according to the application domain. During the last few years, the *bag of features* approach has attracted great attention from the computer vision community. This approach is an evolution of texton-based representations and is also influenced by the bag of words assumption in text classification. In text documents, a word dictionary is defined and all documents are processed so that the frequency of each word is quantified. This representation ignores word relationships in the document, i.e., it does not take into account the document structure. An analogy is defined for images in which a feature dictionary is built to identify visual patterns in the collection. This representation has shown to be effective in different image classification, categorization and retrieval tasks [1,2,3].

The *bag of features* representation is an adaptive approach to model image structure in a robust way. In contrast to image segmentation, this approach does

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 126–135, 2009.

not attempt to identify complete objects inside images, which may be a harder task than the image classification itself. Instead, the *bag of features* approach looks for small characteristic image regions allowing the representation of complex image contents without explicitly modeling objects and their relationships, a task that is tackled in another stage of image content analysis. In addition, an important advantage of the *bag of features* approach is its adaptiveness to the particular image collection to be processed. In the same way as text documents, in which the appropriate word-list to be included in the dictionary may be identified earlier in the process, the *bag of features* approach allows to identify visual patterns that are relevant to the whole image collection. That is, the patterns that are used in a single image representation come from the analysis of patterns in the complete collection. Other important characteristics of this approach are the robustness to occlusion and affine transformations as well as its computational efficiency [2].

Some of these properties are particularly useful for medical image analysis and, in fact, the *bag of features* representation has been successfully applied to some problems in medical imaging [4,5]. Histopathology images have a particular structure, with a few colors, particular edges and a wide variety of textures. Also, objects such as glands or tissues may appear anywhere in the image, in different proportions and at different zoom levels. All those properties make the *bag of features* a potentially appropriate representation for that kind of visual contents. Up to our knowledge, the *bag of features* representation has not been evaluated yet on histopathology images and that is the main goal of this paper.

This paper presents a systematic evaluation of different representations obtained from the *bag of features* approach to classify histopathology images. There are different possibilities to design an image descriptor using the *bag of features* framework and each one lead to different image representations that may be more or less discriminative. In addition, the obtained image descriptors are processed using two kernel functions for Support Vector Machine classifiers. The performed experiments allow to analyze the impact of different strategies in the final classification result. The paper is organized as follows: Section 2 presents the previous work on histopathology image classification. Section 3 describes the *bag of features* methodology and all applied techniques. Section 4 presents experimental results, and finally the concluding remarks are in Section 5.

2 Histopathology Image Classification

2.1 Previous Work

In different application contexts, medical images have been represented using a wide variety of descriptors including colors, textures, regions and transformation-based coefficients. Such descriptors are usually motivated by visual properties that can be identified by domain experts in target images. For instance, Long et. al [6] developed an algorithm based on active contours to segment vertebrae in x-ray spine images, and then match nine semantic points to evaluate a particular disease. Although the algorithm is very effective in such a task, it has two main

disadvantages: first, the computational effort required to process each image, and second, the method is devised to work only on that particular kind of medical images. Other more generic descriptors have been proposed for classification and retrieval of medical images. Guld et. al [7] proposes to down-scale medical images up to 32x32 pixels and use that as a feature vector for classification tasks. In [8] the use of global histogram features are used to retrieve a wide variety of medical images. Even though that descriptors are by nature simple and generic, they lack of direct semantics and may lead to poor results in large-scale real applications. Hence, there is a trade-off between the semantics and the generality of the representation.

In histopathology images that tendency can also be observed. Most of the previous works in the context of histology, pathology and tissue image classification have approached the problem using segmentation techniques [9,10]. They first define the target object to be segmented (e.g. cells, nuclei, tissues) and then apply a computational strategy to identify it. Global features have also been used to retrieve and classify histology images [11,12]. Those two global approaches are in one extreme of the balance between explicit semantics and practical generalization or adaptation. Other kind of works have oriented the image content analysis by window-based features, under the observation that histology images are "*usually composed of different kinds of texture components*" [13]. In [14] those sub-images are classified individually and then a semantic analyzer is used to decide the final image classification on the complete image. This approach is close to the *bag of features* one since the unit of analysis is a small sub-image and a learning algorithm is applied to evaluate the categorization of small sub-images. However, the approach requires the annotation of example sub-images to train first-stage classifiers, a process that is performed in an unsupervised fashion under the bag of features framework.

2.2 Histopathology Image Dataset

The image dataset has been previously used in an unrelated clinical study to diagnose a special skin cancer known as basal-cell carcinoma. Basal-cell carcinoma is the most common skin disease in white populations and its incidence is growing world wide [15]. It has different risk factors and its development is mainly due to ultraviolet radiation exposure. Pathologists confirm whether or not this disease is present after a biopsied tissue is evaluated under microscope. The database is composed of 1,502 images annotated by experts into 18 categories. Each label corresponds to a histopathology concept which may be found in a basal-cell carcinoma image. An image may have one or several labels, that is to say, different concepts may be recognized within the same image and the other way around.

3 The Bag of Features Framework

The *bag of features* framework is an adaptation of the *bag of words* scheme used for text categorization and text retrieval. The key idea is the construction of a

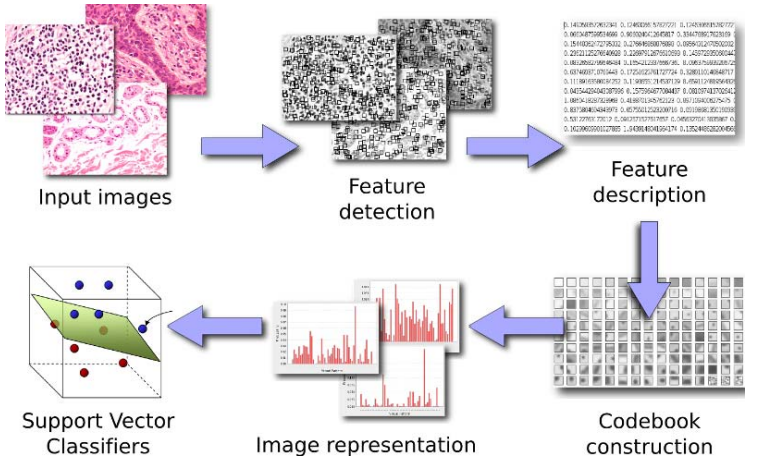


Fig. 1. Overview of the Bag of Features framework

codebook, that is, a visual vocabulary, in which the most representative patterns are codified as *codewords* or visual words. Then, the image representation is generated through a simple frequency analysis of each *codeword* inside the image. Csurka et. al [2] describe four steps to classify images using a *bag of features* representation: (1) Feature detection and description, (2) Codebook generation, (3) the *bag of features* construction and finally (4) training of learning algorithms. Figure 1 shows an overview of those steps. The *bag of features* approach is a flexible and adaptable framework, since each step may be determined by different techniques according to the application domain needs. The following subsections present the particular methods and techniques that have been evaluated in this work.

3.1 Feature Detection and Description

Feature detection is the process in which the relevant components of an image must be identified. Usually, the goal of feature detection is set to identify a spatially limited image region that is salient or prominent. Different strategies have been proposed by the computer vision community to detect local features, that are motivated by different visual properties such as corners, edges or saliency. Once local features are detected, the next step is to describe or characterize the content of such local regions. Ideally, two local features should have the same descriptor values if they refer to the same visual concept. That motivates the implementation of descriptors that are invariant to affine transformations and illumination changes.

In this work two feature detection strategies with their corresponding feature descriptor have been evaluated. The first strategy is dense random sampling.

The goal of this strategy is to select points in the image plane randomly and then, define a block of pixels around that coordinate. The size of the block is set to 9×9 pixels, and the descriptor for these blocks is the vector with explicit pixel values in gray scales. This descriptor will be called *raw block*, but it is also known as texton or raw pixel descriptor. The advantage of this strategy is its simplicity and computational efficiency. In addition, a large number of blocks may be extracted from different image scales, and that sample is a good approximation of the probability distribution of visual patterns in the image [16].

The second strategy is based on Scale-Invariant Feature Transform (SIFT) points [17]. This strategy uses a keypoint detector based on the identification of interesting points in the location-scale space. This is implemented efficiently by processing a series of difference-of-Gaussian images. The final stage of this algorithm calculates a rotation invariant descriptor using predefined orientations over a set of blocks. We use SIFT points with the most common parameter configuration: 8 orientations and 4×4 blocks, resulting in a descriptor of 128 dimensions. The SIFT algorithm has demonstrated to be a robust keypoint descriptor in different image retrieval and matching applications, since it is invariant to common image transformations, illumination changes and noise.

3.2 Codebook Construction

The visual dictionary or *codebook* is built using a clustering or vector quantization algorithm. In the previous stage of the *bag of features* framework, a set of local features has been extracted. All local features, over a training image set, are brougth together independently of the source image and are clustered to learn a set of representative visual words from the whole collection. The k-means algorithm is used in this work to find a set of centroids in the local features dataset. An important decision in the construction of the *codebook* is the selection of its size, that is, how many *codeblocks* are needed to represent image contents. According to different works on natural image classification, the larger the *codebook* size the better [16,2]. However, Tomassi et. al [4] found that the size of the codebook is not a significant aspect in a medical image classification task. We evaluated different *codebook* sizes, to analyze the impact of this parameter in the classification of histopathology images.

The goal of the *codebook* construction is to identify a set of visual patterns that reflects the image collection contents. Li et. al [18] have illustrated a *codebook* for natural scene image categorization that contains several visual primitives, such as orientations and edges. The result is consistent with the contents of that image collection. In the same way, we illustrate a *codebook* extracted from the collection of histopathology images. It is composed of 150 *codeblocks* as is shown in Figure 2. In this case the *codeblocks* are also reflecting the contents of the histopathology image collection. Visual primitives in this case may be representing cells and nuclei of different sizes and shapes. That codebook has been generated using the raw-block descriptor described in the previous Subsection.

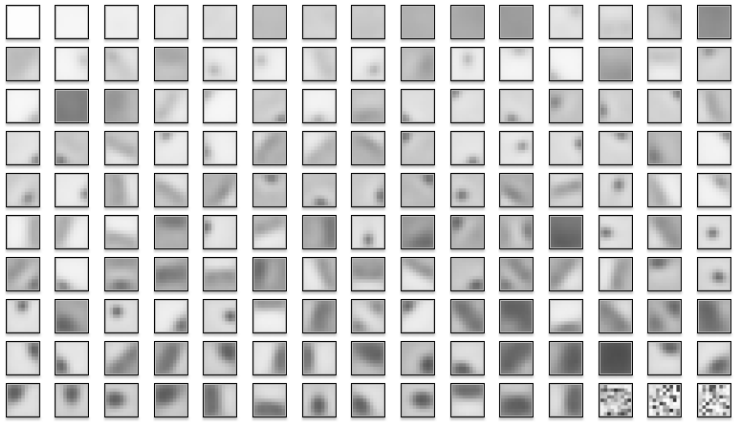


Fig. 2. A *codebook* with 150 *codeblocks* for the histopathology image collection. Codeblocks are sorted by frequency.

3.3 Image Representation

The final image representation is calculated by counting the occurrence of each *codeblock* in the set of local features of an image. This is why the framework receives its name, the *bag of features*, since the geometric relationships of local features are not taken into account. This representation is known as term frequencies (TF) in text applications and is also adopted in this work. Normalized term frequencies can also be calculated by normalizing according to the number of features extracted from the image. This may be specially useful for the SIFT points strategy, in which the number of features from image to image may have a large variation. Those are the standard image representations, commonly used for image categorization. In addition, the *inverse document frequency* (IDF) has also been used as weighting scheme to produce a new image representation.

3.4 Kernel Functions

Classifiers used in this work are Support Vector Machines (SVM), that receives as input a data representation impliciltly defined by a kernel function [19]. Kernel functions describe a similarity relationship between the objects to be classified. The image representation that we are dealing with are histograms with term frequencies. In that sense, a natural choice of a kernel function would be a similarity measure between histogram structures. The histogram intersection kernel is the first kernel function evaluated in this work:

$$D_{\cap}(H, H') = \sum_{m=1}^{M} min(H_m, H'_m)$$

Where H and H' are the term frequency histograms of two images, calculated using a *codebook* with M *codeblocks*. In addition, a Radial Basis Function composed with the histogram intersection kernel has been also implemented and tested:

$$K(H, H') = \exp(-\gamma D_{\cap}(H, H'))$$

4 Experimental Evaluation

4.1 Experimental Setup

The collection has 1,502 histopathology images with examples of 18 different concepts. The collection is split into 2 datasets, the first one for training and validation, and the second one for testing. The dataset partition is done using stratified sampling in order to preserve the original distribution of examples in both datasets. This is particularly important due to the high imbalance of classes. In the same way, the performance measures reported in this work are precision, recall and F-measure to evaluate the detection rate of positive examples, since the class imbalance may produce trivial classifiers with high accuracy that do not recognize any positive sample. In addition, since one image can be classified in many classes simultaneously, the classification strategy is based on binary classifiers following the one-against-all rule. Experiments to evaluate different configurations of the bag of features approach have been done. For each experiment, the regularization parameter of the SVM is controlled using 10-fold cross validation in the training dataset, to guarantee good generalization on the test dataset. Reported results are calculated on the test dataset and averaged over all 18 classes.

4.2 Results

The first evaluation focuses on the *codebook* size. We have tested six different *codebook* sizes, starting with 50 *codeblocks* and following with 150, 250, 500, 750 and 1000. Figure 3 shows a plot for *codebook* size vs. F-measure using two different *bag of features* configurations. The first strategy, is based on SIFT points and the second is based on raw blocks. Perhaps surprisingly, the plot shows that the classification performance decreases while the *codebook* size increases. This behaviour is explained by the intrinsic structure of histopathology images: they are usually composed of some kinds of textures, that is, the number of distinctive patterns in the collection is limited. This fact can also be seen in the *codebook* illustrated in Figure 2, which shows several repeated patterns, even with just 150 *codeblocks*. In the limit, it is reasonable that a large codebook size does not have any discriminative power because each different pattern in an image has its own *codeblock*.

The nature of the descriptor is also a determinant factor in this behaviour since the performance of the SIFT points decreases faster than the performance of raw blocks. This suggests that a SIFT-based *codebook* requires less *codeblocks*

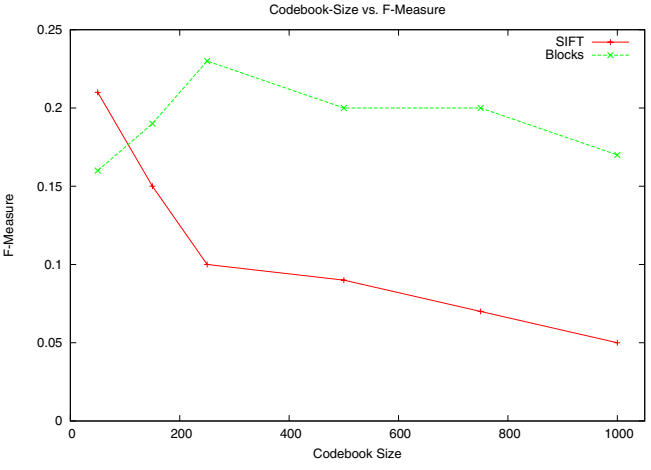


Fig. 3. Codebook size vs. F-Measure for two *bag of features* representation using SIFT points and Blocks

to express all different patterns in the image collection, which is consistent with the rotation and scale invariance properties of that descriptor. On the other hand, a block-based *codebook* requires a larger size because it is representing the same visual patterns using different *codeblocks*.

The second factor to evaluate is the feature descriptor. As is shown in Figure 3, the raw-block descriptor has obtained a better performance in terms of F-measure among all *codebook* sizes. Table 1 shows the performance summary of the different configurations evaluated in this work. In bold are the best values of precision, recall and F-measure, showing that block-based strategies are more effective in general. An important question here is why SIFT points, that are provided with robust invariant properties, are not a good choice with this type of histopathology images. First, there is some evidence that they were not designed to find the most informative patches for image classification [16], and second, it is possible that all the attempts to increase the invariance of features in histopathology images, lead to a loss of discriminative information.

The next aspect in this evaluation is the image representation, i.e. the use of absolute term frequencies (TF) or the use of the weighted scheme provided by inverse document frequencies (IDF). According to the results presented in Table 1, it is not clear when IDF improves the classification performance. In the case of the SIFT descriptor, IDF produces a poorer performance in most of the cases. In constrast, in the raw-block strategy, the IDF is increasing the importance of discriminative codeblocks, resulting in an improvement of the classification performance. Finally, the use of the RBF kernel in general shows an improvement in precision, either for SIFT points or blocks. However, the recall value is in general hurted by the use of the RBF kernel.

Table 1. Performance measures for the evaluated configurations of the *bag of features*

Kernel function	BOF	SIFT			Raw-Blocks		
		Precision	*Recall*	*F-Measure*	*Precision*	*Recall*	*F-Measure*
Hist. Intersection	*TF*	0.480	0.152	0.207	0.610	0.162	0.234
Hist. Intersection	*IDF*	0.473	0.128	0.189	0.634	0.152	0.231
RBF Kernel	*TF*	0.393	0.146	0.205	0.647	0.123	0.190
RBF Kernel	*IDF*	0.506	0.136	0.165	**0.673**	**0.155**	**0.237**

5 Conclusions and Future Work

This paper presented an evaluation of the *bag of features* approach to classify histopathology images. This is the first systematic evaluation of this representation scheme on this particular medical image domain. The developed evaluation includes a comparative study of different methods and techniques in each stage of the *bag of features* approach. The main advantage of the proposed approach is its adaptiveness to the particular contents of the image collection. Previous work in histology and pathology image analysis used global-generic features or segmentation-based approaches that are not easily extended to other applications even in the same domain.

The adaptiveness property of this framework is obtained with an automated codebook construction, which should have enough patterns to describe the image collection contents. In this domain, we found that the codebook size is very small compared with the codebook size required in other applications such as natural scene classification or even in other kinds of medical images (i.e. mamography and x-rays). The main reason of this smaller size is due to the structure of histopathology images which exhibit homogeneous tissues and the representative visual patterns among the whole collection tends to be uniform.

The bag of features representation is a flexible framework that may be adapted in different ways, either in the visual feature descriptors and the codebook construction. The future work includes a more extensive evaluation of codebook construction methods and different strategies to include color information into visual words as well as more robust texture descriptors. This evaluation will also include a comparison against other commonly used representation schemes.

Acknowledgments

This work has been partially supported by Ministerio de Educación Nacional de Colombia grant 1101393199, according to the COLCIENCIAS call 393-2006 to support research projects using the RENATA network.

References

1. Bosch, A., Muñoz, X., Martí, R.: Which is the best way to organize/classify images by content? Image and Vision Computing 25, 778–791 (2007)
2. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision (2004)

3. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos 2, 1470–1477 (2003)
4. Tommasi, T., Orabona, F., Caputo, B.: CLEF2007 Image annotation task: An SVM-based cue integration approach. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
5. Iakovidis, D.K., Pelekis, N., Kotsifakos, E.E., Kopanakis, I., Karanikas, H., Theodoridis, Y.: A pattern similarity scheme for medical image retrieval. IEEE Transactions on Information Technology in Biomedicine (2008)
6. Long, L.R., Antani, S.K., Thoma, G.R.: Image informatics at a national research center. Computerized Medical Imaging and Graphics 29, 171–193 (2005)
7. Guld, M.O., Keysers, D., Deselaers, T., Leisten, M., Schubert, H., Ney, H., Lehmann, T.M.: Comparison of global features for categorization of medical images. Medical Imaging 5371, 211–222 (2004)
8. Deselaers, T., Keysers, D., Ney, H.: FIRE - Flexible Image Retrieval Engine: imageCLEF 2004 evaluation. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 688–698. Springer, Heidelberg (2005)
9. Datar, M., Padfield, D., Cline, H.: Color and texture based segmentation of molecular pathology images using hsoms. In: 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008, pp. 292–295 (2008)
10. Comaniciu, D., Meer, P., Foran, D.: Shape-based image indexing and retrieval for diagnostic pathology. In: Proceedings on Fourteenth International Conference on Pattern Recognition, vol. 1, pp. 902–904 (1998)
11. Caicedo, J.C., Gonzalez, F.A., Romero, E.: A semantic content-based retrieval method for histopathology images. In: Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., Zhou, G. (eds.) AIRS 2008. LNCS, vol. 4993, pp. 51–60. Springer, Heidelberg (2008)
12. Zheng, L., Wetzel, A.W., Gilbertson, J., Becich, M.J.: Design and analysis of a content-based pathology image retrieval system. IEEE Transactions on Information Technology in Biomedicine 7(4), 249–255 (2003)
13. Lam, R.W.K., Ip, H.H.S., Cheung, K.K.T., Tang, L.H.Y., Hanka, R.: A multi-window approach to classify histological features. In: Proceedings on 15th International Conference on Pattern Recognition, vol. 2, pp. 259–262 (2000)
14. Tang, H.L., Hanka, R., Ip, H.H.S.: Histological image retrieval based on semantic content analysis. IEEE Transactions on Information Technology in Biomedicine 7(1), 26–36 (2003)
15. Fletcher, C.D.M.: Diagnostic Histopathology of tumors. Elsevier Science, Amsterdam (2003)
16. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification, pp. 490–503 (2006)
17. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
18. Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR 2005: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), Washington, DC, USA, vol. 2, pp. 524–531. IEEE Computer Society, Los Alamitos (2005)
19. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)

Improving Probabilistic Interpretation of Medical Diagnoses with Multi-resolution Image Parameterization: A Case Study

Matjaž Kukar and Luka Šajn

University of Ljubljana, Faculty of Computer and Information Science,
Tržaška 25, SI-1001 Ljubljana, Slovenia
{matjaz.kukar,luka.sajn}@fri.uni-lj.si

Abstract. Clinicians strive to improve established diagnostic procedures, especially those that allow them to reach reliable early diagnoses. Diagnostics is frequently performed in a stepwise manner which consists of several consecutive tests (steps). The ultimate step in this process is often the "gold standard" reference method. In stepwise testing, results of each diagnostic test can be interpreted in a probabilistic manner by using prior (pre-test) probability and test characteristics (sensitivity and specificity). By using Bayes' formula on these quantities, the posterior (post-test) probability is calculated. If the post-test probability is sufficiently high (or low) to confirm (or exclude) the presence of a disease, diagnostic process is stopped. Otherwise, it proceeds to the next step in sequence. Our case study focuses on improving probabilistic interpretation of scintigraphic images obtained from the penultimate step in coronary artery disease diagnostics. We use automatic image parameterization on multiple resolutions, based on texture description with specialized association rules. Extracted image parameters are combined into more informative composite parameters by means of principle component analysis, and finally used to build automatic classifiers with machine learning methods. Experiments show that the proposed approach significantly increases the number of reliable diagnoses as compared to clinical results in terms.

1 Introduction

Image parameterization is a technique for describing bitmapped images with numerical parameters – features or attributes. Popular image features are first- and second-order statistics, structural and spectral properties, and several others. Over the past few decades we observe extensive use of image parameterization in medical domains where texture classification is closely related to diagnostic process [4]. This complements medical practice, where manual image parameterization (evaluation of medical images by expert physicians) frequently plays an important role in diagnostic process.

Coronary artery disease (CAD) is one of the world's most premier causes of mortality, and there is an ongoing research for improving diagnostic procedures. The usual clinical process of coronary artery disease diagnostics is stepwise, consisting of four diagnostic levels: (1) evaluation of signs and symptoms of the disease and ECG (electrocardiogram) at rest, (2) ECG testing during the controlled exercise, (3) myocardial scintigraphy and (4) coronary angiography. In this process, the fourth diagnostic level

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 136–145, 2009.

(coronary angiography) is considered as the "gold standard" reference method. As this diagnostic procedure is invasive and rather unpleasant for the patients, as well as relatively expensive, there is a tendency to improve diagnostic performance of earlier diagnostic levels, especially of myocardial scintigraphy [9]. Approaches used for this purpose include applications of neural networks [15], expert systems [6], subgroup mining [5], statistical techniques [19], and rule-based approaches [11]. In our study we focus on various aspects of improving the diagnostic performance of myocardial scintigraphy.

Results of myocardial scintigraphy consist of a series of medical images that are taken both during rest and a controlled exercise. This imaging procedure does not represent a threat to patients' health. In clinical practice, expert physicians use their medical knowledge and experience as well as the image processing capabilities provided by various imaging software to manually describe (parameterize) and evaluate the images.

We propose an innovative alternative to manual image evaluation – automatic multi-resolution image parameterization, based on texture description with specialized association rules, coupled with image evaluation with machine learning methods. Since this approach yields a large number of relatively low-level features (though much more informative than simple pixel intensity values), we recommend using it in conjunction with additional dimensionality reduction techniques, either by feature selection or feature extraction. Our results show that multi-resolution image parameterization equals or even outperforms the physicians in terms of the quality of image parameters. By using automatically generated image description parameters and evaluation with machine learning methods, diagnostic performance can be significantly improved with respect to the results of clinical practice.

2 Methods

An important issue in image parameterization in general, and in our approach with association rules in particular, is to select appropriate resolution(s) for extracting most informative textural features. Structural algorithms use descriptors of some local relations between image pixels where the search perimeter is bounded to a certain size. This means that they can give different results at different resolutions. The resolution used for extracting parameters is important and depends on the observed domain.

We developed the algorithm (ARes – ArTex with Resolutions) that finds suitable resolutions at which image parameterization algorithms achieve more informative features. From our experiments with synthetic data we observe that using parameterization-produced features at several different resolutions usually improves the classification accuracy of machine learning classifiers [20]. This parameterization approach is very effective in analyzing myocardial scintigraphy images used for CAD diagnostics in the stepwise process.

The obtained high quality image parameters can be used for several purposes, among others to describe images with a relatively small number of features. They are subsequently used in machine learning process in order to build a model of diagnostic process. Images corresponding to patients with known correct final diagnosis are used as learning data that, in conjunction with the applied machine learning methods, produces a reliable model (and/or classifier) for the diagnostic problem at hand.

2.1 Stepwise Diagnostic Process

Stepwise diagnostic process[16] is frequently used in clinical practice. Diagnostic tests are ordered in a sequence according to some pre-determined criteria, such as increasing invasiveness, cost, or diagnostic accuracy. Diagnostic process continues until some utility criteria are fulfilled (such as sufficiently high reliability of a diagnosis). Test results can be analyzed by sequential use of the Bayes' conditional probability theorem. The obtained post-test probability accounts for the pre-test probability, sensitivity and specificity of the test, and may later be used as a pre-test probability for the next test in sequence (Figure 1). The first pre-test probability is typically estimated from tables concerning some broader population sample. The process results in a series of tests where each test is performed independently. Its results may be interpreted with or without any knowledge of the other test results.

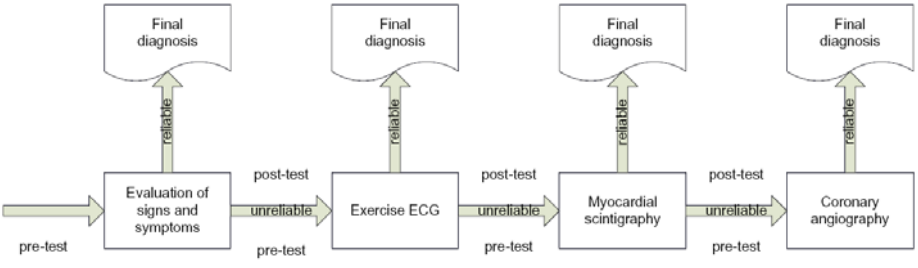


Fig. 1. Increasing the diagnostic test levels in stepwise diagnostic process of CAD

In diagnostic problems, performance of a diagnostic test is described with diagnostic accuracy (*Acc*), sensitivity (*Se*), and specificity (*Sp*). These quantities are subsequently used for post-test probability calculation with Bayes' theorem [16]. Test results from earlier levels are used to obtain the final probability of disease. Diagnostic tests are performed until the post-test probability of disease's presence or absence exceeds some pre-defined threshold value (e.g., 90%). This approach may not only incorporate several test results but also the data from the patient's history [2]. The Bayes' theorem is applied to calculate the conditional probability of the disease's presence, when the result of a diagnostic test is given. For positive or negative test result the respective post-test probabilities $P(d|+) = P(disease|positive\ test\ result)$ or $P(d|-) = P(disease|negative\ test\ result)$ are calculated:

$$P(d|+) = P \cdot Se/(P \cdot Se + (1-P) \cdot (1-Sp)) \quad (1)$$

$$P(d|-) = P \cdot (1-Se)/(P \cdot (1-Se) + (1-P_1) \cdot Sp) \quad (2)$$

2.2 Image Classification with Machine Learning Methods

The ultimate goal of medical image analysis and image mining is decision about the diagnosis. When images are described with informative numerical attributes, we can use various machine learning algorithms for generating a classification system (classifier)

that produces diagnoses of the patients, whose images are being processed [8]. Based on our previous experience in medical diagnostics [9], we decided to use decision trees, naive Bayesian classifiers, Bayesian networks, K-nearest neighbours, and Support Vector Machines.

Our early work in the problem of diagnosing CAD from myocardial scintigraphy images [10] indicates that the naive Bayesian classifier gives best results. Our results conform with several others [7, 8] who also find out that in medical diagnosis the naive Bayesian classifier frequently outperforms other, often much more complex classifiers.

2.3 The ArTex Algorithm and Multi-resolution Image Parameterization

Images in digital form are normally described with spatially complex data matrices. Such data, however, are insufficient to uniformly distinguish between the predefined image classes. Determining image features that can satisfactorily discriminate between observed image classes is a difficult task for which several algorithms exist. [14]. They transform the image from the matrix form into a set of numeric or discrete features (parameters) that convey useful high-level (compared to simple pixel intensities) information for discriminating between classes.

Most texture features are based on structural, statistical or spectral properties of the image. For the purpose of diagnosis from medical images it seems that structural description is most appropriate [21]. We use the ArTex algorithm to obtain textural attributes [20], which are based on spatial association rules. The association rules algorithms can be used for describing textures if an appropriate texture representation formalism is used. Association rules capture structural and statistical information and are very convenient to identify the spatial relations that occur frequently, and have most discriminative characteristic. Texture representation with association rules has several desirable properties: invariance to affine transformations as well as to global brightness. Association rules capture structural and statistical information and identify spatial relations that occur most frequently and have the highest descriptive power.

It is often beneficial to obtain association rules from the same image at several different resolutions, as they may convey different kinds of useful information. This means that we may get completely different image parameterization attributes for the same image at different scales. In existing multi-resolution approaches [1], authors are using only a few fixed resolutions independent of the image contents. Others, however report better results when using more than one (although not more than three) relevant resolutions [20]. For automatic selection of relevant resolutions we use the ARes algorithm [20] that builds upon the well-known SIFT algorithm [13]. ARes proposes several most informative resolutions by counting local intensity peaks ordering them by this count. The user (or an automatic heuristic criterion) then selects how many of the proposed resolutions are to be used.

3 Materials

In our case study we use a dataset of 288 patients with performed clinical and laboratory examinations, exercise ECG, myocardial scintigraphy (including complete image

Table 1. CAD data for different diagnostic levels. Of the attributes belonging to the coronary angiography diagnostic level, only the final diagnosis – the two-valued class – was used.

Diagnostic level	Number of attributes		
	Nominal	Numeric	Total
1. Signs and symptoms	22	5	27
2. Exercise ECG	11	7	18
3. Myocardial scintigraphy (+9 image series)	8	2	10
4. Coronary angiography	1	6	1
Class distribution	129 (46.40%)		CAD negative
	149 (53.60%)		CAD positive

sets) and coronary angiography because of suspected CAD. The features from the ECG and the scintigraphy data were extracted manually by the clinicians. 10 patients were excluded for data pre-processing and calibration required by ArTex/ARes, so only 278 patients (66 females, 212 males, average age 60 years) were used in actual experiments. In 149 cases the disease was angiographically confirmed and in 129 cases it was excluded. The patients were selected from a population of several of thousands patients who were examined at the Nuclear Medicine Department between 2001 and 2006. We selected only the patients with complete diagnostic procedures, and for whom the imaging data was readily available. Some characteristics of the dataset are shown in Table 1.

The myocardial scintigraphy group of attributes consists of evaluation of myocardial defects (no defect, mild defect, well defined defect, serious defect) that could be observed in images either while resting or during a controlled exercise. They are assessed for four different myocardial regions: LAD, LCx, and RCA vascular territories, as well as ventricular apex. Additional two attributes concern effective blood flow and volumes in myocardium: left ventricular ejection fraction (LVEF) and end-diastolic volume (EDV).

3.1 Scintigraphic Images

For each patient a series of images was taken with the General Electric eNTEGRA SPECT camera, both at rest and after a controlled exercise, thus producing the total of 64 grayscale images in resolution of 64×64 and 8-bit pixels. Because of patients' movements and partial obscuring of the heart muscle by other internal organs, these images are not suitable for further use without heavy pre-processing. For this purpose, the ECToolbox workstation software [3] was used, and one of its outputs, a series of 9 polar map (bull's eye) images were taken for each patient. Polar maps were chosen since previous work in this field [12] had shown that they have useful diagnostic value.

Unfortunately, in most cases (and especially in our specific population) the differences between images taken during exercise and at rest are not as clear-cut as shown in Figure 2. Interpretation and evaluation of scintigraphic images therefore requires considerable knowledge and experience of expert physicians. Although specialized tools such as the ECToolbox software can aid in this process, they still require special training and in-depth medical knowledge for evaluation of results.

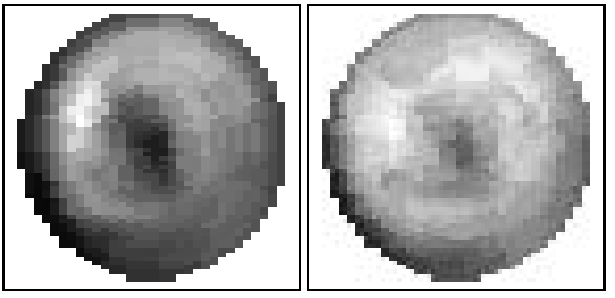


Fig. 2. Typical polar maps taken after exercise (left), and at rest (right). Shadows in the center of both images suggest inadequately perfused myocardial tissue, especially during exercise (left image). Images shown in this figure correspond to the patient with a very clear manifestation of CAD. Taken from [10] with permission from authors.

4 Results

Experiments were performed in the following manner. First, 10 learning examples (images or sets of nine images for CAD) were excluded for data preprocessing and calibration of ArTex/ARes. Images from the remaining examples were parameterized; only the obtained parameters were subsequently used for evaluation. Further testing was performed in the ten-fold cross-validation setting: at each step 90% of examples were used for building a classifier, and the remaining 10% of examples for testing. In each cross-validation step the real-valued attributes were discretized in advance by using the Fayyad-Irani algorithm, if the applied method (such as the naive Bayesian classifier) required only discrete attributes.

In CAD diagnostics that combines generated parameters for nine images, number of parameters (attributes) was reduced with feature extraction – by applying the principal component analysis (PCA) and retaining only the best principal components (those that together accounted for not less than 70% of data variance, amounting to 10 best components). Besides the described 10 components, an equal number of best attributes provided by physicians was used (as estimated by the ReliefF algorithm [18]).

We applied four popular machine learning algorithms: naive Bayesian classifier, tree-augmented Bayesian network, support vector machine (SMO using RBF kernel), and J4.8 (C4.5) decision tree. We performed experiments with both Weka and Orange machine learning toolkits. For CAD diagnostics, aggregated results of the coronary angiography (CAD negative/CAD positive) were used as the class variable. The results of clinical practice were validated by careful blind evaluation of images by an independent expert physician. Significance of differences to clinical results was evaluated by using the McNemar's test.

4.1 Results in CAD Diagnostics

Out of the 278 patients with 9 images, each of them was parameterized for three resolutions in advance. ARes proposed three[1] resolutions: $0.95\times$, $0.80\times$, and $0.30\times$ of the

[1] A resolution of $0.30\times$ means $0.30 \cdot 64 \times 0.30 \cdot 64$ pixels instead of 64×64 pixels.

Table 2. Experimental results of machine learning classifiers on parameterized images obtained by selecting only the best 10 attributes from PCA on ArTex/ARes (also combined with 10 best attributes provided by physicians). Classification accurracy results that are significantly better ($p < 0.05$) than clinical results are emphasized.

	PCA on ArTex/ARes			PCA on ArTex/ARes+physicians		
	Accurracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity
Naive Bayes	**81.3%**	83.7%	79.2%	**79.1%**	82.9%	75.8%
Bayes Net	**71.9%**	69.0%	74.5%	**79.1%**	83.7%	75.2%
SMO (RBF)	**78.4%**	76.0%	80.1%	**76.6%**	77.5%	75.8%
J4.8	**75.2%**	78.3%	72.5%	**74.1%**	73.6%	74.5%
Clinical	64.0%	71.1%	55.8%	64.0%	71.1%	55.8%

original resolution, producing together 2944 additional attributes. Since this number is too large for most practical purposes, it was reduced to 10 by applying feature extraction (with PCA). We also enrich the data representation by using the same number (10) of best physicians' attributes as evaluated by ReliefF and compare the results of machine learning with diagnostic accuracy, specificity and sensitivity of expert physicians after evaluation of scintigraphic images. It is gratifying to see that without any special tuning of learning parameters, the results are in all cases significantly better than the results of physicians in terms of classification (diagnostic) accuracy. Especially good results are that of the naive Bayesian classifier (Table 2 and Figure 3), that improve in all three criteria: diagnostic accuracy, sensitivity and specificity. This is not unexpected, as it conforms with ours and others previous experimental results [7, 8, 10]. Another interesting issue is that including the best physician-provided attributes does not necessarily improve diagnostic performance (SMO, J4.8 in Table 2). It seems that there is some level of redundancy between physicians' and principal components generated from ArTex/ARes attributes, that bothers some methods more than the others. Consequently, it seems that some of automatically generated attributes are (from the diagnostic performance point of view) at least as good as the physician-provided ones, and may therefore represent new knowledge about CAD diagnostics.

4.2 Assessing the Diagnostic Power

We experimented with different methods for assessing reliability (probability of a correct diagnosis) of machine learning classifications in stepwise diagnostic process, as described in [16]. To determine the pretest probability we applied tabulated values as given in [17]. For each patient, the table was indexed by a subset of "signs and symptoms" attributes (age, sex, type of chest pain).

For both physicians and machine learning methods we calculated the post-test probabilities in the stepwise manner, starting from the pre-test probability and proceeding with evaluation of signs and symptoms, exercise ECG, and myocardial scintigraphy. For myocardial scintigraphy, physicians achieved 64% diagnostic accuracy, 71.1% specificity, and 55.8% sensitivity. For the reliability threshold of 90%, 52% of diagnoses could be considered as reliable (their post-test probability was higher than 90% for positive, or lower than 10% for negative diagnoses). On the other hand, naive Bayesian

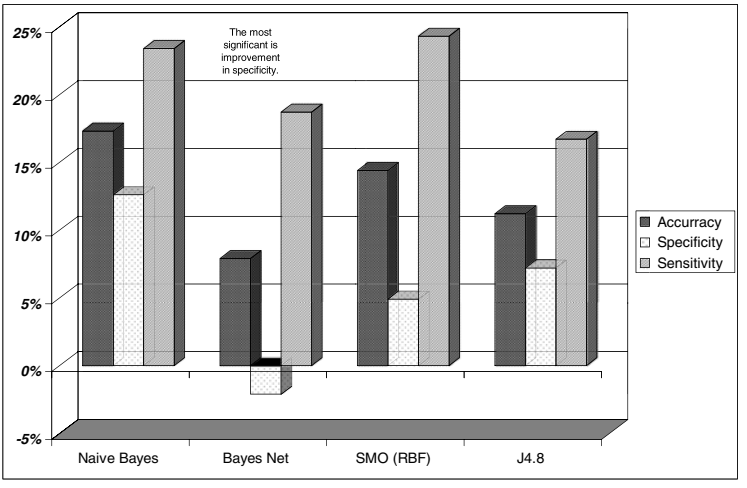


Fig. 3. Impovements of machine learning classifiers on parameterized images from Table 2 relative to clinical results (baseline 0%)

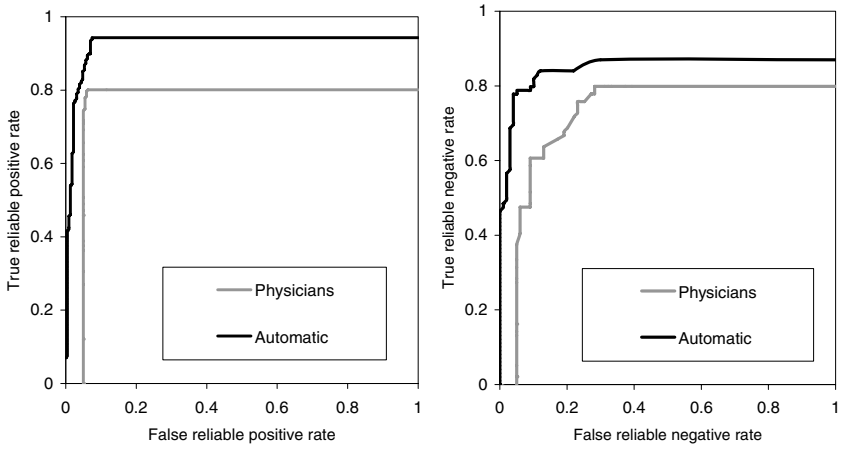


Fig. 4. ROC curves, obtained by varying reliability threshold between 0 and 1, for reliable positive diagnoses (left) and reliable negative diagnoses (right)

classifier achieved for myocardial scintigraphy 81.3% diagnostic accuracy, 83.7% specificity, and 79.2% sensitivity. For the reliability threshold of 90%, 69% of diagnoses could be considered as reliable. Improvement of 17% overall is a result of 19% improvement for positive diagnoses, and 16% for negative diagnoses.

We also depict results of both approaches in ROC curves, obtained by varying reliability threshold between 0 and 1 (Figure 4). Fully automatic approach (Naive Bayes on parameterized images) has considerably higher ROC curve, both for reliable

positive (AUC=0.87 vs. 0.77) and reliable negative patients (AUC=0.81 vs. 0.72). Of both improvements in positive and negative reliable diagnoses, by far the more important is the 16% improvement for reliable negative diagnoses. The reason for this is that positive patients undergo further pre-operative diagnostic tests in any case, while for negative patients diagnostic process can reliably be finished on the myocardial scintigraphy level.

5 Discussion

We describe an innovative alternative to manual image evaluation - automatic multi-resolution image parameterization based on spatial association rules (ArTex/ARes) supplemented with feature selection or (preferably) feature extraction. Our results show that multi-resolution image parameterization equals or even betters the physicians in terms of diagnostic quality of image parameters. By using these parameters for building machine learning classifiers, diagnostic performance can be significantly improved with respect to the results of clinical practice. We also explore relations between newly generated image attributes and physicians' description of images. Our findings indicate that ArTex/ARes with PCA is likely to extract more useful information from images than the physicians do, as it significantly outperforms them in terms of diagnostic accuracy, specificity and sensitivity, as well as in the number of reliable diagnoses.

Utilizing machine learning methods can help less experienced physicians evaluate medical images and thus improve their performance both in accuracy as well as in sensitivity and specificity. From the practical use of described approaches two-fold improvements of the diagnostic procedure can be expected. Higher diagnostic accuracy (up to 17.3%) is by itself a very considerable gain. Due to higher specificity of tests (up to 12%), fewer patients without the disease would have to be examined with coronary angiography which is invasive and therefore dangerous method. Together with higher sensitivity and more reliable diagnoses (17% improvement) this would save money and shorten the waiting times of the truly ill patients. Also, new attributes, generated by ArTex/ARes with PCA had invoked considerable interest from expert physicians, as they significantly contribute to increased diagnostic performance and may therefore convey some novel medical knowledge of the CAD diagnostics problem. This could represent a significant improvement in the diagnostic power as well as in the rationalization of the existing CAD diagnostic procedure without danger of incorrectly diagnosing more patients than in current practice.

Last but not least, we again emphasize that the results of our study are obtained on a significantly restricted population and therefore may not be generally applicable to the normal population, i.e. to all the patients coming to the Nuclear Medicine Department, University Clinical Centre Ljubljana, Slovenia.

Acknowledgements. We thank dr. Ciril Grošelj, M. D., University Clinical Centre Ljubljana, for comments and cooperation. This work was supported by the Slovenian Ministry of Higher Education, Science, and Technology.

References

[1] Comer, M.L., Delp, E.J.: Segmentation of textured images using a multiresolution gaussian autoregressive model. IEEE Transactions on image processing 8(3), 408–420 (1999)

[2] Diamond, G.A., Forester, J.S.: Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. New England Journal of Medicine 300(1350) (1979)

[3] General Electric. Ectoolbox protocol operator's guide (2001)

[4] Fitzpatrick, J., Sonka, M.: Handbook of Medical Imaging, Medical Image Processing and Analysis, vol. 2. SPIE, Bellingham (2000)

[5] Gamberger, D., Lavrac, N., Krstacic, G.: Active subgroup mining: a case study in coronary heart disease risk group detection. Artif. Intell. Med. 28(1), 27–57 (2003)

[6] Garcia, E.V., Cooke, C.D., Folks, R.D., Santana, C.A., Krawczynska, E.G., De Braal, L., Ezquerra, N.F.: Diagnostic performance of an expert system for the interpretation of myocardial perfusion spect studies. J. Nucl. Med. 42(8), 1185–1191 (2001)

[7] Kononenko, I.: Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine 3, 89–109 (2001)

[8] Kononenko, I., Kukar, M.: Machine Learning and Data Mining: Introduction to Principles and Algorithms. Horwood publ. (2007)

[9] Kukar, M., Kononenko, I., Grošelj, C., Kralj, K., Fettich, J.: Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artificial Intelligence in Medicine 16(1), 25–50 (1999)

[10] Kukar, M., Šajn, L., Grošelj, C., Grošelj, J.: Multi-resolution image parametrization in sequential diagnostics of coronary artery disease. In: Bellazzi, R., Abu-Hanna, A., Hunter, J. (eds.) AIME 2007. LNCS, vol. 4594, pp. 119–129. Springer, Heidelberg (2007)

[11] Kurgan, L.A., Cios, K.J., Tadeusiewicz, R.: Knowledge discovery approach to automated cardiac spect diagnosis. Artif. Intell. Med. 23(2), 149–169 (2001)

[12] Lindahl, D., Palmer, J., Pettersson, J., White, T., Lundin, A., Edenbrandt, L.: Scintigraphic diagnosis of coronary artery disease: myocardial bull's-eye images contain the important information. Clinical Physiology 6(18) (1998)

[13] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)

[14] Nixon, M., Aguado, A.S.: Feature Extraction and Image Processing, 2nd edn. Academic Press, Elsevier (2008)

[15] Ohlsson, M.: WeAidU–a decision support system for myocardial perfusion images using artificial neural networks. Artificial Intelligence in Medicine 30, 49–60 (2004)

[16] Olona-Cabases, M.: The probability of a correct diagnosis. In: Candell-Riera, J., Ortega-Alcalde, D. (eds.) Nuclear Cardiology in Everyday Practice, pp. 348–357. Kluwer Academic Publishers, Dordrecht (1994)

[17] Pollock, B.H.: Computer-assisted interpretation of noninvasive tests for diagnosis of coronary artery disease. Cardiovasc. Rev. Rep. 4, 367–375 (1983)

[18] Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning 53, 23–69 (2003)

[19] Slomka, P.J., Nishina, H., Berman, D.S., Akincioglu, C., Abidov, A., Friedman, J.D., Hayes, S.W., Germano, G.: Automated quantification of myocardial perfusion spect using simplified normal limits. J. Nucl. Cardiol. 12(1), 66–77 (2005)

[20] Šajn, L., Kononenko, I.: Multiresolution image parametrization for improving texture classification. EURASIP J. Adv. Signal Process 2008(1), 1–12 (2008)

[21] Šajn, L., Kononenko, I.: Image segmentation and parametrization for automatic diagnostics of whole-body scintigrams. In: Computational Intelligence in Medical Imaging: Techniques & Applications, pp. 347–377. CRC Press, Boca Raton (2009)

Segmentation of Lung Tumours in Positron Emission Tomography Scans: A Machine Learning Approach

Aliaksei Kerhet[1], Cormac Small[2], Harvey Quon[2], Terence Riauka[3], Russell Greiner[4], Alexander McEwan[1], and Wilson Roa[2]

[1] Department of Oncology, University of Alberta,
11560 University Avenue, T6G 1Z2 Edmonton AB, Canada
{kerhet,mcewan}@ualberta.ca
[2] Department of Radiation Oncology, Cross Cancer Institute,
11560 University Avenue, T6G 1Z2 Edmonton AB, Canada
{cormacsm,harveyqu,wilsonro}@cancerboard.ab.ca
[3] Department of Medical Physics, Cross Cancer Institute,
11560 University Avenue, T6G 1Z2 Edmonton AB, Canada
terencer@cancerboard.ab.ca
[4] Alberta Ingenuity Centre for Machine Learning,
Department of Computing Science, University of Alberta,
3-59 Athabasca Hall, T6G 2E8 Edmonton AB, Canada
greiner@cs.ualberta.ca

Abstract. Lung cancer represents the most deadly type of malignancy. In this work we propose a machine learning approach to segmenting lung tumours in Positron Emission Tomography (PET) scans in order to provide a radiation therapist with a "second reader" opinion about the tumour location. For each PET slice, our system extracts a set of attributes, passes them to a trained Support Vector Machine (SVM), and returns the optimal threshold value for distinguishing tumour from healthy voxels in that particular slice. We use this technique to analyse four different PET/CT 3D studies. The system produced fairly accurate segmentation, with Jaccard and Dice's similarity coefficients between 0.82 and 0.98 (the areas outlined by the returned thresholds vs. the ones outlined by the reference thresholds). Besides the high level of geometric similarity, a significant correlation between the returned and the reference thresholds also indicates that during the training phase, the learning algorithm effectively acquired the dependency between the extracted attributes and optimal thresholds.

Keywords: Support Vector Machine (SVM), Positron Emission Tomography (PET), Radiation Treatment, Lung Cancer, Gross Tumour Volume (GTV).

1 Introduction

According to the Canadian Cancer Society reports [1,2], lung cancer represents the second most common type of cancer (approximately 23,900 new cases were

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 146–155, 2009.

expected in 2008 in Canada alone), and the one most fatal to both men and women (it is responsible for more than 1/4 of all cancer-associated deaths). Even for younger adults (aged 20-44), it is ranked first (men) and second (women) with respect to the potential number of years of life lost.

Radiation therapy involves applying beams of ionizing radiation to irradiate the tumour volume. Present-day equipment allows these beams to be directed very accurately. However, this is only effective if one can define (segment) a tumour with a similar accuracy. Unfortunately, this is not currently the case.

The conventional way to define lung tumours is based on the analysis of *computed tomography* (CT) images. The sensitivity and specificity of this imaging modality is not always high enough, which leads to significant levels of the intra- and inter-observer variability. Introduction of the *positron emission tomography* (PET) imaging modality to the process of lung tumour definition has already been shown to alter the results and decrease the variability above. However, due to some challenges related to the analysis of PET scans, the role of this modality in radiation treatment planning has not yet been well established.

This paper uses a machine learning approach to address some of these challenges. For each PET slice, our system extracts a set of attributes, passes them to a trained Support Vector Machine (SVM), returns the optimal threshold value, and applies it for distinguishing tumour from healthy voxels in that particular slice. This automatically provides a radiation therapist with a "second reader" opinion about the tumour location.

The remaining sections of this paper are organised as follows. The next section reviews the state-of-the-art for tumour segmentation in PET scans. Section 3 describes the proposed approach and provides a brief introduction into SVM for regression estimation. The experimental part and results are described in Sect. 4, followed by Sect. 5, dedicated to the discussion and conclusion.

2 PET in Radiation Therapy: Background

Unlike CT imaging, which provides an anatomical description of a scanned body, PET imaging visualises the functionality of the body cells. Prior to a PET imaging study, the patient is injected with a radioactively marked substance, which is absorbed and metabolised differently by different types of cells. The radioactivity emitted from each region of the body is then registered, and the reconstructed images visualise quantities of the substance uptake, measured in counts of radioactive decays or some other uptake values. As cancerous cells are known to absorb more sugar than surrounding healthy tissue for many organs, most PET studies use a radioactively labelled analogue of sugar called *fluorodeoxyglucose* to visualise tumours. Today PET is primarily used in diagnostics as an indispensable technique for characterizing neoplasms and detecting distant metastases.

Besides diagnostics applications, adding PET to radiation treatment planning is also considered beneficial, as compared to using CT alone. As these two modalities are built on completely different underlying phenomena, they supply

a radiation oncologist with two different and complementary perspectives on the problem of tumour segmentation. In a recent review on lung cancer, authors of all 18 different studies (involving 661 patients) agreed that PET adds essential information, affecting the results of tumour segmentation [3].

However, the role of PET in radiation treatment planning is not well established, mainly due to the following challenges. First, PET images lack the sharpness and clearness of CT scans (Fig. 1). Second, sizes of objects in PET images strongly depend on the visualisation setup (contrast and brightness value, known as window/level setup in medical imaging; see Fig. 1, three rightmost images), and the optimal setup can vary across patients, and even across slices within a patient. Finally, using both PET and CT scans for tumour definition implies that the two scans must be co-registered, which is challenging due to PET's blurriness.

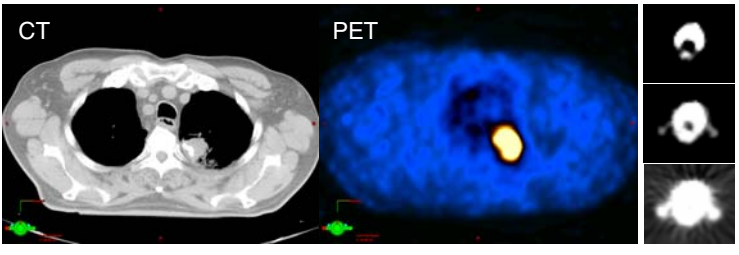


Fig. 1. Corresponding CT and PET slices of a human thorax. While PET image is much more blurred than CT, the tumor area is much more evident in PET than in CT. Right: a slice of a mouse PET scan displayed using three different visualisation setups, which lead to different sizes of objects.

In the most basic case, an experienced nuclear medicine physician and/or radiation oncologist will visually interpret both PET and CT images to determine the tumour borders. Several automatic segmentation techniques have been proposed to make the interpretation of PET scans observer-independent and to cope with the challenge of choosing the right visualisation setup. They can be broadly divided into two groups. The first broad group aims to segment the tumour by searching for some inhomogeneity throughout the PET scan. Although there are some interesting examples from this group, such as gradient-based (watershed) methods [4,5] and a multimodal generalisation of level set method [6], they are not as well established nor as frequently cited in current reviews as the methods from the second group, which aim to define the optimal threshold value of the uptake in order to segment a tumour. This second group includes approaches that define the optimal threshold as some fixed uptake value, or a fixed percentage of the maximum uptake value; other more sophisticated approaches determine the optimal threshold as the weighted sum of mean target uptake and mean background uptake, among other tecniques [7,8,9,10,11,12]. Note that methods from the second group define a *single* optimal threshold value for the whole PET 3-D scan.

3 Proposed Approach

Our approach falls into the second group above: our system defines the optimal threshold value of the uptake, and declares each voxel to be cancerous if its uptake value is above that threshold. However, given the complexity of segmenting lung tumours in PET scans, we consider the optimal threshold as a *non-linear* function of *more* than just one or two attributes. We therefore extract a richer set of attributes from PET scans, and use a machine learning algorithm, capable of incorporating more complex dependencies based on these attributes, to find the (potentially different) optimal threshold value for *each* slice in a PET scan.

3.1 PET Attributes

Several works that compared and reviewed different threshold-based lung tumour delineation algorithms suggested the use of contrast-oriented algorithms [3,8]. Nestle *et al.* [8] defines the optimal threshold value *for the whole* PET 3-D scan as $U_{\text{bg}} + 0.15 \times U_{70}$, where the scalar U_{bg} is the mean uptake of a background (some homogeneous 3-D area near a tumour, e.g. mediastinum), and the scalar U_{70} is the mean uptake inside the 3-D isocontour of 70% of the maximum uptake. However, other studies suggest that even within the same 3-D scan, the optimal threshold value can vary from slice to slice with the tumour volume/cross-sectional area [11]. Our own observations also support this claim (Sect. 4.2 and Fig. 3). In line with the two considerations above, we aim to define the optimal threshold value *for each PET slice individually*, based on U_{bg} and the following 6 scalar attributes extracted from the given PET slice:

- the area and mean uptake inside the $0.10 \times U_{70}$ contour
- the area and mean uptake inside the $0.15 \times U_{70}$ contour
- the area and mean uptake inside the $0.20 \times U_{70}$ contour.

3.2 SVM Regressor

Support Vector Machine (SVM) [13,14] is a very successful machine learning algorithm which has been used effectively for a wide variety of tasks, ranging from optical character recognition and electricity load prediction to biomedicine and face detection/recognition. In this work we use SVM for *regression estimation.*

During the training phase, a *training set* (available examples of the values of the seven attributes above (vector $\mathbf{x}$) and the corresponding optimal threshold values (scalar y) is analysed by the SVM. As a result of this analysis, a subset of the most important, characteristic examples (called *support vectors*) is identified and used to build the regression function, which is then used to predict optimal thresholds for new attribute vectors. In the most basic case, this is a linear function of the attributes. However, such a linear regressor is not always "expressive" enough to reflect the complexity of real-world problems. This obstacle is overcome using the so-called *kernel trick* [14], which consists in implicit mapping the vector of attributes $\mathbf{x}$ onto a higher-dimensional space: $\mathbf{\Phi} : \mathbf{x} \rightarrow \mathbf{\Phi}(\mathbf{x})$, where

the data is more likely to allow linear approximation. Thanks to kernels, all calculations actually occur in the lower-dimensional space where vectors $\mathbf{x}$ live, which is extremely beneficial from the computational point of view. Explicitly this leads to the following form of the regression function:

$$y(\mathbf{x}) = \sum_{i=1}^{N_{\text{sv}}} a_i K(\mathbf{x}_i, \mathbf{x}) + b \ , \tag{1}$$

where N_{sv} is the number of support vectors $\mathbf{x}_i$, and $K(\mathbf{u}, \mathbf{v}) = \mathbf{\Phi}(\mathbf{u}) \cdot \mathbf{\Phi}(\mathbf{v})$ is a kernel function that implicitly calculates dot product in the higher-dimensional space where vectors $\mathbf{\Phi}(\mathbf{x})$ live. One of the most used and studied kernels is the so-called Gaussian kernel: $K(\mathbf{u}, \mathbf{v}) = e^{-\gamma \|\mathbf{u}-\mathbf{v}\|^2}$. For linear SVMs, the corresponding kernel is simply a dot product in the space where samples live: $K(\mathbf{u}, \mathbf{v}) = \mathbf{u} \cdot \mathbf{v}$. For a brief introduction into SVM regression, the reader is referred to [15].

A comprehensive scheme of our system is shown in Fig. 2.

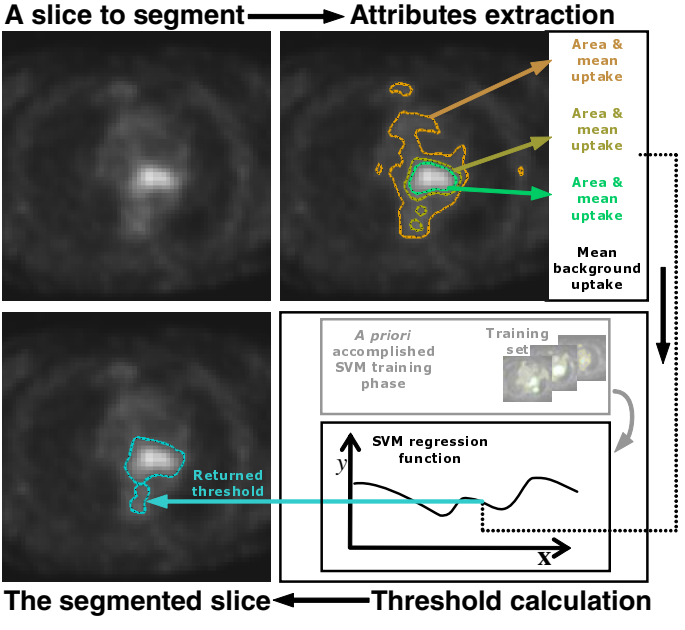


Fig. 2. A comprehensive scheme of the proposed approach

4 Experiments and Results

4.1 Initial Data

In this work we analyzed the data of two patients. Each patient underwent two same-day studies (free-breathing one and gated one [16]); and each study

comprised three scans: a diagnostic CT and a fluorodeoxyglucose PET obtained with a hybrid PET/CT scanner (Philips Gemini); and a separate treatment planning CT scan. Using a hybrid scanner ensured that the corresponding PET and CT scans were automatically and perfectly co-registered, eliminating the PET-CT co-registration challenge described in Sect. 2. This is important since it allows us to attribute the obtained results solely to the attributes extraction and the algorithm used, rather than to unwanted artifacts of the PET-CT co-registration.

For each study, a treatment planning CT scan was then manually co-registered with a diagnostic CT scan (a CT-CT co-registration is not as challenging as a PET-CT one), thus spatially linking all three scans. Two experienced radiation oncologists then used this rich and high-quality information in order to manually draw a tumour volume for the radiation treatment planning. This volume is referred to as a *gross tumour volume* (GTV). Using GTVs produced by the consensus of two radiation oncologists based on co-registered PET and CT data ensures that the obtained GTVs are of high quality.

4.2 Data Sets Generation

Each study of each patient was analysed separately and independently (i.e. we used a *study-specific* scenario, where both training and test sets are obtained from the same scan). Therefore, U_{bg}, which characterizes the whole scan rather than a specific slice (see Sect. 3.1), vanished from the consideration as a constant. Tumour-containing PET slices and 8 adjacent tumour-free slices were extracted. For each of these PET slices our system computed the values for 6 attributes described in Sect. 3.1. Also, for each of these slices a *reference uptake threshold* was assigned by the consensus of two radiation oncologists as the threshold that produced the segmentation most closely approximating the corresponding GTV contour. For each tumour-free slice the maximum uptake of that slice was used as the reference uptake threshold. The slice-to-slice variation of reference uptake thresholds for patient 2 is shown in Fig. 3.

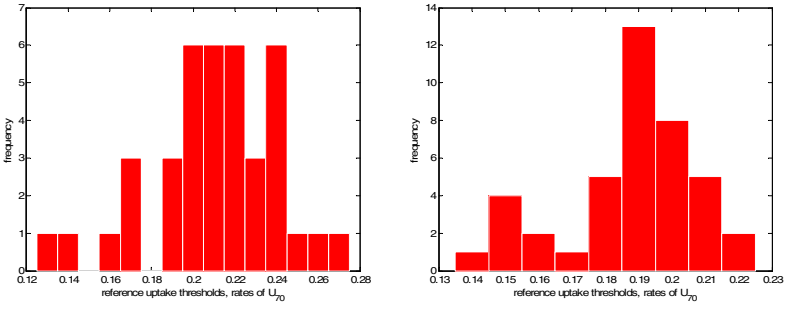


Fig. 3. Slice-to-slice variation of reference uptake thresholds, as demonstrated by their histograms for free-breathing (left) and gated (right) studies of patient 2

PET slices were randomly split in two groups in order to form a training set (75% of slices) and a test set (the remaining 25% of slices). This random splitting was repeated 5 times (resampling), resulting in 5 different pairs of training and test sets for the same study. The characteristics of data sets for different patients/studies are summarized in Table 1.

Table 1. Summary of the data sets

Patient/study	N	N_+	N_-	N_{train}	N_{test}
1/gated	32	24	8	24	8
1/fb	27	19	8	20	7
2/gated	41	33	8	30	11
2/fb	39	31	8	29	10

fb: free-breathing; N: total number of the slices extracted for the given patient/study; N_+: number of slices containing tumour; N_-: number tumour-free slices; N_{train}: number of slices used to form the training set (randomly selected from N slices); N_{test}: number of slices used to form the test set (remaining $N - N_{\text{train}}$ slices)

4.3 SVM Training and Model Selection

We used μ-SVM regression estimation [15] with Gaussian kernel. This variety of SVM algorithm has three parameters (γ, C and μ) that must be set during the training phase. We approached the problem of finding their optimal values (*model selection*) by performing a five-fold cross-validation [17] on a logarithmic grid. The total training time per study was about 15 minutes (selection from 1764 different combinations of the parameters by means of 5-fold cross-validation for five different training sets).

4.4 Results

The following three metrics were used to evaluate the results. First, the correlation coefficient was calculated between the reference thresholds and those predicted by the algorithm. The other two measures evaluate the quality of the results in terms of geometric similarity of the regions contoured with the reference thresholds, and the regions outlined by the algorithm-predicted thresholds. To this end, Jaccard and Dice's similarity indices were calculated

$$J = |R \cap A| / |R \cup A| \tag{2}$$

$$D = 2|R \cap A| / (|R \cup A| + |R \cap A|) \ , \tag{3}$$

where R and A stand for the regions contoured by the reference and algorithm-predicted thresholds, respectively. Both Jaccard and Dice's indices are equal to zero when two regions have no common area, and are equal to unity when the regions match perfectly.

The results are summarized in Table 2 (since no significant difference was found between the results of the gated and free-breathing studies in the same

patient, we present their averages). Several segmentation examples are presented in Fig. 4. Besides the high level of geometric similarity, a correlation between the predicted and the reference uptake thresholds also indicates that during the training phase, the learning algorithm effectively acquired the dependency between the attributes and the reference uptake threshold.

Table 2. Summary of the results

Patient	C	J	D
1	0.71	0.82 (0.61)	0.89 (0.74)
2	0.83	0.96 (0.77)	0.98 (0.87)

C: correlation coefficient between the reference and algorithm-predicted thresholds; J and D: Jaccard and Dice's similarity indices between the regions contoured by the reference and algorithm-predicted thresholds (the values in parentheses represent the results obtained with the contrast-oriented algorithm [8], see Sect. 3.1)

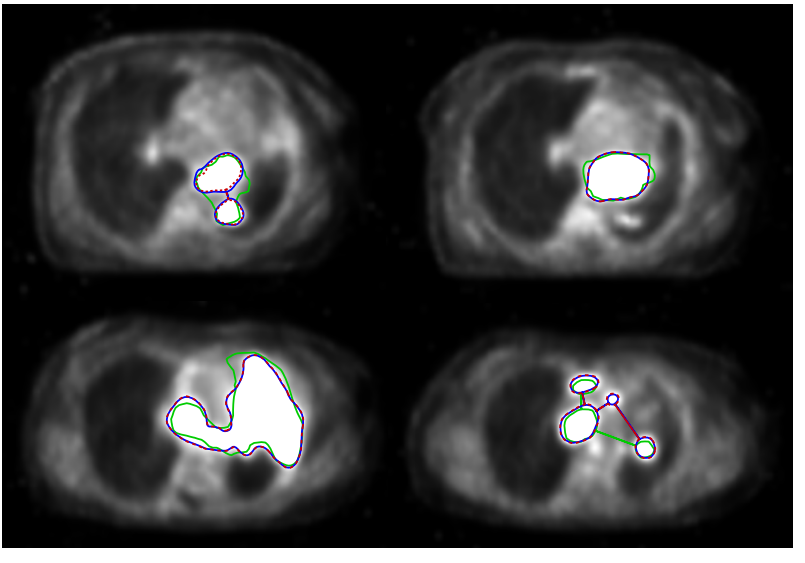


Fig. 4. Examples of GTV, reference, and predicted uptake thresholds. Grey contour (green in colour version): GTV; solid black contour (blue in colour version): region contoured by the reference uptake threshold; dashed black contour (dashed red in colour version): region contoured by the algorithm-predicted uptake threshold.

Table 2 shows that we obtained better results for the second patient. We think this is because the second patient had a bigger tumour, occupying about 30% more PET slices (see Table 1), which resulted in a bigger training set, and hence better generalisation during training.

5 Discussion, Conclusion and Future Work

Our approach to PET-based lung tumour definition extends standard threshold-based approaches in three important ways. First, we base the definition of the optimal thresholds on a richer set of attributes extracted from the PET scans. Second, we use an "adaptable" machine learning algorithm capable of approximating data in a complex nonlinear way. Finally, we estimate the optimal threshold for each PET slice (instead of assigning a single threshold to all slices in the study).

The two threshold contours (reference and predicted) shown in Fig. 4 look very similar. However, this does not guarantee a high similarity between them and the GTV, as in case of the upper leftmost image, where both predicted and reference regions are composed of two contours, whereas the corresponding GTV is a single contour including some additional area. This illustrates an inherent limitation of *any* approach that is based on thresholding. This is simply because the shape of GTV contour can be whatever a radiation oncologist draws. In contrast, the shape of any particular threshold contour for a given image is fixed; and therefore choosing from even infinite number of different thresholds means choosing from an infinite number of different, but FIXED shapes.

This work was performed using a *study-specific* scenario, to cope with the slice-to-slice variation of the optimal threshold values within a study, and 75% of slices should have been defined manually in order to automatically define the remaining 25%. In order to prove its fitness for the real clinical use, an *inter-study/inter-patient* scenario with a sufficiently high number of patients is necessary; and we are currently exploring the challenges of collecting the necessary cases (i.e. checking and confronting both radiation treatment and diagnostic data available at out institution). In addition, we are looking for another informative PET attributes.

This work has demonstrated the potential advantages and applicability of the machine learning methodology as a tool to help plan radiation treatment for lung cancer.

Acknowledgments. This project has been made possible through a grant from the Alberta Cancer Board and the Alberta Cancer Foundation. R.G. was partially funded by NSERC and the Alberta Ingenuity Centre for Machine Learning.

References

1. Canadian cancer statistics 2008. Canadian Cancer Society (2008)
2. Parkin, D.M., Bray, F., Ferlay, J., Pisani, P.: Global cancer statistics, 2002. CA Cancer J. Clin. 55(2), 74–108 (2005)
3. Nestle, U., Kremp, S., Grosu, A.L.: Practical integration of [18F]-FDG-PET and PET-CT in the planning of radiotherapy for non-small cell lung cancer (NSCLC): the technical basis, ICRU-target volumes, problems, perspectives. Radiother Oncol. 81(2), 209–225 (2006)

4. Drever, L.A., Roa, W., McEwan, A., Robinson, D.: Comparison of three image segmentation techniques for target volume delineation in positron emission tomography. J. Appl. Clin. Med. Phys. 8(2), 93–109 (2007)
5. Geets, X., Lee, J.A., Bol, A., Lonneux, M., Gregoire, V.: A gradient-based method for segmenting FDG-PET images: methodology and validation. Eur. J. Nucl. Med. Mol. Imaging 34(9), 1427–1438 (2007)
6. Naqa, I.E., Yang, D., Apte, A., Khullar, D., Mutic, S., Zheng, J., Bradley, J.D., Grigsby, P., Deasy, J.O.: Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning. Med. Phys. 34(12), 4738–4749 (2007)
7. Black, Q.C., Grills, I.S., Kestin, L.L., Wong, C.Y.O., Wong, J.W., Martinez, A.A., Yan, D.: Defining a radiotherapy target with positron emission tomography. Int. J. Radiat Oncol. Biol. Phys. 60(4), 1272–1282 (2004)
8. Nestle, U., Kremp, S., Schaefer-Schuler, A., Sebastian-Welsch, C., Hellwig, D., Rube, C., Kirsch, C.M.: Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-Small cell lung cancer. J. Nucl. Med. 46(8), 1342–1348 (2005)
9. Nestle, U., Schaefer-Schuler, A., Kremp, S., Groeschel, A., Hellwig, D., Rube, C., Kirsch, C.M.: Target volume definition for 18F-FDG PET-positive lymph nodes in radiotherapy of patients with non-small cell lung cancer. Eur. J. Nucl. Med. Mol. Imaging 34(4), 453–462 (2007)
10. Daisne, J.F., Sibomana, M., Bol, A., Doumont, T., Lonneux, M., Gregoire, V.: Tridimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. Radiother. Oncol. 69(3), 247–250 (2003)
11. Drever, L., Robinson, D.M., McEwan, A., Roa, W.: A local contrast based approach to threshold segmentation for PET target volume delineation. Med. Phys. 33(6), 1583–1594 (2006)
12. Drever, L., Roa, W., McEwan, A., Robinson, D.: Iterative threshold segmentation for PET target volume delineation. Med. Phys. 34(4), 1253–1265 (2007)
13. Vapnik, V.N.: The Nature of Statistical Learning Theory. In: Statistics for Engineering and Information Science, 2nd edn. Springer, Heidelberg (1999)
14. Scholkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge (2002)
15. Smola, A.J., Scholkopf, B.: A tutorial on support vector regression. Neurocolt2 technical report series (October 1998)
16. Boucher, L., Rodrigue, S., Lecomte, R., Benard, F.: Respiratory gating for 3-dimensional PET of the thorax: feasibility and initial results. J. Nucl. Med. 45(2), 214–219 (2004)
17. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. In: Data Mining, Inference, and Prediction. Springer, Heidelberg (2001)

A System for the Acquisition, Interactive Exploration and Annotation of Stereoscopic Images

Karim Benzeroual[1,2], Mohammed Haouach[1,2], Christiane Guinot[1,2], and Gilles Venturini[1]

[1] Computer Science Laboratory, University François-Rabelais of Tours
64 Avenue Jean Portalis, 37200 Tours, France
{benzeroual,haouach,venturini}@univ-tours.fr
[2] CE.R.I.E.S., Biometrics and Epidemiology unit,
20 Rue Victor Noir, 92521 Neuilly sur Seine, France
christiane.guinot@ceries-lab.com

Abstract. We present in the paper a system that integrates all hardware and software to extract information from 3D images of skin. It is composed of a lighting equipment and stereoscopic cameras, a camera calibration algorithm that uses evolutionary principles, virtual reality equipment to visualize the images and interact with them in 3D, a set of interactive features to annotate images, to create links between them and to build a 3D hypermedia. We present an experimental study and an application of our tool on faces skin.

Keywords: Stereoscopic acquisition, camera calibration, genetic algorithms, 3D visualization, image annotation, hypermedia, skin relief.

1 Introduction

Relief is a complex and important data for many domains. In medicine, numerous methods have been developed in order to acquire relief of various parts of the human body with the aim of discovering information and knowledge. In this paper we are especially interested with the acquisition of a surface, like the skin. We have conceived a complete and operational system (see an overview in figure 1) which is compound of three main modules: (1) an acquisition module that takes stereoscopic photographs of people with skin problems or other specific pathologies, (2) a camera calibration module that estimates the cameras parameters which are necessary for computing 3D

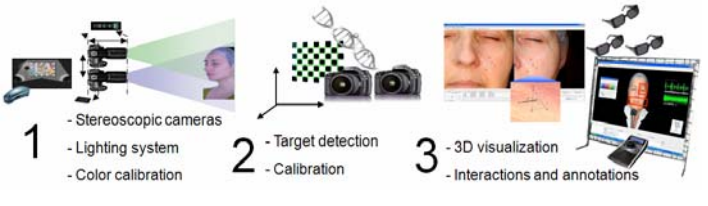


Fig. 1. Overview of our system

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 156–160, 2009.

information, and (3) a visualization and exploration module which can be used by dermatologists for instance to perform 3D measurements, to create annotations as well as a 3D hypermedia, and to share the extracted knowledge with others.

In this paper, we respectively detail each module in sections 2, 3 and 4, and we present our motivations and the state of the art for each of them. In section 5 we described the obtained results on the precision of camera calibration and a first example of an annotated 3D hypermedia build on 3D photographs of faces.

2 Acquisition and Calibration Module

In the skin domain, two types of relief acquisition methods can be distinguished, the so-called active and passive methods. Active methods consist in combining an optical sensor with a source of light, like for example Laser scanners, sensors that use structured lights or profilometry. Passives methods rather use one or more images like in [8] or like stereophotogrammetry [7]. For this module of our tool, we have conceived an acquisition system on the basis of two cameras assembled together and which are triggered in a synchronized way. We have designed a specific lighting system and we have used an optical sensor to calibrate all graphic devices (cameras, screens, video projectors, etc).

Camera calibration determines the accuracy of the acquired relief. It consists in estimating the intrinsic and extrinsic parameters of the cameras. Numerous methods exist in this context [9] without a real consensus, even if some algorithms are relatively common [11]. The types of methods we have selected consist in taking pictures of a calibration target with known dimensions, and then to estimate the parameters that minimize a target « reconstruction » error. We have developed a new calibration method, based on genetic algorithms (GA) [2][10], and which distinguishes itself from the others on the following points: it is specific to stereovision; it uses the notion of distance between points in its evaluation function.

The input data of our method is a set of known distances d_1, …, d_n between points detected in one or several couples of stereo images, and the pixel coordinates of these points on each CCD. We also provide initial bounds for the parameters to be estimated, but those bounds do not need to be precise at all. The genetic algorithm tries to find the set of parameters which minimizes the difference between the real and estimated distances.

3 Visualization and Interactive Exploration

3.1 Virtual Reality and 3D Measurements

Using virtual reality is necessary in order to visualize the relief in stereoscopy but also to let the expert navigate in the 3D images and, for instance, make annotations. For the stereoscopic visualization of images, we have used two types of projection hardware: on the one hand, standard cathodic screens that alternatively visualize the left and right images using active shuttering glasses, and on the other hand, two video projectors with passive glasses. These projectors have allowed us to project skin 3D images on a $25m^2$ screen in front of more than 20 people.

One may compute the 3D coordinates of a point P on the skin thanks to the parameters estimated by the calibration module of our system. Let Pl denote the projection of P in the left image, and let us suppose that this projected point was selected by the user. In order to compute the 3D coordinates of P, one has to find the point Pr, i.e. the correspondent of Pl in the right image. This is performed using a pattern matching algorithm that tries to maximize the correlation between Pl and Pr.

This correlation is computed using the color values of pixels on two small images centered respectively on Pl and Pr [6]. Then, using Pl and Pr, the 3D coordinates of P are known. The expert may thus measure 3D distances between two selected points. In order to measure depths (or heights), the expert selects 3 points in the image. These 3 points represents a plane, and the distance between this plane and a fourth point can be computed, which results in a height or depth measurement.

3.2 Annotations of Stereoscopic Images with 3D Stereoscopic Pointer

During the exploration of images, the expert may select some regions of interest in order to annotate them. Image annotation is currently the object of many researches, especially for automatic methods. For interactive or manual methods, one may cite for instance the work on VirtualLab [1] where microscopic pictures can be annotated, or [5] where images can be annotated using a web interface. Our system includes the interactive tools necessary to associate textual or voice annotations to selected areas (see figure 2). For this purpose, the user selects a specific area (wrinkle, specific symptoms, etc.) with a 3D pointer and may define for this annotation, a title, a text and a recording of his voice. Furthermore, annotations have specific parameters: a name, a color, a shape, and specific pointer events can be associated to it.

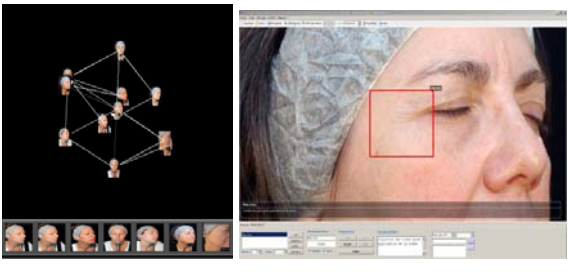


Fig. 2. On the left: Exploring the 3D hypermedia. On the right: Example of an annotation.

3.3 Interactive Tour and 3D Hypermedia

The author of annotations may generate in an intuitive way an interactive guided tour of the 3D picture. For this purpose, he may determine an ordering of the annotations and the corresponding selected areas. Our tool may then automatically scan these annotations in the specified order, with an adjusted zoom, and with playing the recorded voice. The expert's annotations are turned into an interactive movie. In this way, the expert may underline some facts and present them first, and then he may

explain their consequences. Several image databases exist in dermatology like DermAtlas [3]. Our system manages 3D images, and allows the expert to define links between annotations. Each annotation may thus point to several others, either in the same image or in other images. These links allow the expert to create a graph of relations between images (see figure 2).

4 Results

4.1 Experiments with Camera Calibration

We have compared our calibration method with a camera calibration toolbox implemented in MatLab [6]. 6 pictures of calibration target have been taken with different target orientation. In order to compare the two approaches, we have used a cross validation technique: each picture is isolated in turn and is used as an unseen test case, while the 5 others are used for learning the parameters. Both methods are evaluated with the same set of detected points, and with the same error measure (difference between real and estimated distances). The results are thus very encouraging and we have planned additional comparative tests.

Table 1. Evaluation of calibration accuracy using a cross validation technique over 6 pictures (48 points and 82 distances per picture). In underline and italic are presented the results of MatLab's « Camera Calibration ToolBox », and in bold the results of our tool.

Images	img1	img2	img3	img4	img5	img6
Mean Error	*142 µm*	*392 µm*	*270 µm*	*943 µm*	*330 µm*	*241 µm*
	45 µm	**53 µm**	**73 µm**	**41 µm**	**42 µm**	**55 µm**
Std deviation	*86 µm*	*72 µm*	*162 µm*	*83 µm*	*83 µm*	*120 µm*
	36 µm	**40 µm**	**53 µm**	**35 µm**	**35 µm**	**48 µm**
Max Error	*387 µm*	*559 µm*	*570 µm*	*1147 µm*	*546 µm*	*579 µm*
	175 µm	**192 µm**	**265 µm**	**165 µm**	**141 µm**	**225 µm**

4.2 Real Study

In order to evaluate our system in a real world application, we have conducted a study involving 18 women from 20 to 65 years old who presented skin specificities. For each woman, we have taken 3D pictures of their face (front and both sides). For some women who presented specific symptoms, we have also taken picture of their hands and of their back. In order to analyze the pictures, we have presented them to a panel of international dermatologists. They have used our tool to visualize the pictures in stereoscopy, to perform 3D measurements and to annotate the pictures. They have defined a guided tour. The possibilities offered by our tool (like 3D visualization and annotations) have improved the diagnostic of different skin symptoms by making the identification of specific information easier than in standard photographs.

5 Conclusions

We have developed a complete system for the acquisition, visualization and interactive exploration of stereoscopic pictures in the domain of dermatology. We have described its 3 main modules. We have defined a new calibration method which, after a first experimental comparison, seems to be efficient and well adapted to our application. We have proposed the use of specific virtual reality hardware in order to visualize stereo images and to navigate through them. We have developed several ways to perform 3D measurements, annotations and to share the discovered knowledge.

References

1. Alfonso, B.: Science Magazine, a publication by the AAAS vol. 308 featured Virtual Lab in their NetWatch section (2005)
2. Baeck, T., Hoffmeister, F., Schwefel, H.-P.: A Survey of Evolution Strategies. In: Proc. Fourth Int. Conf. Genetic Algorithms, pp. 2–9. Morgan Kaufmann, San Francisco (1991)
3. Bernard, A., Cohen, M., Christoph, U., Lehmann, M.D.: DermAtlas, Johns Hopkins University (2008), http://www.dermatlas.org
4. Bouguet, J.: Camera Calibration Toolbox for Matlab, http://www.vision.caltech.edu/bouguetj/calib_doc/
5. Chalam, K.V., Jain, P., Shah, V.A., Shah, G.Y.: Evaluation of web-based annotation of ophthalmic images for multicentric clinical trials. Indian Journal of Ophthalmology 54, 126–129 (2006)
6. Chambon, S., Crouzil, A.: Dense matching using correlation: new measures that are robust near occlusions. In: British Machine Vision Conference, BMVC 2003, vol. 1, pp. 143–152 (2003)
7. D'Apuzzo, N.: Modeling human faces with multiimage photogrammetry. In: Three-Dimensional Image Capture and Applications, vol. 4661, pp. 191–197 (2002)
8. Hernandez Esteban, C., Schmitt, F.: Silhouette and Stereo Fusion for 3D Object Modeling. Computer Vision and Image Understanding 96(3), 367–392 (2003)
9. Tsai, R.Y.: A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses. IEEE Journal of Robotics and Automation, 323–344 (1987)
10. Zhang, Y., Ji, Q.: Camera Calibration With Genetic Algorithms. In: IEEE International Conference on Robotics and Automation, pp. 2177–2182 (2001)
11. Zhang, Z.: A Flexible New Technique for Camera Calibration. Technical Report MSR-TR, Microsoft Research, pp. 98–71 (1998)

Implementing a Clinical Decision Support System for Glucose Control for the Intensive Cardiac Care

Rogier Barendse, Jonathan Lipton, Maarten van Ettinger, Stefan Nelwan, and Niek van der Putten

Erasmus MC, 's-Gravendijkwal 230, 3015 CE Rotterdam, The Netherlands
r.barendse@erasmusmc.nl

Abstract. Adherence to guidelines and protocols in clinical practice can be difficult to achieve. We describe the implementation of a Clinical Decision Support System (CDSS) for glucose control on the Intensive Cardiac Care Unit (ICCU) of the Erasmus MC. An existing paper protocol for glucose control was used for the CDSS rule set. In the first phase we implemented a proof of concept of a CDSS: a web 2.0 AJAX-driven web screen, which resulted in an improved adherence to the glucose guideline. This paper will reflect on the technical implementations and challenges of our experience with this process. The end product will allow: storage of guidelines in a shareable and uniform matter, presentation of guidelines in a more clear way to physicians, a more flexible platform to maintain guidelines, the ability to adjust guidelines to incorporate changes based on collected evidence from the CDSS and/or literature review, and be able to better review the outcome.

Keywords: Glucose management, CDSS, ICCU, CCU, cardiology, nurse-driven guideline, web 2.0, guideline implementation.

1 Introduction

The use and effects of CDSS systems in clinical practice have been studied extensively and have shown to be an effective mean to improve healthcare [1, 2]. At the Thoraxcentre of the Erasmus MC we have started to implement CDSS by automating the glucose protocol of the ICCU. Glucose regulation is difficult to achieve and may have significant implications for clinical outcome [3]. Though the clinical problem is complex, the nature of the paper protocol was very straightforward and therefore a good starting point.

The ICCU of the Thoraxcentre treats cardiology patients who require intensive care. These patients have continuous monitoring of vital signs which are registered, along with other clinical data in a Patient Data Management System (PDMS), Innovian [4].

1.1 Paper Protocol

A simple, rule based, sliding scale glucose protocol was used and was available at each patient bedside. The protocol was nurse-driven and dependent on glucose measurements determined by the laboratory. Compliance was low regarding advised insulin

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 161–165, 2009.

dosage and timing of measurements: there was a lack of notification when new lab results were available and there was no reminder on when to re-determine glucose values. These factors were given as the main reasons for not adhering to the protocol.

The paper protocol uses the most recent glucose measurement to advise an action of starting, adjusting or stopping insulin pump, and advises to measure glucose again within a certain amount of time.

The lab results are sent to the patient monitor, the PDMS and the Electronic Patient Record (EPR). A retrospective study of the data in the PDMS system revealed low compliance the protocol [5].

The protocol rules could not be defined as a gold standard: users suggested that the protocol could be improved with regard to certain points.

2 Methods

To achieve higher protocol compliance we decided to implant a CDSS that would resolve some of the previously mentioned problems. We deployed a medical touch screen computer at the nurse desk which displays the 8 beds of the ICCU with patient characteristics, previous glucose measurements and insulin pump settings (Figure 1). When a new glucose measurement for a patient arrives, a popup appears on the "bed" of the corresponding patient. The popup displays the glucose value, time of measurement, generated advice regarding insulin treatment and advised time for the next glucose measurement.

Fig 1 shows the Glucose Screen. This is a web 2.0 Ajax-driven web interface that polls the glucose web service every 10 seconds using SOAP. The web service

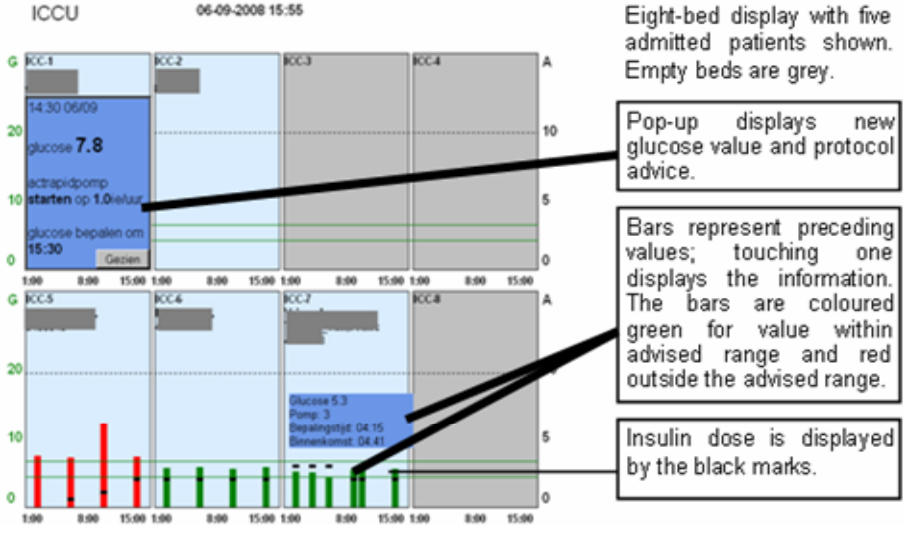


Fig. 1. Screen shot of the Glucose Screen with explanation

component runs on a web server and caches the lab values, the insulin pump settings and the generated decision of each lab value, every minute.

The database runs on SQL Server 2000 and is a real-time replicated database of the PDMS database. The database has extra tables for the glucose lab values, the generated advice and audit information. Figure 2 shows the dataflow of the application.

The guideline engine consists of an if-else structure, hard coded into the web service. The values needed for calculation of the generated advice are entered into the decision tree and a corresponding advice is returned.

We collected the data, the glucose value, the time of measurement, time of display and the time of reaction into this database.

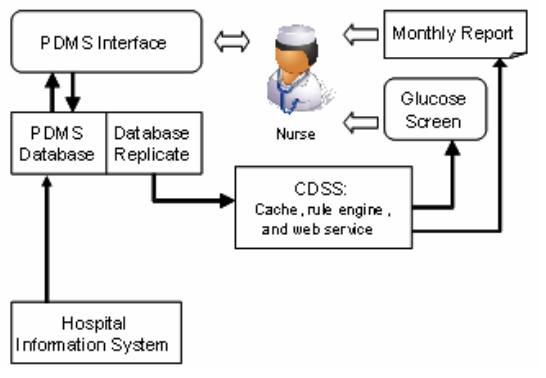


Fig. 2. Dataflow schema of the Glucose Screen

3 Results

In our setup the nurse no longer is required to actively look in the PDMS or EPR system to retrieve the latest measurement. The nurse can now easily discover new measurements and the generated advice by glancing at a fixed screen at the nursing station of the ward.

After implementation of the CDSS adherence to the glucose protocol increased when compared to baseline[5]. During a 4 month period we collected 3418 glucose measurements. Retrospectively we analyzed 15360 glucose measurements from the same ICCU from 18 months before the implementation of the CDSS. Patients that had less than 2 glucose measurements were not included in the analyses.

The percentage of glucose measurements performed on time (next measurement not later than the advised time + 10%) increased after implementation from 41% to 55%, an increase of 13.2% (95%CI 11.4% to 15.1% P<0.001). Compliance with advised insulin dosage also improved from 48% to 58%, increase of 9.8% (95% CI 7.9% to 11.6% P<0.001).

4 Future Work

One of the challenges in generating this application was retrieving the necessary data. Several sources, such as the hospital information system (HIS), the EPR built on the

HIS and the PDMS provide the necessary data elements. The PDMS in itself receives data from the HIS (lab and patient demographics). Getting the necessary data from 3rd party applications can be challenging.

Currently we are extending the project with a third-party commercial decision support tool Gaston [6]. The tool consists of a guideline executer and an interface to visually design guidelines. Also it has built-in support for data acquisition and several other features. Figure 3 shows the guideline editor. In this program physicians can specify the guidelines themselves. These guidelines are immediately available from a web service when published. This gives us a clear distinction between guidelines and corresponding advice and the display of these guidelines on the screen.

With this extension we can focus our research more on implementation of CDSS and on how we can deliver the generated guidelines to the nurse or physician in the most efficient way possible. We want to extend the current application with this rule engine in our webservice. In a later phase we plan to implement a framework for transporting guidelines to other screens, applications and devices.

Many aspects of the implementation would be facilitated by an improved data integration of the different products and/or systems. A data warehouse solution would not work in the current setup, since the extraction would only be daily at most and not continuously. At the moment we are implementing HL7 to receive the lab data to be less dependent on the lab data in the PDMS.

The new improvements, Gaston and HL7 lab will facilitate and speed up the implementation of new guidelines in a faster and more flexible fashion.

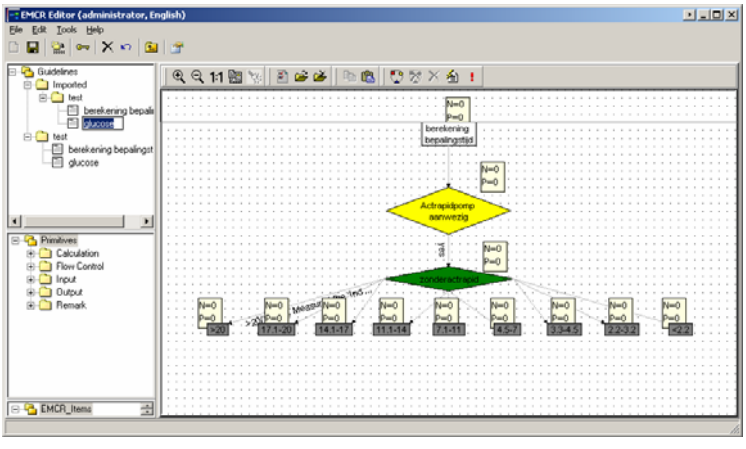


Fig. 3. A screenshot of the KA-tool of Gaston with Glucose Decision Tree

5 Discussion

We would like to expand CDSS into our organization. This will consist of working with 3-party software vendors that are capable of integrating CDSS into their application. Also we want to be able to extend CDSS to other platforms at the point of care e.g. PDA's.

Validating the outcome of our research is challenging as it is an iterative process with many different alterations: we have been upgrading software periodically on one side and also been improving the guideline on the other side. Each change has been documented and data has been collected until each point of the update. We chose to use different outcome measures for evaluating technical aspects, protocol compliance and clinical outcomes to be able to investigate the effect of each of these changes we made.

When interpreting the results it is important to consider that it is possible that changes to the guideline may result in increased adherence, but not always in improved clinical measures, and that technical improvements may lead to improved outcomes as well, irrespective of guideline adherence (e.g. a better graphical display of certain laboratory values may lead to earlier detection of abnormal values). Finally one must always be on the lookout for 'bugs' (both technical as inconsistencies in guidelines) that can adversely affect patient care.

References

1. Garg, A.X., Adhikari, N.K.J., McDonald, H., et al.: Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review. JAMA 293(10), 1223–1238 (2005)
2. Kaplan, B.: Evaluating informatics applications–clinical decision support systems literature review. International Journal of Medical Informatics 64(1), 15–37 (2001)
3. Weston, C., Walker, L., Birkhead, J.: National Audit of Myocardial Infarction Project NIFCOR. Early impact of insulin treatment on mortality for hyperglycaemic patients without known diabetes who present with an acute coronary syndrome. Heart 93(12), 1542–1546 (2007)
4. Nelwan, S., van Dam, T., Meij, S., van der Putten, N.: Implementation and use of a patient data management system in the intensive care unit: A two-year experience. Computers in Cardiology 2007, 221–224 (2007)
5. Lipton, J., Barendse, R., Eenkhoorn, E., et al.: Glucose Control as a Model for Implementation of a Clinical Decision Support System. In: CIC Proceedings, Bolonga, vol. 35, pp. 661–664 (2008)
6. de Clercq, P.A., Hasman, A., Blom, J.A., Korsten, H.H.M.: Design and implementation of a framework to support the development of clinical guidelines. International Journal of Medical Informatics 64(2-3), 285–318 (2001)

Steps on the Road to Clinical Application of Decision Support – Example TREAT

Steen Andreassen[1], Alina Zalounina[1], Knud Buus Pedersen[2], John Gade[2], Mical Paul[3], and Leonard Leibovici[3]

[1] Center for Model-Based Medical Decision Support, Aalborg University, Denmark
{sa,az}@hst.aau.dk
[2] Judex Datasystemer AS, Aalborg, Denmark
[3] Department of Medicine E, Rabin Medical Center, Beilinson Hospital, Petah-Tiqva, Israel

Abstract. The decision support system TREAT advices on antibiotic treatment of severe infections. A multicenter randomized clinical trial has demonstrated that Treat reduces inappropriate treatment by 50%. This paper will show that TREAT satisfies several features closely correlated with decision support systems's ability to improve clinical practice. Examples of such criteria are: providing recommendations, not just assessments; transparent line of reasoning; convenience in use. Additional design features, such as transferability and addressing an important clinical problem, will also be discussed.

1 Introduction

Medical Decision Support System (MDSS) has been slow to fulfill expectations in terms of actual improvements of diagnosis and therapy, making some workers feel like "waiting for Godot" [1]. A systematic review [2] identified trials of one hundred MDSS between 1973 and 2004. Both the number of trials and the methodological rigor increased over the years, but disappointingly only 7% of the trials demonstrated any improvement of patient outcome.

In another critical review Kawamoto et al. [3] identified several features of MDSS that statistically predicted that a given system was capable of improving clinical practice. Some of those are related to the **Style** of the MDSS, while others are related to the **Convenience** of use of the MDSS. Based on recent experience with an MDSS called Treat for advising on antibiotic therapy of severe infections, we will suggest two additional criteria, labeled **A Clinical Problem** and **Transferability**, mainly based on economic considerations. In addition we would like to emphasize our belief that randomized controlled **Clinical Trials** are required for clinical acceptance of the as yet largely unproven MDSS technology. We hypothesize that systems that fulfill these criteria in general and Treat in particular will terminate the period in which we have been "waiting for Godot".

Our 5 criteria for a successful MDSS are:

1) **A Clinical Problem.** The MDSS should address a clinical area that satisfies 2 criteria: the area has an unsolved clinical problem; the unsolved problem is big in

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 166–170, 2009.

terms of number of patients and impact on patients. These features ensure that there is a large health economic benefit associated with the use of the MDSS.

2) **Clinical Trials.** The MDSS should solve this problem, and patient benefit should be proved with the same rigor as required for other new medical methods, i.e. through randomized controlled clinical trials.
3) **Style.** The MDSS should be computer-based [3]. It should provide recommendations, not just assessments [3], and furthermore the line of reasoning should be transparent to the user [3]. It should also provide periodic performance feedback. [3]
4) **Convenience.** The logistics should be simple and integrate into the clinical workflow [3], support should be provided at the time and location of the clinical decision [3] and impose little or no extra work, in particular if "annoying", such as double entry of patient data.
5) **Transferability.** The MMDS should be transferable between different environments. Otherwise the substantial intellectual and economical investment required to build an MDSS will not be justifiable.

The following sections will illustrate these criteria by describing them in relation to TREAT.

2 Infectious Diseases as an Example

2.1 A Clinical Problem

As it turns out, infectious diseases actually present two important problems, a short term and a long term problem. The short term problem has to do with the current clinical prescription of antibiotics in hospitals for moderate to severe infections. *"Studies have shown that about half the time physicians are not prescribing antibiotics properly"* [4]. Infections are the 6th leading cause of death even in high-income countries, and the first or second cause of preventable deaths [5]. Given these numbers, in the EU an estimated number of 850.000 deaths annually are due to infection.

There is an equally important long term problem: We witness the emergence of pathogens resistant to almost all useful antibiotics, endangering the life of patients. Methods that would enable clinicians to reduce inappropriate use of antibiotics, in particular broad spectrum antibiotics would slow down this threatening development.

2.2 Clinical Trials

In this section we shall examine to which extent TREAT provides a solution to the short and long term problems in infectious diseases.

The short term problem is essentially the high rate of inappropriate antibiotic treatments. TREAT has been tested on over 5000 patients in different hospital settings (Table 1). The clinical trials show an average reduction in inappropriate treatment of 50%, for TREAT relative to clinical practice. In the interventional study in Germany, Italy and Israel [6] the improved treatment resulted in a significant reduction of 0.6 bed-days for each patient where the advice from TREAT was made available to the

clinicians, which represents a substantial clinical and economic incentive for using TREAT. TREAT also reduced total cost of treatment by 50%.

TREAT can also contribute towards the long term problem, the steady rise of resistance to antibiotics. The large reduction of inappropriate treatment was mentioned above, but in addition the observational study in Germany, Italy and Israel show that TREAT reduced the ecological impact of antibiotics on the development of antibiotics by advising significantly less broad-spectrum drugs.

Table 1. Clinical trials of TREAT

Type of trial	Period	Catchment area, hospital settings	No. of patients	Inappropriate antibiotics Doctor. vs. TREAT		Relative reduction of inappropriate by TREAT
Observational [7]	2005-2006	Copenhagen, Denmark	161	34%	14%	59%
Interventional, randomized, controlled [6]	2004	Germany, Italy and Israel	2326	36%	27%	25%
Prospective cohort, observational [6]	2002-2003	Germany, Italy and Israel	1203	43%	30%	30%
Observational [8]	1995-1996	North Jutland County, Denmark	1597	39%	5%	87%
Total/average			5287	38%	19%	50%

2.3 Style

Treat is computer–based [3] and provides answers to the following questions, which all clinicians considering antibiotic treatment must attempt to answer: Is there a bacterial infection? How severe is it? What is the most likely site of infection? Which pathogens are responsible? What are the benefits and costs of potential antibiotic regimens? The answer to the last question takes the form of a ranked list of recommendation for antibiotic therapy [3].

The output display in TREAT allows the clinician to follow the steps in TREAT's reasoning, all the way to the recommendation of antibiotics, making it possible to disagree at every step and if required to choose a treatment different from TREAT's recommendation. TREAT therefore satisfies the requirements of transparency of reasoning [3] and allows the clinician the freedom of easily choosing to deviate from the advice.

The TREAT package includes a program that provides periodic feedback on performance of the system [3].

2.4 Convenience

No decision support system is likely to be successful, irrespective of its clinical merit, if the operation of the system is not convenient. In the case of TREAT this requirement mainly translates into availability of the system at appropriate times in

the workflow [3] and into minimization of data entry, the design goal being that operating of TREAT should take no more than 5 minutes.

The question of availability has been addressed by implementing TREAT as a web-service, available on every computer with a browser (e.g. Microsoft Internet Explorer) installed. Therefore TREAT is available whenever patient data is entered into an EPR system, or whenever the clinician is accessing the hospital's lab-information system for test results.

The question of minimization of data entry has been addressed by allowing TREAT to access the hospital's data repositories, primarily the Hospital Information System for the patient's demographic data, along with clinical chemistry and microbiology. TREAT automatically searches for patient data relevant for infectious diseases and makes this data available to the clinician through TREAT's user interface. Since the TREAT requires a considerable amount of clinical data, this strategy is a key element in making TREAT user friendly.

2.5 Transferability

The development and testing of TREAT to its present state has taken a little over 50 man-years. This investment would be difficult to justify, if application of TREAT was only envisioned at a single hospital. Obstacles to transferability between hospitals fall in three categories, which we shall call integration, adaption and calibration.

Integration involves building TREAT's integration engine. Adaption involves a range of tasks, ranging from conversion of units for lab results to listing antibiotics available at the hospital pharmacy. Calibration involves adjusting probabilities for factors which affect the performance of the system, if not adjusted to local conditions (e.g., the local resistance of pathogens to antibiotics) [9]. Integration, adaption and calibration require several man-months.

In the case of TREAT benefits of the system are sufficient to pay for the costs associated with transferability. We can estimate from the clinical trials [6] that the combined annual costs of operating TREAT, including costs of transferability and recalibration will be repaid 2 times per year due to savings on antibiotics and 22 times per year due to savings on bed-days. On top of this, TREAT could save a considerable number of lives due to the improved treatment, about 16/100.000 per year [6].

3 Discussion

In the sections above we have argued that TREAT satisfies the five criteria listed in the introduction. These criteria are based on the large body of scientific evidence [3], but also include extra features. Additionally, a study on a MDSS can be evaluated on a 10-point scale for methodological quality [2]. Such evaluation of TREAT gave presumably the highest score (10).

While we believe that these criteria are necessary conditions for widespread clinical application of an MDSS, they are not necessarily sufficient conditions. We might add two more conditions which we could call Motivation and Availability.

Motivation is related to the different sensitivities different decision makers may have to the arguments for introduction of an MDSS. Annual costs of running and

maintaining an MDSS can be beyond the budget of any single department. The repayment of this cost in the form of saved bed-days may not represent a saving from the point of view of the hospital; it may simply represent a loss of revenue.

Availability is related to the commercial availability of an MDSS, which is a condition for widespread clinical application. Universities are not ideally suited for promoting and maintaining a large and complicated system. This usually requires a company, but such a company must be willing to make major investments.

Given this set of conditions for an MDSS it is not so surprising that MDSSs with widespread clinical acceptance are few and far between – and whether TREAT will make it all the way still remains to be seen.

References

1. Wears, R., Berg, M.: Computer Technology and Clinical Work: Still Waiting. JAMA 293(10), 1261–1263 (2005)
2. Garg, A., Adhikari, N., McDonald, H., Rosas-Arellano, M., Devereaux, P., Beyene, J., Sam, J., Haynes, R.: Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 293(10), 1223–1238 (2005)
3. Kawamoto, K., Houlihan, C., Balas, E., Lobach, D.: Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 330(7494), 765 (2005)
4. Davey, P., et al.: Review: Interventions to improve antibiotic prescribing practices for hospital inpatients. Cochrane Database of Systematic Reviews 1 (2008)
5. Kung, H., et al.: Deaths: final data for 2005. Natl. Vital Stat. Rep. 56, 1–119 (2008)
6. The TREAT Working Group: Improving empirical antibiotic treatment using TREAT, a computerised decision support system: a cluster randomised trial. J. Antimicrob. Chemother 58, 1238–1245 (2006)
7. Kofoed, K., Zalounina, A., Andersen, O., Lisby, G., Paul, M., Leibovici, L., Andreassen, S.: Performance of the TREAT decision support system in an environment with a low prevalence of resistant pathogens. J. Antimicrob. Chemother. 63(2), 400–404 (2009)
8. Kristensen, B.: Construction and evaluation of a decision support system for empirical antibiotic treatment of bacteraemia. Aalborg Hospital. Ph.D-thesis (2003)
9. Zalounina, A., Andreassen, S., Leibovici, L., Paul, M.: Transferability modelling in the TREAT decision support system. In: 17th World Congress of the International Federation of Automatic Control, pp. 809–8102. Elsevier, Amsterdam (2008)

Integrating Healthcare Knowledge Artifacts for Clinical Decision Support: Towards Semantic Web Based Healthcare Knowledge Morphing

Sajjad Hussain and Syed Sibte Raza Abidi

NICHE Research Group, Faculty of Computer Science, Dahousie University, Canada

Abstract. Healthcare decision making demands the systematic integration of knowledge from multiple sources, such as clinical guidelines, clinical pathways, knowledge of practitioners and so on. We present a semantic web based approach for synthesizing health knowledge through the semantic modeling of healthcare knowledge as ontologies and reasoning over the ontologies to derive a morphed knowledge object. We demonstrate the application of our approach by generating morphed knowledge about prostate cancer clinical pathways.

1 Introduction

Healthcare decision making during the diagnostic-treatment cycle is a complex activity. Health professionals make clinical decisions by applying healthcare knowledge that includes their experiential knowledge and explicit knowledge 'artifacts', such as clinical practice guidelines, best evidence, clinical pathways and so on [1]. One may note that each healthcare knowledge artifact captures specific conceptual, contextual and operational aspects of a disease and corresponding diagnostic/therapeutic procedures. Health professionals, guided by the patient's context, are able to select the relevant 'knowledge objects' from these different artifacts and then inter-relate these specific knowledge objects whilst satisfying clinical relevance and pragmatics constraints. For instance, a health professional generating a treatment plan for a patient with hypertension and diabetes will refer to the relevant sections of (a) clinical guidelines for recommendations; (b) clinical pathways for procedural protocols to exercise these recommendations; and (c) medical literature to determine the best evidence and outcomes of treatment options. In our work, we attempt to pursue a context-sensitive selection and integration of medical knowledge from multiple knowledge artifacts to generate a comprehensive knowledge object for clinical decision support.

We are developing the concept of healthcare knowledge morphing that entails "the intelligent and autonomous fusion/integration of contextually, conceptually and functionally related knowledge objects that may exist in different representation modalities and formalisms, in order to establish a comprehensive, multifaceted and networked view of all knowledge pertaining to a domain-specific problem"–Abidi 2005 [2]. In this paper, we present our healthcare knowledge

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 171–175, 2009.

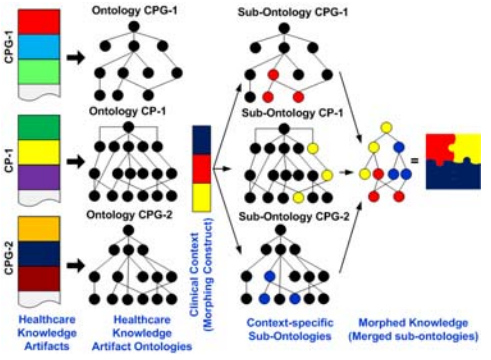


Fig. 1. Healthcare Knowledge Morphing

morphing framework $\mathcal{K}$-$\mathcal{MORPH}$ that is based on the semantic web approach that entails: (a) Developing *Knowledge Artifact Ontologies* (KAOs) to represent knowledge within CPG, CP and CM [1]; (b) Specifying the clinical context of the knowledge morphing activity through a rich *morphing construct*; (c) Generating the morphed knowledge by (i) selecting a *contextualized sub-ontology*, corresponding to the clinical context, from the KAO; and (ii) merging the selected contextualized sub-ontologies, using reasoning algorithms applied to a set of domain-specific *context-specific axioms*, to generate a new sub-ontology that represents the 'morphed' knowledge artifact. We demonstrate the working of our knowledge morphing framework $\mathcal{K}$-$\mathcal{MORPH}$ by morphing three different location-specific clinical pathways to generate a comprehensive knowledge about treatments and follow-ups for a clinical context *therapeutic decision support* (see Figure 1) [1].

2 Related Work

The literature suggests other approaches for knowledge integration problem from different perspectives. For instance, the ECOIN framework performs semantic reconciliation of independent data sources, under a defined context, by defining *conversion functions* between contexts as a network. ECOIN takes the *single ontology, multiple views* approach [3], and introduces the notion of *modifiers* to explicitly describe the multiple specializations/views of the concepts used in different data sources. It exploits the modifiers and conversion functions, to enable context mediation between data sources, and reconcile and integrate source schemas with respect to their conceptual specializations. Another recent initiative towards knowledge integration is the OpenKnowledge project [4] that supports the knowledge sharing among different knowledge artifacts, not by sharing their asserted statements, instead by sharing their *interaction models*. An interaction model provides a context in which knowledge can be transmitted between two (or more) knowledge sources (peers).

3 $\mathcal{K}$-$\mathcal{MORPH}$ Architecture

The $\mathcal{K}$-$\mathcal{MORPH}$ approach is shown in Figure 1, and its main elements are described in the following subsections. For further details see [5].

3.1 Knowledge Representation and Annotation via Ontologies

In $\mathcal{K}$-$\mathcal{MORPH}$, knowledge artifacts are represented using two different (but inter-related) ontologies, namely: (i) *Domain Ontology*; and (ii) *Knowledge Artifact Ontology* (KAO). A domain ontology serves two purposes: (i) standardization of the domain-specific concepts and relations defined in the knowledge artifact ontologies; and (ii) specification of abstract knowledge links between contextually and functionally congruent knowledge components in different KAOs. A knowledge artifact ontology (KAO) serves as a lower-level ontology that captures both the structure and content of a particular knowledge artifact–such as CPG, CP [1], clinical cases etc. As a test-case, we used three location-specific (Halifax, Calgary and Winnipeg) Prostate Cancer (PC) pathways as medical knowledge artifacts, and represented them in different KAOs [1].

3.2 Contextualizing Ontologies

Contextualizing an ontology deals with an adaptation of its ontology model to support a local view [6]. In $\mathcal{K}$-$\mathcal{MORPH}$, each KAO models the procedural knowledge of a knowledge artifact. But, by contextualizing a KAO we are able to provide a specialized view that models (i) a specific interpretation of its ontology concepts, and (ii) an implementation of its procedural knowledge applied in a particular context. A contextualized sub-ontology is extracted from a KAO based on the context-specific concepts, and comprises (i) instances (ii) sub-concepts, (iii) equivalent-concepts, (iv) properties, (v) property domain and range, and (vi) assertions for the context-specific concepts.

3.3 Morphing Constructs

In order to represent the context under which two or more knowledge artifacts can be morphed to solve a specific problem, we defined a *Morphing Construct*. The morphing construct supervises the knowledge morphing process, and provides a context for determining when, where and how two or more knowledge artifacts need to be reconciled. A Morphing construct is a tuple that declares a context-specific knowledge component and its role under a defined context.

3.4 Morphing Engine

The $\mathcal{K}$-$\mathcal{MORPH}$ morphing engine inputs the problem-context, ontology-encoded knowledge artifacts (OKAs), domain ontology, and morphing constructs. It first employs the problem-context to determine the context-specific knowledge components (i.e. contextualized sub-ontologies) from different KAOs. Based on the declarative knowledge of morphing constructs, it identifies correspondences

between the ontology-entities (concepts, properties, and individuals) of different contextualized sub-ontologies. Based on the identified correspondences, the morphing engine employs the ontology reconciliation process that (i) aligns and then merges contextualized sub-ontologies; (ii) identifies and resolves logical inconsistencies, if present; and (iii) generates a morphed ontology, and unresolved inconsistencies in it.

4 $\mathcal{K}$-$\mathcal{MORPH}$ in Action: Morphing Clinical Pathways

We tested the above mentioned processes in $\mathcal{K}$-$\mathcal{MORPH}$ using our *PC Test-case*. The test-case involves (i) three medical knowledge artifacts, describing Prostate Cancer (PC) clinical pathways for three different locations (Halifax, Calgary, and Winnipeg); (ii) a problem-context; and (iii) the morphing constructs. The morphing process for the PC Test-case follows the following steps:

Step # 1: *Knowledge Representation and Annotation of PC Artifacts*: The knowledge of three PC pathway artifacts are encoded into three different KAOs. Each pathway deals with four major types of tasks, namely (a) *Consultation Task*; (b) *Non-consulation Task*; (c) *Referal Task*; and (d) *Followup Task*, represented as concepts/classes in each KAO. Such tasks are supported (via properties) by other concepts such as *Clinician*, *Decision Criteria*, *Frequency*, *Interval Duration*, *Investigation*, *Patient Condition Severity*, *Test Result*, *Followup*, and *Treatment*.

Step # 2: *Defining a Problem-context*: We defined a problem-context *therapeutic decision support* whereby the user is needs to morph all three PC pathways in terms of: (i) the treatments, (ii) their durations, (iii) their follow-ups, (iv) their care-settings and (v) the practitioners involved for them. The given problem-context represents context-specific interpretations, such as (a) Calgary and Halifax both share PC Clinicians; and (b) Treatments in the PC Calgary pathway can be treated as Followups in the PC Winnipeg pathway.

Step # 3: *Identifying Contextualized Sub-ontologies*: Given the *therapeutic decision support* context, morphing constructs and three PC pathway ontologies, three contextualized sub-ontologies were generated. Each contextualized sub-ontology was semantically validated for conceptual consistency and completeness.

Step # 4: *Context-driven Ontology Reconciliation of Sub-ontologies*: The $\mathcal{K}$-$\mathcal{MORPH}$ morphing engine initiated an ontology reconciliation process on the contextualized sub-ontologies, and as a result alignments were found between the classes *Treatment*, *Followup*, *Frequency*, *Interval Duration*, and *Clinician*. The morphing engine processes these alignments using context-axioms and PC domain-axioms to generate potential 'knowledge-links' between aligned PC treatments.

Results: Figure 2 shows the morphed knowledge for the treatment `PC-Halifax: ActiveSurveillance`. The morphed knowledge has determined that `PC-Halifax: ActiveSurveillance` can now be treated by a *Primary Urologist*. Based on the reconciliation of the concepts *Clinician*, *Treatment*, *Followup* and *Interval* a knowledge-base was generated in terms of a contextualized sub-ontology.

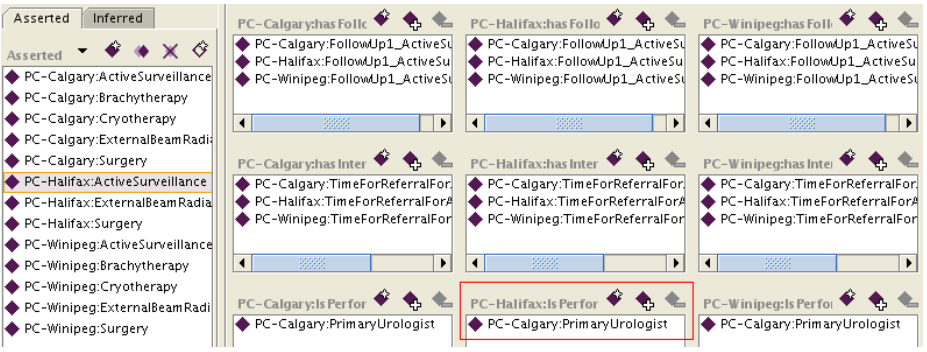


Fig. 2. PC Test-case: Morphed knowledge for PC-Halifax:ActiveSurveillance

5 Concluding Remarks

Clinical decision support needs a knowledge-base that can be designed from scratch or created by synthesizing existing healthcare knowledge existing in different modalities. In this paper, we presented our knowledge morphing approach that allows the systematic synthesis of multiple knowledge artifacts to develop a comprehensive knowledge-base that can be used by decision support systems. Our knowledge morphing approach ensures the semantic correctness of the morphed knowledge to the extent that it is comparable to a knowledge-base created through a knowledge engineering exercise. We showed how our approach is used to develop a unified prostate cancer clinical pathway by synthesizing three different clinical pathways.

This research is funded by a grant from Agfa Healthcare (Canada).

References

1. Abidi, S., Abidi, R., Hussain, S., Butler, L.: Ontology-based modeling and merging of institution-specific prostate cancer clinical pathways. In: Knowledge Management for Healthcare Processes Workshop at ECAI 2008, Patras, Greece, July 21-25 (2008)
2. Abidi, S.S.R.: Medical knowledge morphing: Towards the integration of medical knowledge resources. In: Computer-Based Medical Systems, June 23-24 (2005)
3. Firat, A., Madnick, S., Grosof, B.: Contextual alignment of ontologies in the eCOIN semantic interoperability framework. Inf. Tech. and Management 8(1), 47–63 (2007)
4. Robertson, D., et al.: Open knowledge: Semantic webs through peer-to peer interaction. Technical Report DIT-06-034, University of Trento, Povo, Italy (May 2006)
5. Hussain, S., Abidi, S.S.R.: K-MORPH: A semantic web based knowledge representation and context-driven morphing framework. In: Workshop on Context and Ontologies, at ECAI 2008, Patras, Greece, July 21-25 (2008)
6. Segev, A., Gal, A.: Putting things in context: A topological approach to mapping contexts to ontologies. In: Spaccapietra, S., Atzeni, P., Fages, F., Hacid, M.-S., Kifer, M., Mylopoulos, J., Pernici, B., Shvaiko, P., Trujillo, J., Zaihrayeu, I. (eds.) Journal on Data Semantics IX. LNCS, vol. 4601, pp. 113–140. Springer, Heidelberg (2007)

A Knowledge-Based System to Support Emergency Medical Services for Disabled Patients

Luca Chittaro[1], Roberto Ranon[1], Elio Carchietti[2], Agostino Zampa[3], Emanuele Biasutti[3], Luca De Marco[1], and Augusto Senerchia[1]

[1] Human-Computer Interaction Lab,
University of Udine, via delle Scienze. 206, 33100 Udine, Italy
[2] 118 Regional Emergency Medical Service,
Udine Hospital, 33100 Udine, Italy
[3] Department of Rehabilitation Medicine,
Physical Medicine and Rehabilitation Institute "Gervasutta", 33100 Udine, Italy

Abstract. This paper illustrates a knowledge based system devoted to help nurses and volunteers of Emergency Medical Services (EMS) in dealing with disabled patients during an emergency.

Keywords: emergency medical services, decision support systems, knowledge-based systems, disabled patients, Web-based systems, mobile devices.

1 Introduction

Being able to promptly and accurately choose a proper course of action in the field is a crucial aspect of emergency response. To guarantee that, emergency medical services (EMS) heavily rely on predefined procedures, on which first responders are specifically trained. The procedures are necessarily thought for the most frequently occurring cases. As a result, they may not be optimal and require additional knowledge for special cases, such as the various types of disabilities. To the best of our knowledge, no research has been devoted to using knowledge based systems for helping EMS nurses and volunteers in dealing with disabled patients. This paper focuses on a system that provides recommendations in the field for such cases.

2 Requirement Analysis and Knowledge Acquisition

We started our project by conducting focus groups that involved: (i) EMS physicians and nurses, (ii) rehabilitation clinicians specialized in disabilities, and (iii) representatives of various italian associations of disabled persons[1]. We summarize in the following the main findings that emerged from the focus groups:

[1] Association of the Blind and Visually impaired (UICI), Association of the Deaf and Mute-Deaf (ENS), Autism association (PROGETTO AUTISMO FVG), Dystrophy Association (UILDM), Regional Council of the Disabled (CONSULTA FVG), Spilimbergo Center for the Motor Disabled (PROGETTO SPILIMBERGO).

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 176–180, 2009.

- Although knowing the general class of (sensory, motory, cognitive) disability to which the patient belongs already allows to provide some disability-related advice, for each class there are a large number of descriptive attributes (e.g., detailed anatomical descriptions of motor disabilities) that would allow the system to provide advice which is tailored to the single patient. Therefore, the system needs a detailed representation of the patient's disabilities that comprises all those attributes. From this point of view, our work shares some similarities with the problem of generating personalized information using medical records that has been explored in non-emergency domains [1].
- Since every second counts in EMS operations, it is not conceivable to acquire the detailed description of patient's disabilities during the emergency: the information is needed beforehand, also taking into account that determining the value of the different attributes can require considerable time to an experienced clinician.
- The disabled person and her family should be actively involved in the management of the information stored in the system: although some attributes can be provided only by doctors, allowing the disabled to access their full record and keep some personal fields up-to-date contributes to build trust in the system and make patients aware (for privacy and legal reasons) of the data stored about them and who can access it.
- The system should provide advice to the phone operators of the EMS center (to help them choose which team and which ambulance is most appropriate to the context) as well as to the EMS first responders on the field (to provide advice about the course of actions to take). For this reason, the system should run on desktop as well as mobile devices, and the mobile interface should take into account peculiar limitations of mobile data visualization [2].
- An important contextual factor to be taken into account is the severity of the emergency which is formalized by EMS with standard codes (e.g., the standard employed by all Italian EMS is based on 4 codes of increasing severity: white, green, yellow, red). As severity increases, the system should restrict the number of recommendations, focusing on those which are crucial to preserve life.
- The advice provided to different classes of medical first responders (physicians, nurses, volunteers) should not necessarily be the same.

After the analysis of requirements, the knowledge acquisition process has been organized to take advantage of three different kinds of knowledge sources:

- Available general documents about safety response concerning the disabled, produced by different organizations, e.g. the National Department of Firefighters in Italy and the Center for Development and Disability in the US [3]. The analysis of such documents allowed us to derive basic rules about how to communicate and behave with blind, deaf, mute people or people with mental disorders, and how to transport motor-impaired persons in emergency situations such as fires or underground train evacuations. This knowledge was not specific to EMS so it was reviewed with clinicians to adapt it to the EMS context, e.g. some recommendations were considered to be trivial for professional EMS personnel.
- Expert knowledge, provided by the medical authors of this paper (an EMS physician and two clinicians specialized in disabilities). Each expert analyzed the problem from a different perspective, the acquired rules were formulated in natural language in a draft document and we carried out periodical panel meetings involving all the

experts to review the individually proposed rules. These panel meetings helped point out and correct some differences in the terminology used by the different experts. Changes in rules were typically made to reconcile the clinical approach of thoroughly reasoning from very precise diagnoses with the EMS approach where priority is given to preserve life, stabilizing the patient and transporting him quickly and safely to the hospital. When the two approaches could not be reconciled, the rule was rejected: it would indeed be impossible on the field to carry out evaluations which require considerable time and are technically more appropriate for a hospital environment.

- Knowledge acquired from representatives of the associations of disabled persons. Semi-structured interviews were carried out to gather information about previous experiences (if any) as EMS patients or let them imagine (as a role-playing exercise) being rescued and think about which kind of first responders' actions should be avoided or should be undertaken to make the whole operation more acceptable and comfortable to them. This was especially useful to more thoroughly investigate communication-related and social aspects of the interaction between first responders and disabled patients (e.g., ways to appropriately get the attention of a deaf person, verbal expressions that should be avoided with blind persons,…). The acquired knowledge was always submitted to the previously mentioned panel meetings for final approval.

Table 1. Main GEM II elements and values for a guideline that applies to a motor disability

Identity	*GuidelineTitle*	First Aid of Motor Disabled Patients – Forearm impairment - Transportation
Developer	*DeveloperName*	Physical Medicine and Rehabilitation Institute "Gervasutta"
	Sponsor	118 Regional Emergency Medical Service, Udine Hospital
Purpose	*Main Focus*	Transport of Motor Disabled people
	Exception	none
	Objective	Prevent transport injuries and provide comfort
Intended Audience	*Users*	nurses, volunteers, physicians, relatives
	Care Setting	red, yellow, green, and white emergency codes
TargetPopulation	*InclusionCriterion*	Motor disabled people
Knowledge Components	*Conditional Recommendation*	IF (b710.s7103="Moderate Impairment" OR b710.s7103 = "Complete Impairment") THEN (avoid forced movement of b710)
	ActionType	Transportation
	Recommendation Strength	4

3 The Knowledge Based System

Identifying and representing all the impairments of each disabled patient to generate guidelines for EMS operations is a challenging task because severely disabled patients can be affected by many different and unrelated conditions which are not taken into

account by general disability stereotypes (e.g., blind, deaf, …). To represent patient's disabilities, we started from the ICF international standard of the World Health Organization (WHO) for measuring health and disability, and defined a specific Disabled User Profile (DUP) for the EMS context. The DUP is described in detail in [4].

The knowledge base has been represented using the GEM II document model [5]. For example, Table 1 shows a guideline that applies to motor disabilities where the impairment is located on the patient's forearm (b710 is the ICF code for Mobility of joint functions and s7301 is the ICF code for Structure of Forearm). The table shows only the GEM elements and values that are more important in this example. From the GEM II documents, we derive rules in the format executable by Drools (jboss.org/drools), a Rule Management System based on the well-known RETE algorithm. We included the knowledge base into a Web system we developed with the Jboss framework (jboss.org).

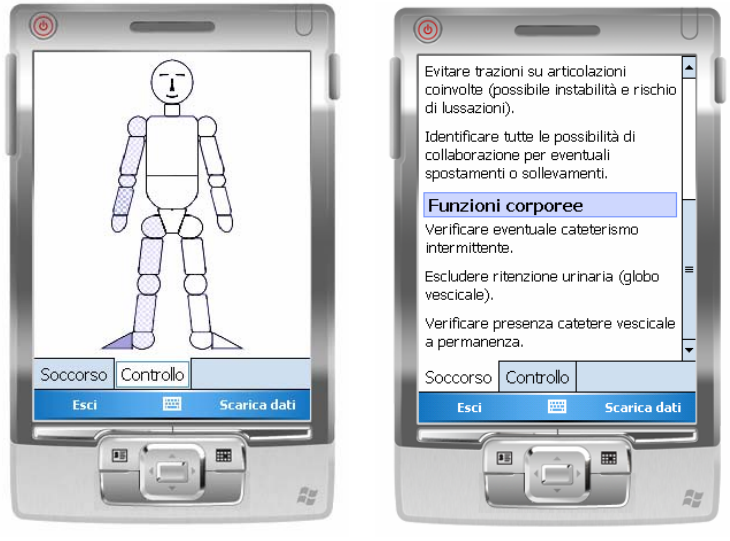


Fig. 1. The screen of the Web interface devoted to motor control representation

The first step in using the system is to fill the DUP for the considered person, through a Web interface. The personal data section of the DUP can be filled by the disabled person or her relatives. The medical sections have instead to be filled by the general practitioner or the specialists who follow the disabled person.

The second step concerns the emergency call. When the phone operator in the EMS center receives the call, the system first tries to match the calling number with the DUP database to automatically display the caller's personal data on the phone operator's screen. If caller's automatic identification fails, the system provides the phone operator with a quick search functionality to retrieve the right DUP from the typical information that is requested anyway during an emergency call. When the phone operator dispatches an ambulance to the emergency destination, she assigns that DUP to that ambulance run.

The third step concerns sending system recommendations to the first responders. Since each ambulance run has an associated team of first responders, assigning the DUP to that ambulance run will enable the members of that team to read the system's recommendation from their mobile devices. Figure 1 shows two screens of the mobile interface used in the field: the screen on the left shows the motor control graphical representation for the considered patient, the screen on the right displays the recommendations, organized into sections and ordered according to their Recommendation Strength. Team members can thus examine recommendations while traveling to the emergency destination, so that they do not need to use the mobile device when they reach the patient.

4 Conclusions

At the time of writing, the project has entered a validation phase: each member of a pool of clinicians, who were not involved in the focus groups and in the expert panel, is now separately entering DUPs of real patients into the system. Each considered patient case is given to every involved clinician, to detect possible misunderstandings in the DUP forms as well as analyze consistency among clinicians in filling the DUP. Moreover, these clinical cases are being used to have the EMS physicians and nurses assess the usefulness of the recommendations provided by the system.

We have also started an exploration of using the DUP and the knowledge base for training purposes. In particular, we aim at integrating them in a serious game [6] to provide visual realism and user immersion in simulated EMS training scenarios.

Acknowledgements

Our research is partially supported by the Friuli Venezia Giulia region project "Servizi avanzati per il soccorso sanitario al disabile basati su tecnologie ICT innovative".

References

1. Binsted, K., Cawsey, A., Jones, R.B.: Generating Personalised Information using the Medical Record. In: Wyatt, J.C., Stefanelli, M., Barahona, P. (eds.) AIME 1995. LNCS, vol. 934, pp. 29–41. Springer, Heidelberg (1995)
2. Chittaro, L.: Visualizing Information on Mobile Devices. IEEE Computer 39(3), 40–45 (2006)
3. Center for Development and Disability. Tips for First Responders, 3rd edn.
4. Chittaro, L., Ranon, R., De Marco, L., Senerchia, A.: User modeling of disabled persons for generating instructions to medical first responders. In: Proc. UMAP 2009 Internat. Conf. on User Modeling, Adaptation, and Personalization. Springer, Berlin (2009)
5. Shiffman, R.N., Karras, B.T., Agrawal, A., Chen, R., Marenco, L., Nath, S.: GEM: A proposal for a more comprehensive guideline document model using XML. Journal of the American Medical Informatics Association 7, 488–498 (2000)
6. Cabas Vidani, A., Chittaro, L.: Using a Task Modeling Formalism in the Design of Serious Games for Emergency Medical Procedures. In: Proc. IEEE VS-GAMES Internat. Conf. Games and Virtual Worlds for Serious Applications, pp. 95–102. IEEE Computer Society Press, Los Alamitos (2009)

A Mobile Clinical Decision Support System for Clubfoot Treatment

Weiqin Chen and Dag Skjelvik

Department of Information Science and Media Studies,
University of Bergen, P.O. Box 7802,
N-5020 Bergen, Norway
{weiqin.chen,dag.skjelvik}@infomedia.uib.no

Abstract. In current congenital clubfoot treatment, clinicians use paper forms to register and monitor the treatment process. Routines for registration and archiving are scarce, and the guideline for treating clubfoot is not always followed strictly. This paper presents a PDA-based system (GenSupport) that can support the registration of patient information, supervise the treatment process, as well as provide advice during treatment. GenSupport has been evaluated in a pilot study.

Keywords: Clinical decision support systems, mobile health information, clubfoot treatment.

1 Introduction

Clinicians suffer almost universally from the problem of poor data quality, difficulty of access and bad communication. In addition, some individuals need support in decision-making. Therefore well-designed patient oriented information systems which improve the routines of registration and archiving of patient data and decision support systems which monitor and support treatments are needed throughout the health service [1]. In recent years, with the development of Internet and mobile technologies, research in healthcare has been shifted towards mobile Electronic health record and clinical decision support systems.

Clubfoot (Talipes equinovarus) is a congenital condition where the foot is deformed and turns inward and downward. It is the most common birth defect, and in most cases it is treated using mainly non-surgically methods. The Ponseti-Pirani method is now considered to be the worldwide standard of treating clubfoot [2]. In this method the Pirani score is for classifying the severity of the clubfoot and the Ponseti method for correction.

The treatment of clubfoot using this method is as of today not computer supported. The clinicians use paper forms to monitor the treatment. After the foot is scored, the results are plotted with a pen into a graph on paper. Classification results vary depending on the clinician performing it. Information registered about the patient is usually unstructured and archived in an ad-hoc manner and sometimes not archived at all. Thus, there are few possibilities to perform statistical analysis. The treatment

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 181–185, 2009.

process is in some cases ineffective because the Ponseti guideline is not followed strictly. Mistakes made by clinicians during treatment are often discovered too late and this can either corrupt or prolong the treatment process.

According to Osheroff et al. [3], the best opportunity for a computer-based system to deliver interventions is usually when the pertinent persons can be reached with the intervention and are prepared to act upon the information immediately. Handheld computers are the most versatile in stressful clinical environments, especially in those that are lack of infrastructure. Therefore we believe that a PDA-based system could improve the treatment by controlling registration of patient information, supervising the treatment process as well as providing advice during treatment. In this paper we present the design and evaluation of such a system (GenSupport).

2 The Design and Development of GenSupport

The design and development of GenSupport follows a user-centred approach. The domain expert has been closely involved in the process. The requirements were gathered through meetings and low-fidelity mock-ups. The system should be able to: 1) Allow clinicians to register core patient information; 2) Allow clinicians to register attributes of the clubfoot; 3) Provide treatment recommendations based on clubfoot treatment guideline and information registered by the clinicians using it; 4) Run on a handheld device (e.g. PDA with Windows mobile OS). The system has been designed as a generic framework for supporting clinical guidelines.

2.1 Clubfoot Treatment Guideline

The guideline for clubfoot treatment is represented in a decision tree (Fig. 1). The tree is composed of two types of nodes; internal nodes and leaf nodes. Internal nodes have a list of child nodes, while the leaf nodes have a list of statements. Each recommendation is a suggestion from GenSupport to the clinician to perform a certain action such as "Cast right foot for three weeks" or "Perform tenotomy on left foot" In addition to suggesting which actions to perform, GenSupport can also provide warning (e.g. "Warning! Check treatment of right foot") and error messages (e.g. "ERROR! Check treatment of left foot"). A warning message is given when treatment could be wrong, i.e. when there are reasons for suspecting that the treatment is not progressing as normal and special measures must be taken to prevent the treatment going wrong. When something indicates that the treatment most likely has gone wrong, an error message is provided to the user. The decision tree is implemented in GenSupport as an XML file. To support another guideline, one only needs to replace the configuration XML file with a new one which contains the new guideline.

2.2 Rule Engine

A specific rule engine (Eval3RulesEngine) was implemented for PDAs. This rule engine is based on an external library Eval3 which parses expressions represented by strings, and returns the truth value of the expression. When the Eval3RulesEngine runs, it parses the decision tree depth-first and evaluates the conditions of the rule at

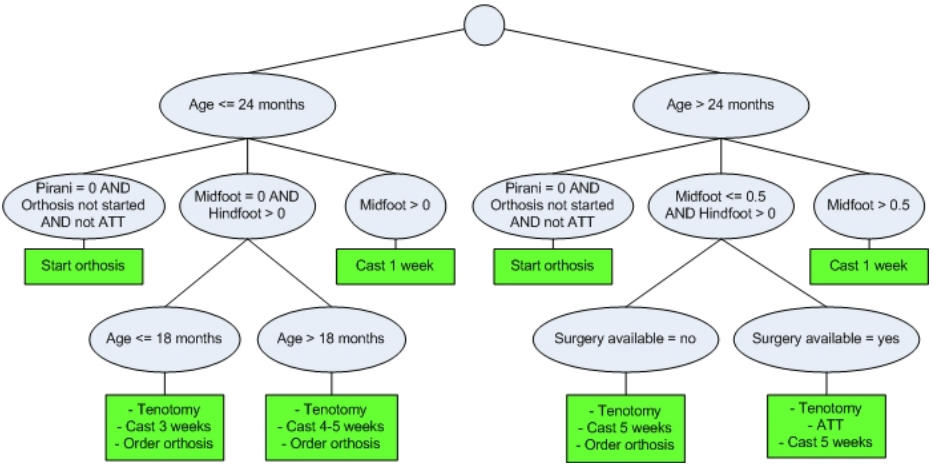


Fig. 1. Decision tree for clubfoot treatment (partial)

each node. Internal nodes evaluated as true will be expanded, whilst the tree will be cut at those internal nodes having a condition evaluated as false. When a leaf node is evaluated as true, its statement will be placed on the agenda of the Eval3RulesEngine. The agenda is a list of the statements contained in the leaf nodes evaluated as true. This list is available after the execution is finished. The rules engine parses the full decision tree, not stopping at the first leaf node evaluated as true.

2.3 Recommendations

When the clinician is provided with the recommendations for treatment, s/he can choose to reject the recommendations, plan actions according to the recommendations or add actions manually. According to Osheroff et al. [3] there are standard reasons for rejecting treatment recommendations: MD disagrees with recommendation, recommendation already implemented, alert fired inappropriately, patient ineligible for recommended intervention, patient refuses recommended intervention, and others. These standard reasons are implemented in GenSupport. The clinician should also provide a comment explaining why the overriding was done.

It is not always possible or feasible to perform the recommended action instantly. This calls for the need to be able to plan the actions. In GenSupport it is possible for the clinician to postpone an action, and plan for when to perform it. A date must be specified, and the clinician should state a comment about why the action is postponed.

The clinicians are able to manually invoke an action if they believe that a certain action is correct to perform under the given conditions even though GenSupport has not suggested it. They can choose amongst the actions which are specified in the current guideline. When choosing an action manually, they should specify the reason for the choice.

3 Evaluation

A pilot study was conducted to evaluate GenSupport focusing on whether it is able to support the clubfoot treatment. This study included a quantitative and a qualitative evaluation.

3.1 Quantitative Evaluation of Recommendations

A data set containing full treatment history on 17 patients having congenital clubfoot on right foot, left foot or both feet was used.

As shown in Table 1, GenSupport provided the same recommendations of the treatment as performed by the clinician on 5 of the 17 patients. On the other patients, the system advised to perform the tenotomy either before (7 of 17) or after (5 of 17) it was actually performed by the orthopaedian.

Table 1. Evaluation of the recommendations by GenSupport

Status	Cases
Recommedations identical with the treatment performed by the clinician	5
Tenotomy recommended before actually performed by the clinician	7
Tenotomy recommended later than actually performed by the clinician	5

The Ponseti expert group specifically recommends performing tenotomy as soon as the midfoot score is 0. When the clinicians are in doubt about whether the procedure should be performed, they should perform it [2]. In seven of the cases investigated in this test, tenotomy was not performed according to the recommendations from the Ponseti expert group. In these cases, the recommendations were correct and the clinicians provided a sub-optimal treatment.

In five of the cases, the clinicians performed tenotomy earlier than GenSupport recommended. In these cases, there is no apparent pattern describing why the clinicians have acted as they have. The clinicians' actions are most likely based on factors not documented in the patient data available in this evaluation.

3.2 Qualitative Evaluation of Functionality and Usefulness

Qualitative evaluation methods such as think aloud, observation and semi-structured interviews were used in this part of the evaluation. All three clinicians are experienced in treating clubfoot.

Functionality and perceived usefulness. All the clinicians were generally satisfied with GenSupport. Due to their high level of expertise, they did not believe they could benefit from getting treatment advice from this clinical decision support system. They believed that GenSupport would be best suited as a tool for training novice clinicians.

All the clinicians identified an area which GenSupport could help improving: the current routines of registration and archiving data about the treatment. GenSupport

can help to improve the registration by "forcing" the clinicians to register proper treatment data while treating the patient.

Experience with PDA. One interesting findings from the evaluation is that none of the clinicians in the evaluation had difficulties using the handheld computer although none of them have any prior experience with PDAs. Observation showed that the clinicians in the evaluation had no problems at all using the soft keyboard, even though they were not used to such a small user interface.

When asked to compare mobile devices with desktop computers in the daily practice, they emphasized that a handheld computer is easier to use and transport in a hectic clinical environment. They also pointed out that the handheld computers are more robust than a regular computer. They are resistant to dust and shock. Another advantage is the quick start-up time of the handheld computers, compared to a regular computer.

4 Conclusion and Future Work

This paper presents the development and evaluation of GenSupport—a mobile decision support system for clubfoot treatment. The system was found to be able to improve and simplify the registration process and "force" the medical personnel to follow routines more strictly. It is also considered to be a useful tool for less experienced clinicians [4].

According to the domain expert, tenotomy is only performed on approximately 50% of the patients by inexperienced clinicians. However, Ponseti estimates that 90% of the patients need tenotomy [2]. Reliable registration of treatment data would make it possible to investigate whether the need for tenotomy is actually as low as 50%, or whether the procedure is not performed often enough.

References

1. Garg, A.X., Adhikari, N.K., McDonald, H., Rosas-Arellano, M.P., Devereaux, P.J., Beyene, J., Sam, J., Haynes, R.B.: Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. Journal of the American Medical Association 293, 1223–1238 (2005)
2. Staheli, L.: Clubfoot: Ponseti Management, 2nd edn. Oxford Medical Publications (2005)
3. Osheroff, J.A., Pifer, E.A., Teich, J.M., Sittig, D.F., Jenders, R.A.: Improving out-comes with clinical decision support: an implementer's guide. Healthcare Information and Management Systems Society Press, Chicago (2005)
4. Godin, P., Hubbs, R., Woods, B., Tsai, M., Nag, D., Rindfleisch, T., Dev, P., Mel-mon, K.L.: A New Instrument for Medical Decision Support and Education: The Stanford Health Information Network for Education. In: Proceedings of the 32nd Hawaii International Conference on System Sciences. IEEE Computer Society Press, Maui (1999)

An Ambient Intelligent Agent for Relapse and Recurrence Monitoring in Unipolar Depression

Azizi Ab Aziz, Michel C.A. Klein, and Jan Treur

Agent Systems Research Group, Department of Artificial Intelligence
Vrije Universiteit Amsterdam, De Boelelaan 1081a,
1081 HV Amsterdam, The Netherlands
{mraaziz,michel.klein,treur}@few.vu.nl

Abstract. Mental healthcare is a prospective area for applying AI techniques. For example, a computerized system could support individuals with a history of depression in maintaining their well-being throughout their lifetime. In this paper, the design of an ambient intelligent agent to support these individuals is presented. It incorporates an analysis and support model for diagnostics based on observed features and for suggested actions. The model used is based on dynamic relations that describe the occurrence of relapse in unipolar depression. By incorporating this model into an ambient agent system, the agent is able to reason about the state of the human and the effect of possible actions. Several simulation experiments have been conducted to illustrate the functioning of the proposed model in different scenarios.

Keywords: Ambient Agent Modeling, Relapse in Unipolar Depression, Temporal Dynamics, Decision Support Systems.

1 Introduction

In many cases, depression is a recurring condition; a subsequent depressive episode is called a relapse or recurrence. In principle, the depressive relapse stage can be defined as "episode of major depressive disorder that occurs within six months after either response or remission (no longer meeting the depression criteria)", while, recurrence is a depressive episode occurs after six months have elapsed [3]. Several related works on depression relapse suggested that at least 50 percent of patients who recover from an initial episode of depression would experience at least one subsequent depressive episode throughout their lifetime [4]. Before a relapse happens, there might be changes in the usual symptoms of the illness, or changes in behaviour, thoughts or feelings. Therefore, the earlier those symptoms can be identified, the better chance there is of stopping a relapse / recurrence or reducing the severity of it. To envisage this possibility, there are several conditions to be evaluated, namely; (1) the *neuroticism* of a patient (exaggerating ordinary situations as threatening), (2) the *immunity* against negative feelings (which can be low because of residual symptoms and a history with onset), (3) *lack of social support* (disengagement from social interactions), (4) the *assertiveness* (if it is low, it results in a lack of self esteem and poor control over anger), and (5) *avoidance coping* (a tendency to solve a problem by avoiding it,

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 186–190, 2009.

which can be signalled by e.g. substance abuse) [3][4]. Essentially, stressful events are one of the most dominant factors that will trigger relapse or recurrence. These events may cause from trauma, grief, pressure, or even from typical daily hassles (such as traffic congestion). Based on the factors described above, a domain model for the occurrence of relapse or recurrence of a depression has been developed [1]. The simulation results have shown the model exhibits important patterns between the events and the course of relapse and recurrence.

In the past, intelligent agent technology has become an important means for increasing decision-making ability and communication. With the advent of wearable devices, and mobile applications, new ways are created for agents to interact, and react about human related information gathered from sensors. Such kind of agents, known as intelligent ambient agents, will be able to contribute to the development of personal care and human wellbeing applications by harnessing vital information from human itself [6]. In this paper, an intelligent ambient agent is presented that could support people that have recovered from a depression to maintain a healthy state, using the domain model for relapse or recurrence of a depression described above.

The remainder of the paper is organized as follows. Section 2 provides the main design of the model. Later, simulation results for selected observable features using the model are illustrated and described in Section 3. Finally, Section 4 concludes the paper.

2 Modelling Approach

The key contribution of this paper is the design of an ambient agent to support people recovered from a depression. In order to achieve this, an approach has been followed in which the domain model for depression functions as starting point for the model that describes the functioning of the ambient agent. Thus, by integrating the domain model, the ambient agent will be able to reason about the processes and its environment. It is obviously important to have such capabilities, since an ambient agent should be aware of human behaviours and states. Through this mechanism, the agent will use this knowledge to provide related actions related to the predicted state of the human and the environment. Figure 1 presents the overview of the integrated model.

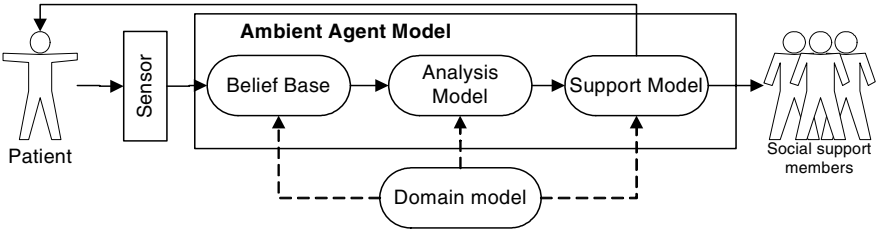


Fig. 1. Overview of the Integrated Model

The functioning of the ambient agent can be largely described by three components: an analysis, a belief base, and a support model. For the detailed design, see (http://www.few.vu.nl/~mraaziz/AIME09/AIME2009-model.pdf). In general, the ambient agent interacts with a patient through a set of non-obtrusive ambient devices

(i.e.; medicine box that registers medication intake (MEMS), a passive alcohol sensor, a mobile phone/ personal digital assistant (PDA), and a blood volume pressure sensor) [5][7]. This model is developed using a temporal specification language called LEADSTO. Consider the format of $\alpha \rightarrow\!\!\!\!\rightarrow_{e,f,g,h} \beta$, where α and β are state properties in form of a conjunction of atoms (conjunction of literals) or negations of atoms, and *e,f,g,h* represents non-negative real numbers, then it can be interpreted as follows [2]:

> If state property α holds for a certain time interval with duration *g,* after some delay (between *e* and *f*), state property β will hold a certain time interval of length *h*.

To specify properties on dynamics relationship, the ontology of the model is designed using predicate calculus. The detailed ontology of the model can be found at (http://www.few.vu.nl/~mraaziz/AIME09/ontology.pdf). Using this pre-determined ontology, the Belief-Desire-Intention (BDI) approach regulates action selection process (internal processing) [6]. The temporal rules of an ambient agent have been specified using the ontology. To utilize the specification, a forward method for belief generation is used. It allows the time sequence and causality, to generated new belief from previous properties. The ambient agent functionality is described by three actions; belief generation in belief base, evaluation of risk, and action selection for the support. Below are several related specifications for social withdrawal case.

BB4: Generating basic belief on phone/PDA usage
When the ambient agent observes there is no phone/PDA usage, then the agent will believe that a patient is not using phone/PDA to communicate with the others.

observed(agent, phone_usage(negative)) $\rightarrow$ belief(agent, phone_usage(negative))

DB5: Derived belief on social support from the phone usage belief
If the ambient agent believes that there is no phone usage then the agent will believe there is no social interaction between social support network members.

belief(agent, phone_usage(negative)) $\rightarrow$ belief(agent, social_support(negative))

GE2: Evaluation on social withdrawal condition
If it is believed that patient is not interacting with any social network support members, and having difficulty to control anger and it is believed that patient is vulnerable for the future onset then the agent will conclude that the condition of the patient is having social withdrawal.

belief(agent, social_support(negative)) $\wedge$ belief(agent, assertiveness(low)) $\wedge$ belief(agent, immunity(low)) $\rightarrow$ assessment(agent, social_interaction(low))

PCB2: Predicting the risk of relapse from social withdrawal condition
If the patient is having social withdrawal then the ambient agent will assess the patient as having potential risk of relapse.

assessment(agent, social_interaction(low)) $\rightarrow$ prediction(agent, stage(risk_relapse, positive))

BOR: Belief on relapse
When the ambient agent predicts the patient is having a risk in relapse, then the ambient agent will believe the patient is in the risk of relapse.

prediction(agent, stage(risk_relapse, positive)) $\rightarrow$ belief(agent, stage(risk_relapse, positive))

ANR1: Action to notify social support networks
When the ambient agent believes the patient in the risk of relapse then the ambient agent will notify all friends and family within the social support network.

belief(agent, stage(risk_relapse, positive)) ⇛
performed(agent, notify(risk_relapse, friends_family))

ANR2: Action to notify the patient
When the ambient agent believes the patient in the risk of relapse then the ambient agent will notify the patient.

belief(agent, stage(risk_relapse, positive)) ⇛ performed(agent, notify(risk_relapse, patient))

DSI: Desire to improve social interaction
If the ambient agent assesses the patient is having social withdrawal then the ambient agent will desire to improve patient's social interaction by advising the patient about suitable social activities.

assessment(agent, social_interaction(low)) ∧ desire(agent, reduced(risk_relapse))
⇛ desire(agent, improved(social _activities))

ISIA: Intention to advice on social interaction
When the ambient agent desires to improve patient's social interaction through social activities and ambient agent believes there is no social interaction between a patient and social support network members, then ambient agent will have an intention to advice patient on suitable social activities.

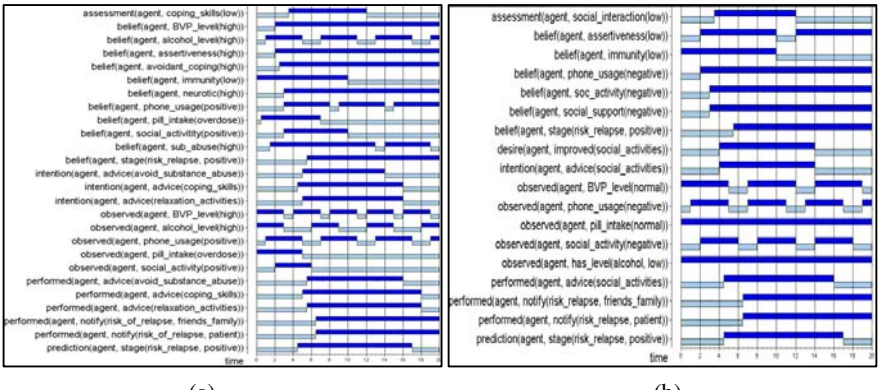
desire(agent, improved(social _activities)) ∧ belief(agent, social_support(negative))
⇛ intention(agent , advice(social_activities))

ASIA: Action to advice on social interaction activities
When the ambient agent intends to advice the patient regarding to social activities to the patient, then the ambient agent will advice the patient about those social activities.

intention(agent , advice(social_activities)) ⇛ performed(agent , advice(social_ activities))

3 Simulation Results

This section describes the simulation results for two scenarios in which the ambient agent monitors the risk on relapse and provides support by illustrating the functioning of the analysis and support model in the agent. In the figures below, time is shown on the horizontal axis, and the state properties are on the vertical axis; a dark box indicates that a state property is true.



(a) (b)

Fig. 2. Simulation traces (a) coping skills deficiencies (b) social withdrawal

Fig.2 (a) depicts a scenario where the ambient agent observes no activities in social interaction, a low assertiveness, and concludes that the patient is highly vulnerable to future onset. The patient is strongly advised to initiate social interaction with others. People within the social support network will be notified by an ambient agent. While in Fig, 2(b), the ambient agent observes a high blood volume pressure, high alcohol level, and an overdose pill intake. Based on this, the agent assesses that the person is having a risk of relapse, and that this is related to coping skills problem. Therefore, the desires to give advice to improve coping skills, to reduce anxiety and to eliminate substance abuse are translated into intentions to do so, as the beliefs about the conditions are true.

4 Conclusion

In this paper, an ambient agent model was presented for automated relapse and recurrence monitoring, developed using a modelling approach in which the domain model of a process (in our case depression recurrence) forms the basis for the functional model of the agent. By compiling knowledge from the domain model into the agent model, the agent is able to reason about the state of the patient. Thus, it is capable to predict the risk of relapse based on several observable features and beliefs. The proposed model is heavily inspired by scientific findings about the relapse and recurrence. The model has been specified using a formal modelling approach, which enables a qualitative specification. The ambient agent model has been applied to several scenarios in a simulation environment. The presented model provides a basic design on how an ambient model can be used to monitor patient in a risk of relapse and recurrence in unipolar depression. Apart from a more thorough evaluation of the proposed system, future work will focus on generalizing the proposed model to a generic model for risk assessment and support in other domains.

References

1. Aziz, A.A., Klein, M.C.A., Treur, J.: An Agent Model of Temporal Dynamics in Relapse and Recurrence in Depression. In: Ali, M., Chen, S.M., Chien, B.C., Hong, T.P. (eds.) IEA-AIE 2009. LNCS (LNAI). Springer, Heidelberg (to appear, 2009)
2. Bosse, T., Jonker, C.M., van der Meij, L., Treur, J.: A Language and Environment for Analysis of Dynamics by Simulation. International Journal of Artificial Intelligence Tools 16, 435–464 (2007)
3. Keller, M.B.: Long-Term Treatment of Recurrent and Chronic Depression. J. Clinical Psychiatry 62(24), 3–5 (2001)
4. Neirenberg, A.A., Petersen, T.J., Alpert, J.E.: Prevention of Relapse and Recurrence in Depression: The Role of Long-Term Pharmacotherapy and Psychotherapy. J. Clinical Psychiatry 64(15), 13–17 (2003)
5. Picard, R.W., Liu, K.K.: Relative Subjective Count and Assessment of Interruptive Technologies Applied to Mobile Monitoring of Stress. International Journal of Human- Computer Studies 65(4) (2007)
6. Sharpanskykh, A., Treur, J.: An Ambient Agent Model for Automated Mindreading by Identifying and Monitoring Representation Relations. In: PETRA 2008. ACM Pub., Athens (2008)
7. Zhai, J., Barreto, A.B.: Instrumentation for Automatic Monitoring of Affective State in Human-Computer Interaction. In: 18th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2005), pp. 207–212 (2005)

An Advanced Platform for Managing Complications of Chronic Diseases

Davide Capozzi and Giordano Lanzola

Biomedical Informatics Laboratory,
Department of Computer and Systems Science, University of Pavia,
Via Ferrata 1, 27100, Pavia, Italy
davide.capozzi@unipv.it, giordano.lanzola@unipv.it
http://labmedinfo.org

Abstract. This paper describes a generic platform for telemedicine services aimed at supporting chronic outpatients. The framework comprises a server agent and several cell phones as mobile agents through which patients and caregivers within their families may input data and receive back suggestions and advice. Mobile agents and server are endowed with domain specific knowledge in order to support users in a flexible way structured on multiple levels.

1 Introducion

Chronic patients are required to manage their disease by themselves [1]. This means that they have to constantly monitor and assess their illness state in order to adopt each time the most appropriate decisions for controlling it, as they lack an immediate advice by trained health care professionals. To this aim, there has been recently a growing concern about complementing traditional education with self-management education for supporting patients in achieving the best possible quality of life despite their chronic conditions [2]. Controlled clinical trials are also suggesting that programs supporting patient self-management are more effective than information-only patient education in improving clinical outcomes [3].

Thus we believe that there is a strong need for tools supporting a comprehensive interaction among patients, relatives and clinical staff. In this paper we propose an advanced telemedicine platform which makes use of mobile network devices such as cell phones, palmtops and PDAs in addition to PC's in order to provide an advanced framework for remote monitoring which also facilitates cooperation between patients and their families as well as with health care staff.

2 The Functional Overview

In this work our goal has been to exploit the latest achievements in ICT for supporting chronic patient management [4,5] from two separate perspectives, corresponding also to different logical information flows into opposite directions,

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 191–195, 2009.

as shown in Figure 1 where all the involved parties are illustrated. In the topmost part of the figure the health care staff responsible for patient monitoring and assessment is indicated, while the bottom part of the figure represents the patient himself and his family. As it transpires, between those two layers a bidirectional information exchange takes place with our platform playing the role of a facilitator. On one side we support patients and their families in integrating the health records traditionally collected at regular interviews with timely information autonomously and proactively entered by them. This is represented through the upward link which carries *Clinical Data* and is what we refer to as the forward information flow.

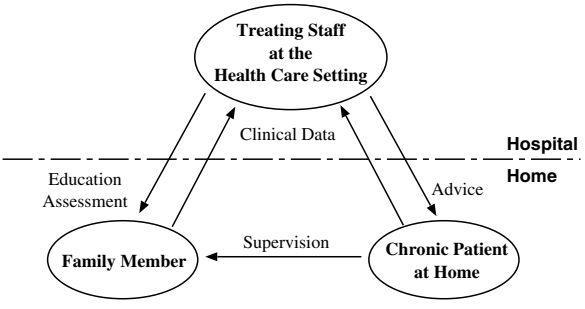


Fig. 1. The logical information flow among the actors involved in the home care scenario supported by the platform

The backward information flow, which is represented by links descending from the upper layer to the lower one provides instead a different kind of support. On the right part with the link labeled *Advice* we point out the pieces of advice mainly meant for patients and helping them in properly self-managing the disease. This flow should be fully configurable according to the treatment requirements and may range from simple reminders to accomplish some specific action to more comprehensive support perhaps aimed at interpreting and assessing the seriousness of a newly acquired evidence, thereby giving rise to a fully *context-aware* system. On the left part of the figure a separate information flow is depicted involving the family. Besides contributing to the input of additional clinical information the family supervises the patient treatment. Thus the information they receive from the staff entails *Education* and *Assessment* while, depending on the situation, they might also receive some feedback straight from the patient.

3 The Architecture

Given the current limitations exhibited by mobile devices along with some organizational issues we focused on a multi-layer architecture for supporting communication between all the actors involved. Figure 2 shows a physical overview

of the platform highlighting the main components required to support the above mentioned logical information flow and the interconnecting links. We have partitioned the platform into two different interconnected agent classes, the *Server Agent* and *Mobile Agent*s. The first one is supposed to be located at a clinical setting, represents the system core and is devoted to hosting the *Electronic Medical Record* where all data concerning any given case are saved. Mobile agents instead are meant to be the primary device for acquiring data from patients and providing feedback to their users, be they patients or relatives. Finally, the bottom part of the figure represents a medical device such as a glucometer, scale, blood pressure monitor or a similar one. Depending on the technology those devices may interface directly with the *Local Short-term Database* located on the mobile agent through a wireless link or, more seldom, behave as remote agents and connect to the server agent through a regular UMTS link. Manual data entering is always permitted through a simple GUI available on mobiles.

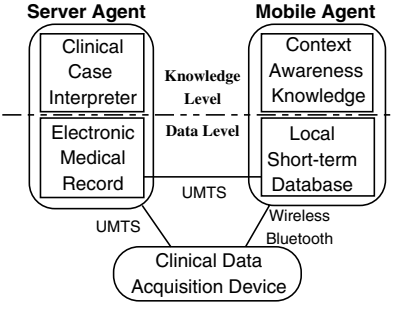


Fig. 2. The physical infrastructure of the platform

Besides the horizontal partitioning there exists also a vertical one among the components. In fact, the bottom part implements the *Data Level* which merely concerns data acquisition and processing. This level entails actions such as the acquisition of new clinical information by mobile agents, feedback provided to the user or the bidirectional data exchange occurring between mobile and server agents. In the higher part of the figure instead the *Knowledge Level* is located, whose purpose is to provide an interpretation of data depending on the context. Both server and mobile agents feature a module at this level, although their functionality will be uncomparable, mainly because of the very different computational environments. On the mobile side the *Context Awareness Knowledge* may be used to provide shallow data interpretation based on patient context before those data are sent to the server agent.

The server exposes on the internet several network services for *Sync*hronizing the EMR between server and mobiles, for accessing and editing patient data also through usual *Web* pages, and for *Report*ing about the patient state. Services are implemented in terms of plugin modules and interact with the framework through a standard interface enabling them to exploit all the agent capabilities from the backend data models to the frontend.

3.1 Mobile Agent

Mobile agent provides local connectivity with a set of supported measurement devices (glucose monitors, scales, blood pressure monitors) exploiting the Bluetooth standard, and internet access through GPRS/UMTS protocols. As displayed in Figure 3 its core includes an *Ontology* that models the patient's EMR and additional domain specific concepts depending on the addressed medical problem. Any other module included in the agent should interact with the ontology through a *Mapper*, whose purpose is to convert data available within the ontology into each specific module representation. For example, the *Customized Knowledge* module is service dependent and implements specific behavior based on data for the particular application (i.e. firing up alarms or rising up a reminder).

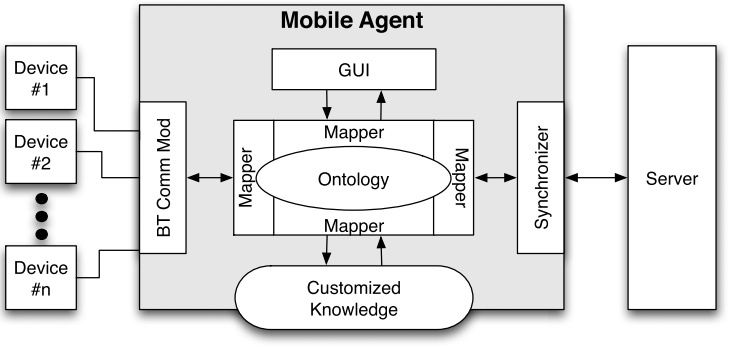


Fig. 3. The structural view of the Mobile Agent

Mobile agent is equipped with a *Bluetooth Communication Module* (BTC) through which peripheral devices not GPRS-enabled may automatically upload the latest data using the mobile as a relay. The availability of a BTC module allows us to distribute system intelligence and create local clusters of devices directly managed by the application running on the mobile, which then takes upon itself the burden of synchronizing data with the server agent or receiving notifications from it. A key feature of the architecture, allowing mobile agents to consistently exchange data, is the *Synchronizer* module that implements the same protocol available on the server agent. Its purpose is to guarantee a regular alignment between mobile and server concerning any data modelled according to the specific ontologies available within those agents. Finally the *Graphical User Interface* (GUI) module provides a tool allowing patients or relatives to be informed about messages from the application (i.e. infos, alarms, reminders, confirmation dialogs etc.), to configure service dependent parameters and also to have an alternative way for manually entering measures.

4 Conclusion

In this paper we described the design and implementation of a telemedicine platform facilitating remote monitoring on chronic outpatients by the treating staff and providing advice back either to them or their caregivers.

The platform has been used for implementing two prototypes addressing different clinical problems. A first one is meant to support young patients affected by Type I Diabetes who use a mobile agent to fill in their daily logbook concerning blood glucose level measurements and insulin doses. Data is entered using the GUI module of the mobile mainly because in Italy there are no commercially available glucometers supporting wireless connectivity [6]. In this case we were able to exploit the plugin architecture for reusing problem specific knowledge chunks developed in previous projects addressing similar problems [7]. The second prototype addresses instead patients undergoing peritoneal dialysis, who need to acquire daily measures concerning their weight and blood pressure. In this case data is automatically acquired from devices through a Bluetooth connection, displayed on the mobile agent for review, and finally sent to the server agent. In both cases feedback is provided either to patient or to relatives concerning the clinical state. An experimental phase concerning this prototype is expected to start before summer at a major hospital located in northern Italy.

References

1. Clark, N.M.: Management of chronic disease by patients. Annu. Rev. Public Health 24, 289–313 (2003)
2. Lorig, K.R., Holman, H.: Self-management education: history, definition, outcomes, and mechanisms. Annals of behavioral medicine: a publication of the Society of Behavioral Medicine 26(1), 1–7 (2003)
3. Bodenheimer, T., Lorig, K., Holman, H., Grumbach, K.: Patient self-management of chronic disease in primary care. JAMA 288, 2469–2475 (2002)
4. Hine, N., Judson, A., Ashraf, S., Arnott, J., Sixsmith, A., Brown, S., Garner, P.: Modelling the Behaviour of Elderly People as a Means of Monitoring Well Being. LNCS, pp. 241–250. Springer, Heidelberg (2005)
5. Welch, G., Shayne, R.: Interactive behavioral technologies and diabetes self-management support: Recent research findings from clinical trials. Current Diabetes Reports 6(2), 130–136 (2006)
6. Lanzola, G., Capozzi, D., D'Annunzio, G., Ferrari, P., Bellazzi, R., Larizza, C.: Going mobile with a multiaccess service for the management of diabetic patients. Journal of Diabetes Science and Technology 1(5), 730–737 (2007)
7. Larizza, C., Bellazzi, R., Stefanelli, M., Ferrari, P., De Cata, P., Gazzaruso, C., Fratino, P., D'Annunzio, G., Hernando, E., Gomez, E.J.: The M2DM Project–the experience of two Italian clinical sites with clinical evaluation of a multi-access service for the management of diabetes mellitus patients. Methods Inf. Med. 45, 79–84 (2006)

One Telemedical Solution in Bulgaria

P. Mihova, J. Vinarova, and I. Pendzhurov
New Bulgarian University, Bulgaria
Montevideo str.21, Sofia, Bulgaria, 1618
pmihova@nbu.bg

Abstract. A new arena of healthcare is emerging, because physicians, hospitals, financial health planners and administrators are coming together in a single highly integrated and coordinated virtual health organization.The mission of Telemedicine is to provide medical services independently of geographical distances between the involved sites. Patients can get access to medical expertise that may not be available at the patients' site through Telemedicine. Experience over the last decade has shown that the goals of Telemedicine are not automatically reached by the introduction and use of particular new technologies per se, but rather require the implementation of *integral services and specialized information systems*.Software Teleconsult aims are to provide logistic and telemedical services between two distant hospitals on the territory of Bulgaria. The objectives of this development are to combine medicine field, e-health and informatics research so as to share information and experience in order to perform the best patient services at a reasonable price.

Keywords: telemedicine, e-health, MIS.

1 Introduction

The transfers of electronic medical files allow medical practitioners to be involved in diagnostic activities without being in the same physical location with the patient.

We introduce a desktop development, a secure and scalable teleconsulation solution, purposefully designed for connection between national hospital, situated in the capital of Bulgaria and one small remote city hospital.

The software is used by 19 authorized physicians from both hospitals and the two hospital managers.

2 Materials and Methods

The presented software for teleconsultations is organized in the following way:

- Main software desktop solution is divided according to the operational level into three main parts (Fig.1.) – separated management modules that are developed according to the requirements and necessary functions for each participant in the telemedical process.
- Video communication by means of recently developed application with individual virtual rooms, locked and password protected meetings.
- Web portal with Forum for participants, patients and opinions.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 196–200, 2009.

This integral solution performs the ability to verify whether a receiving physician is available, whether the receiving system can receive the transmitted files, whether the receiving system has received all prior files, and where there is medical record continuity. According to Bulgarian legislation, each patient is identified only by age, sex and physical conditions, in order to keep the patients` privacy and confidentiality.

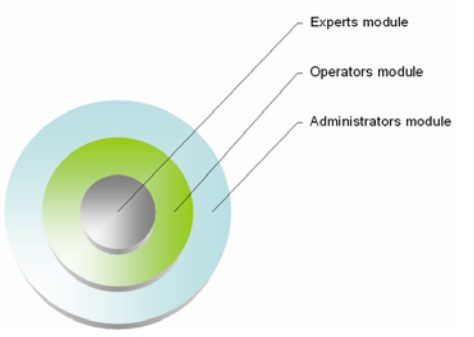


Fig. 1. Three layer architecture

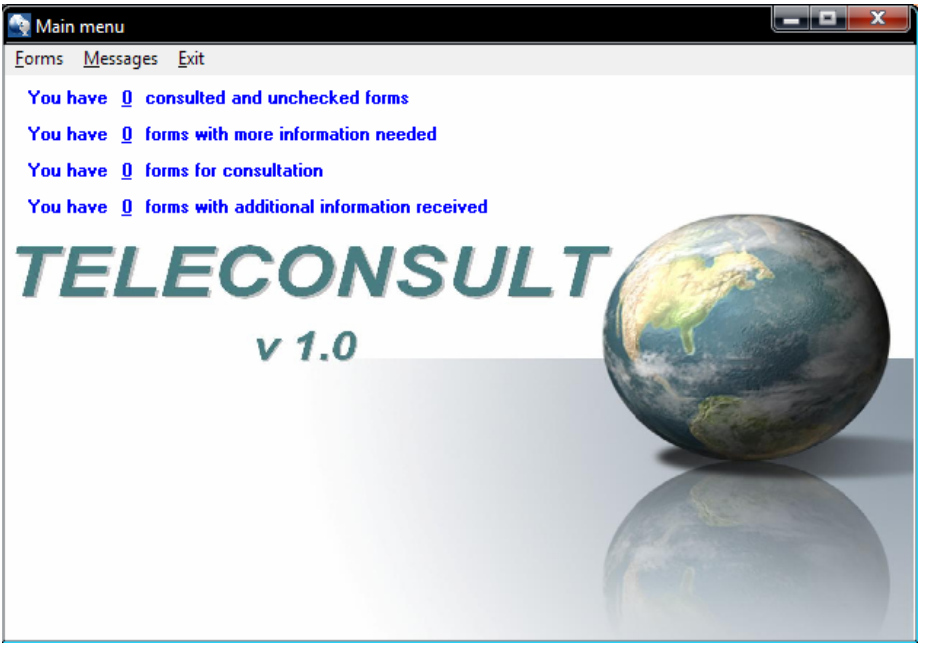


Fig. 2. Entrance screen Teleconsult

The request for consultation consists of several fields within the form, demonstrated at Fig.2:

- **Paraclinic examinations -** Blood tests, Patomorphologic and Urine tests are completely identical to the original paper document.
- **Image examinations** are: ECG, X-ray, Echocardiography, Velo test, Holter, Scanner, Mammography and etc.

These examinations allow uploading unlimited number of images, text description fields – both for the consulting and giving consultations experts.

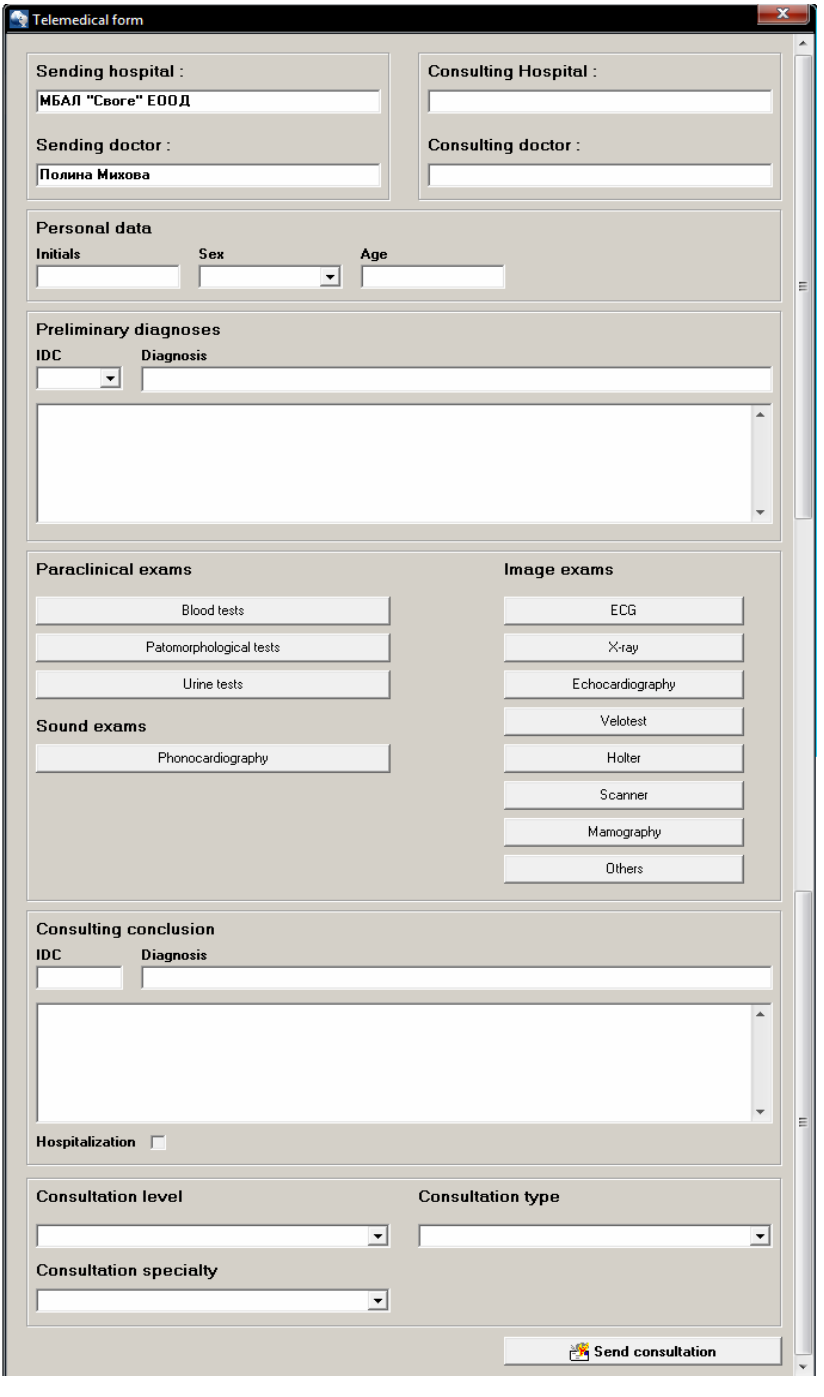


Fig. 3. Telemedicine form for consultations

In order to prove the usage and benefits of telemedical investments, the statistic basis of the system is organized in 69 different sections, due to the requirements of performing well-controlled healthcare services:

1. Doctor & Hospital statistics

- From date to date
- Number of required consultations
- Number of accomplished consultations

2. Reference to Word, Excel and with graphical visualization:

Filtering through starting and ending date, level of consultation:

- Number of consultations per period
- Number of consultations with result - hospitalization
- Percentage distribution according to specialists
- Number of consultations with second consultation
- Percentage distribution of correspondence between working and final diagnoses
- Percentage distribution of final diagnoses according to disease types

One of the most important statistics is actually the chronology of the system and control of each activity of every person.

The system also makes registration of the following user parameter like: name, action, host, address, day, month, year, hour, minutes and seconds. It filters any of the parameters and performs results to Excel, Word and Graphical programs.

Expert's module consists of some specific telemedical consultation characteristics as: required, consulted, not checked and with request for more information, are indicated with its own color.

The main coordinator in the system is the Operator, who manages the **expeditiousness** of the process of giving consultation, and in case of 24 hours delay of the consulting expert answer, the system automatically redirects the form to the available specialists. In case of getting few requirements at the same time to a specific specialist, the system distributes requirements to available specialists in the corresponding specialty.

The Administrator performs functional connection between users and software developers, which is realized within system mailbox. It has the authorization to make any kind of statistics for everybody at any time.

Table 1. Diagnoses, made via TIS "Teleconsult"

Specialty	Number of consulted patients	Without hospitalization	Patients with reduced hospital stay
Surgery	5	2	2
Gastrointestinal section	7	4	2
Obstetrics & Gynecology	2	2	-
Cardiology	15	9	4
Pulmology	4	3	-
Nephrology	15	11	2
Endocrinology	8	4	1

System Usage

The clinical approbation and introduction results show that more than 60% of consulted patients are not hospitalized, in 25% there is a positive influence in the hospitalized – reduced hospital stay. In more than 10% of the cases is demonstrated complete compliance between expert's opinions. The diagnoses made via TIS "Teleconsult" for 56 patients are summarized in Table 1.

3 Lessons Learned

The success or failure of a Telemedical Information System is mainly determined by its acceptance at the various levels of the administrative hierarchy. User may initially respond positively to surveys based on the impression of telemedicine concepts, but when they face any kind of difficulty – the initiatives go fast to the bottom. A mixture of face-to-face consultations and consultations via telemedical solution may be necessary and much more convenient to provide complete diagnoses and treatment to a patient.

4 Conclusion

Successfully demonstrated TIS "Teleconsult" is potential for managing teleconsultations from distance. The system was used regularly to perform diagnoses and follow up for residence. System usage allows patients having a consultation quickly and without visiting a doctor face-to-face. The system avoids mistakes and provides better care about patients reducing information misunderstandings, performing more than one opinion and ensuring experts qualifications at the same time.

According to users' opinion the system has got only high approval and patients' satisfaction that have been using the system for about one year.

The conclusion is that TIS "Teleconsult" can make positive impact on health care and can be expanded to other remote places, community centers and general physicians' offices.

References

1. Vinarova, J., Mihova, P.: Medical informatics, NBU, Sofia (2009) ISBN 13: 978-954-535-515-8
2. Владзимирський А.В.: Оценка эффективности телемедицины", Донецк (2007) ISBN 978-966-335-031-8
3. Mahau, M., Whitten, P., Allen, A.: E-Health, Telehealth and Telemedicine, A guide to start-up, Jossey-bass (2000) ISBN:0-7879-4420-3
4. Mihova P.: Telemedicine software Teleconsult – design, exploitation and results, Том 6, №2. Ukrainian Journal of Telemedicine and Medical Telematics (Online) (Print) ISSN 1811-1688, ISSN 1728-936X
5. Godlevski, L., Kalinchuk, S., Bayazytov, N., Smirnov, I., Adeyinka, M., Samchenko, I., Bayzakov, U.: First results of implementation of telemedical services in the Odesa region, http://strony.aster.pl/pdf/PolJMedPhysEng2007vol13no2_105-114_godlevsky.pdf

A Novel Multilingual Report Generation System for Medical Applications

Kaya Kuru[1,2], Sertan Girgin[1], and Kemal Arda[2,3]

[1] Gülhane Military Medical Academy
kkuru@gata.edu.tr
[2] Middle East Technical University
[3] Ankara Oncology Hospital

Abstract. There has been an increasing demand for high quality medical data that are in a standard electronic format and easily shared. Although a great amount of effort has been invested to ease the process, an effective solution has yet to be found. In this study, we first discuss necessary features of an effective data collection and reporting system, and then reveal the conceptual view of a novel method that aims to encompass these features. We also present the design and implementation details of a Web-based prototype.

1 Introduction

Clinical reports are the primary means of communication between laboratory professionals and referring physicians to guide them for a better health care service. A common complaint by laboratory professionals is that of incomplete, inadequate, or inappropriate clinical information from physicians requesting studies; clinicians, on the other hand, express concerns that interpretations in medical reports are often not relevant to the clinical questions they seek to answer [1,2]. Unfortunately, current reporting methods are insufficient in establishing the required communication medium [2], and this leads to avoidable *medical errors* which cost both human life and substantial amount of money [3].

The results of several studies that have shown that professionals frequently express the need for improvements in report quality at their institutions [4], indicate a necessity for new methods that are both effective and with less cognitive pressure – in between free-text reporting and sophisticated menu-driven structured approaches, which would provide a through communication among professionals, and also facilitate high level operations, such as population based inferences and diagnosis/decision support.

In this study, we propose a novel methodology, called "Structured, Interactive, Standardized, Decision Supporting" (SISDS), to remedy this situation. SISDS combines most of the favorable features of the exiting methods while removing their deficiencies. The interactivity with a versatile, user- and problem-driven, scalable and dynamic reporting scheme allows to avoid inefficiency and reduce the cognitive overload. The feedbacks received from a prototype of SISDS applied to the field of radiology, show that SISDS is more effective in many different perspectives and helps health care professionals practice better medicine.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 201–205, 2009.

2 The SISDS Method

In medical reporting, we can identify two main goals: (i) to make medical reports easily accessible, complete and comprehensible by all users, and (ii) to be able to extract medical data out of them for further analysis. In order to accomplish these goals, abstraction at several layers seems essential.

Data level. The data fields, or *variables*, that constitute a report must be consistent and *well defined*. A typical medical report contains many nominal and numerical values with different measurement units (such as, temperature, length, weight, volume, date, etc.), and without specific data-types for them it is unavoidable to lose some information when working directly with the data afterwards. Specific data-types also enable unit conversion, which facilitates information sharing. The ability to assign *default values* to data fields and to *define constraints* over them, such as a permissible value range, are other useful features that would reduce the cognitive overload and prevent erroneous input by guiding the user during data entry.

Logic level. A data entry can *encapsulate* multiple data fields. In an oesophagus radiology report (ORP), the size of the first ulcerated lesion may be defined in an interval by specifying its lower and upper bounds. Furthermore, there may exist dependencies and relations between data entries; due to the direct relationship between the amount and complexity of information that need to be entered/processed by the user and the cognitive load, *reducing the amount and complexity of information would also reduce the cognitive load*. For example, the shape of the mucosal relief and the narrowness of the oesophagus are typical elements of an ORP. However, in a particular case only a subset of this information is actually relevant. If there isn't any narrowness in oesophagus during the transition of the contrast media, then the mucosal relief should be entered; if the mucosal relief is irregular then the shape of the irregularity should also be specified. In case there is a narrowness, mucosal relief is not important and a completely different set of information should be entered including and depending on the properties of the narrowness. Note that, this inherently leads to a *nested and hierarchical* structure, in which data entered at a certain point determines the information flow. By *interactively walking on the necessary steps* while completing the report, the number of data entries can be reduced considerably. This hierarchical structure is not specific to this particular example, and emerges as a common feature of almost all kinds of medical reports. Furthermore, as several sources point out, in most cases medical reports belong to normal cases in which there are only few fields with abnormal values depending on the case under examination. Ideally, much less time has to be spent to record normality, and for the sake of cognitive simplicity the user should not receive data entries related to abnormal situations; this can simply be achieved by conducting an initial simulated walk on the necessary steps using the default values for the normal cases. This forementioned dynamism can be realized by assigning triggers to data entries, defined in terms of boolean expressions.

Presentation level. Data and logic level can be regarded as the backend that defines the structure of the report; presentation level, on the other hand, is the frontend that defines how the report is rendered for data collection and viewing. The separation of presentation from data and logic would enable to generate different views of the same data based on user requirements; this also brings support for report generation in multiple languages without requiring natural language processing methods, which are not reliable and liable to medical errors.

Now, starting from the data level we will describe the SISDS method and discuss how it possesses the features listed so far. The building block in SISDS is a data field, or *variable*, defined by a tuple $\langle var, type, val, opts \rangle$ where var is the name and $type$ is the type of the variable, val is its initial value, and $opts$ is a set of pairs of the form $\{\langle name_1, val_1 \rangle, \ldots\}$ where $name_i$ denotes the name of the i^{th} option and val_i is its value; typical options include the minimum, maximum and normal values of a variable. $type$ is either one of predefined types or if it is a nominal variable it is a set of possible values, ex. $\{yes, no\}$. For measurement data types, such as length, the initial value should also contain the unit of measurement, ex. $1.2cm$. A *data entry* is a unit of data request and encapsulates one or more variables; it is defined by a tuple $\langle label, vars, defs, triggers \rangle$ where $label$ is a unique identifier denoting the data entry, $vars$ is a set of variable definitions, $defs$ is a set of data request/view definitions (DRVDs) and $triggers$ is a set of triggers that activate related data entries. Each trigger in $triggers$ is a pair of the form $\langle cond, action \rangle$ where $cond$ is a boolean expression with embedded variable references and $action$ specifies an action to be executed when the condition holds, i.e. boolean expression evaluates to true. The boolean expression may include arithmetic and logic operators, function calls, constants and variables references. The variable references in the boolean expression are of the form $\langle label, var \rangle$ where var is the name of the variable, $label$ is the identifier of the data entry that the variable belongs to or $\emptyset$ if it belongs to the current data entry. While evaluating the boolean expression, the variable references are replaced with the current value (default or that entered by the user) of the corresponding variables. Note that, the values of the variables with measurement data types must be normalized, i.e. converted into a common unit, before evaluation since the unit of such variables may be altered by the user. This can be done by calling a unit conversion function within the condition expression. An *action* can be a set of labels that denote the data entries to be activated, a message to be displayed, or a diagnosis prediction; it is important to note that cyclic activations are not allowed, that is a descendant of a data entry cannot re-activate it. Each DRVD is a tuple of the form $\langle type, lang, def \rangle$ where $type$ denotes the type of the DRVD, $lang$ denotes the language of the definition, and def is the body of the definition. The body of the definition is an arbitrary string with embedded variable references of the form $\langle label, var, vals, opts \rangle$ where $label$ and var are defined as above, $vals$ is a set of mappings for nominal variables to map possible values of the variable to string counterparts, and $opts$ is a set of options as in the definition of variables. Typical options include format specifiers to determine the rendering of the variable. DRVDs are used by the presentation

layer to render data entry forms or reports based on their *type*; this gives rise to a unified view in which data collection and viewing are handled similarly.

Finally, a *report* is tuple $\langle E, M, triggers \rangle$ where E is a set of consistent data entries, that is all data entries referred in the trigger conditions of these data entries exists in the report (i.e. are in E), M is an ordered list of data entry identifiers denoting the *main data entries*, and *triggers* is a set of report-wide triggers; for each identifier in M there must be a corresponding data entry in E. The main data entries constitute the initial skeleton of the report. The report-wide triggers enable to provide diagnosis and other suggestions to the user based on the entered data. From a conceptual point of view, our structured design with interactivity looks like a tree with branches growing from a stem such that the branches collapse and expand as needed, main data entries being the initially expanded branches. A dynamic hierarchy of *sections* is built as related data entries logically follow-up depending on the defined conditions. This effectively enables the user to focus on problematic parts and record them in more detail while eliminating other parts to save time, thus avoiding inefficiency and cognitive overload.

3 Design and Implementation of a Prototype

In order to verify the eligibility of the proposed approach, we implemented a Web-based prototype based on the client-server architecture. We opted to use a human-usable textual notation with a simple syntax to realize the abstract variable and data entry definitions given in the previous section; the trigger conditions are also defined using this notation. The main novelty of this particular implementation is a natural free-text like data entry such that the entered data directly corresponds to the content of the final product (i.e. report). One way to ensure this in structured data entry is to let the user see the resulting report while *still entering data*. The solution that we offer is to use inline editing, that is to present the report in a single view but allow the users to directly manipulate the data on the screen simply by clicking on data fields which are displayed as hyperlinks. As the user changes the values of variables, the contents of the report is also rearranged automatically (and notifications are displayed) according to defined trigger conditions. This not only prevents the cognitive overload, but also unifies the data entry and viewing phases[1].

According to some studies about visual cognition, under normal viewing conditions only a minor part of the environment is encoded in detail [6]. Sometimes professionals could not see other pertinent details while concentrating on a specific subject. In order to prevent this, in our implementation the presentation layer is enriched with visual clues; data fields having abnormal values or yet to be entered are automatically highlighted in different ways to warn the user and draw his attention to those sections of the report. We also enabled the user to temporarily hide data entries that are not directly related with a selected data

[1] The syntax of the textual notation and more information about the free-text data entry, including screen-shots, can be found in the technical report [5]. A demo version of the prototype is also available online at http://www.gata.edu.tr/mebs/sisds

entry (i.e. show only selected data entry together with its descendants and those that are involved in the activation of this data entry). The feedback that we received from initial deployment of the system suggests that users find both features effective and useful. Aside free-text like data entry, by taking advantage of the separation of data from its representation the prototype also supports data entry in the form of an enumerated list and additional formats can be added with ease. These are just different representations of the same data, albeit with different cognitive properties; even though the first one is more natural, the enumerated list may be more convenient and preferable in certain cases.

4 Discussion and Conclusion

In this study we propose a new methodology which adopts a systematic approach to improve medical processes by reducing variability and minimizing errors. More specifically, we focus on the process of data entry and report generation. The interactivity with the user in our study, "interactive walk on necessary steps", has many advantages that allow information to be captured at the point of care and eliminate the need for a transcriptionist or auxiliary procedures to write reports. In particular, the end report is automatically generated while structured fields are filled interactively in a natural form which is similar to the final report; it also provides (i) a high degree of timeliness and accuracy, reducing errors, (ii) multifunctional capabilities such as drawing the attention of practitioner to important sections of the report, alerting him about a diagnosis or giving advises at the time of entry, and (iii) an easy way for domain experts to define reports in a textual form without extensive computer knowledge. The initial feedback that we received from the users of the prototype implementation indicates that the proposed method is a promising approach for achieving the aim of effective data collection and reporting. Further studies will concentrate on a wide-scale deployment of the system, and development and integration of a medical decision support system based on the collected data.

References

1. Sistrom, C.L., Langlotz, C.: A framework for improving radiology reporting. J. Am. Coll Radiol. 2(1), 61–67 (2005)
2. Sistrom, C.L.: Conceptual approach for the design of radiology reporting interfaces: The talking template. Journal of Digital Imaging 18(3), 176–187 (2005)
3. IOM06: Preventing Medication Errors: Quality Chasm Series. Institute of Medicine of the National Academies Press (July 2006)
4. Naik, S., Hanbidge, A., Wilson, S.: Radiology reports: examining radiologist and clinician preferences regarding style and content. American Journal of Roentgenology 176(3), 591–592 (2001)
5. Kuru, K., Girgin, S., Arda, K.: A novel multilingual report generation approach for medical applications:the sisds methodology. Technical Report METU-MIN-TR-2009-001-KK, Infomatics Inst (2009)
6. Noë, A., Pessoa, L., Thompson, E.: Beyond the grand illusion: What change blindness really teaches us about vision. Visual Cognition 7(2), 93–106 (2000)

CORAAL – Towards Deep Exploitation of Textual Resources in Life Sciences

Vít Nováček, Tudor Groza, and Siegfried Handschuh

Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway, Ireland
FirstName.LastName@deri.org

Abstract. Prominent biomedical literature search tools like ScienceDirect, PubMed Central or MEDLINE allow for efficient retrieval of resources based on key words. Due to vast amounts of data available in life sciences, key word search is not always sufficient, though. One would often welcome more intelligent search for knowledge, i.e., for concepts and their mutual relations. This is, however, still a major challenge, since getting the necessary machine-readable knowledge manually is virtually impossible in large scale, while its automatic extraction is not particularly reliable. We have researched a novel framework actually enabling practical exploitation of automatically extracted knowledge, though. On the top of the framework, we implemented CORAAL, a prototype for knowledge-based biomedical literature search. This paper describes its essential principles, innovative capabilities and current results.

1 Introduction

Digital content processing has no doubt introduced a whole lot of new possibilities of dealing with scientific publications. It makes knowledge much more open and exploitable than in the old "paper times". However, one still needs to go manually through a lot of possibly irrelevant content very often before actually finding the right answers. If we are to make the next step, it is necessary to process knowledge (i.e., concepts and their mutual relations), and not just data or shallow meta-data (i.e., chunks of free text, titles or author names). Substantial automation of such meaning-intensive information processing is hardly possible with the current industry-strength technologies (e.g., full-text search), since they lack proper support for extraction, representation and processing of knowledge implicitly present in texts. As an illustration, imagine for instance finding a support of the claim that *acute granulocytic leukemia* is different from *T-cell leukemia*. With the current solutions, it is easy to find articles that contain both or either of the terms, however, the number of results may be quite high (e.g., 593 on PubMed). It is tedious or even impossible to go through all of them in order to find out which of them actually mention the two leukemias being different.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 206–215, 2009.

Methods for automated knowledge extraction than can dig more than mere key words from text exist, however, their results are deemed to be to too noisy and sparse to be exploited by the current state of the art without significant manual post-processing [1]. We have recently researched a novel framework for effortless exploitation of automatically extracted knowledge that makes use of similarity-based knowledge representation and respective light-weight inference services [2]. We combined the framework with our repository for semantically inter-linked publications [3], delivering a prototype knowledge-based publication search engine – CORAAL (*COntent extended by emeRgent and Asserted Annotations of Linked publication data*). The tool essentially *extracts* asserted publication meta-data together with the knowledge implicitly present in the respective text, *integrates* the emergent content with existing domain knowledge and *exposes* it via a multiple-perspective search&browse interface. This way we allow for fine-grained publication search combined with convenient and effortless large scale exploitation of the knowledge associated with and hidden in the publication texts.

The rest of the paper is organised as follows. Section 2 describes the data used in the current CORAAL deployment, as well as the tool's essential technological principles and capabilities. Section 3 reports on experiments assessing the applicability of CORAAL and quality of the knowledge served to users. Related work is analysed in Section 4. We discuss the potential of the delivered work, conclude the paper and outline future directions in Section 5.

2 Method

Here we describe the data processed by CORAAL, and summarise the essential technical principles of the prototype.

2.1 Inputs and Outputs

Input. As of March 2009, we have processed 11,761 Elsevier journal articles from the provided XML repositories that were related to cancer research and treatment. The access to the articles was provided within the Elsevier Grand Challenge competition (cf. `http://www.elseviergrandchallenge.com`). The domain was selected so due to the expertise of our sample users and testers from Masaryk Oncology Institute in Brno, Czech Republic. We processed articles evenly distributed across the journals in the following list: 1. *FEBS Letters*; 2. *Biochemical Pharmacology*; 3. *Cancer Genetics and Cytogenetics*; 4. *Cell*; 5. *Trends in Cell Biology*; 6. *Experimental Cell Research*; 7. *Controlled Clinical Trials*; 8. *Molecular Aspects of Medicine*; 9. *Advanced Drug Delivery Reviews*; 10. *Gene*; 11. *Trends in Genetics*; 12. *Genomics*; 13. *Leukemia Research*; 14. *Journal of Microbiological Methods*; 15. *Trends in Microbiology*; 16. *Journal of Molecular Biology*; 17. *Oral Oncology*; 18. *European Journal of Pharmacology*. From the article repository, we extracted the knowledge and publication metadata for further processing by CORAAL. Besides the publications themselves, we employed legacy machine-readable vocabularies for the refinement and extension of

the extracted knowledge (currently, we use the NCI and EMTREE thesauri – see `http://www.cancer.gov/cancertopics/terminologyresources` and `http://www.embase.com/emtree/`, respectively).

Output. CORAAL exposes two data-sets as an output of the publication processing: (1) We used a **triple store** containing publication meta-data (citations, their contexts, structural annotations, titles, authors and affiliations) associated with respective full-text indices. The resulting store contained $7,608,532$ of RDF subject-predicate-object statements [4] describing the input articles. This included $247,392$ publication titles and $374,553$ authors (both from full-texts and references processed). (2) We employed a custom EUREEKA **knowledge base** [2] with facts of various certainty extracted and inferred from the article texts and the seed life science thesauri. Directly from the articles, $215,645$ concepts were extracted (and analogically extended). Together with the data from the initial thesauri, the domain lexicon contained $622,611$ terms, referring to $347,613$ unique concepts. The size of the emergent knowledge base was $4,715,992$ weighed statements (ca. 99 and 334 extracted and inferred statements per publication in average, respectively). The contextual meta-knowledge related to the statements, namely provenance information, amounted to more than $10,000,000$ additional statements (should it be expressed in RDF triples). Query evaluation on the produced content takes usually fractions and at most units of seconds.

2.2 Core Technologies and Capabilities

The publications, their meta-data and full-text were stored and indexed within our KONNEX framework for linked publication data processing [3]. After parsing the input XML article representations, the XML meta-data and structural annotations were quite straightforwardly integrated in the KONNEX RDF repository. Full-text information regarding the articles' content, titles, authors and references were managed using multiple Lucene IR indices (cf. `http://lucene.apache.org/java/docs/`).

Exploitation of the publication knowledge was tackled by our novel EUREEKA framework for emergent (e.g., automatically extracted) knowledge processing [2]. The framework de facto builds on a simple triple model [4], a widely-used part of the Semantic Web [5] standards. However, we extended the subject-predicate-object triples by positive or negative heuristic certainty measures and organised them in so called conceptual matrices, concisely representing every positive and negative relation of an entity to other entities. Metrics can be easily defined on the conceptual matrices. The metrics then serve as a natural basis for gradual concept similarities that define basic light-weight empirical semantics in EUREEKA [2]. On the top of the similarity-based semantics, we implemented simple, yet quite practical inference services of two basic types: 1. *retrieval* of knowledge similar to an input concept, and/or its *extension* by means of similar stored content; 2. fixed-point rule-based *materialisation* of implicit relations, and/or complex *querying* (similarity as a basis for soft variable

unification and for approximate fixed-point computation). The inference algorithms have anytime behaviour and it is possible to programmatically adjust their completeness/efficiency trade-off. Technical details of the solution are out of scope regarding this paper, but one can find them in [2].

We applied our prototype to: (i) automated extraction of machine-readable knowledge bases from particular life science article texts; (ii) integration, refinement and extension of the extracted knowledge within one large emergent knowledge base; (iii) exposure of the processed knowledge via a query-answering and faceted browsing interface, tracking the article provenance of particular statements.

For the initial knowledge extraction, we used a NLP-based heuristics stemming from [6,7] in order to process chunk-parsed texts into subject-predicate-object-score quads. The scores were derived from aggregated absolute and document frequencies of subject/object and predicate terms. The extracted quads encoded three major types of ontological relations between concepts: (1) taxonomical—*type*—relationships; (2) concept difference (i.e., negative *type* relationships); (3) "facet" relations derived from verb frames in the input texts (e.g., *has part*, *involves* or *occurs in*). About $27,000$ facet relations were extracted. A taxonomy was imposed on them, considering the head verb of the respective phrase as a more generic relation (e.g., *involves expression of* was assumed to be a type of *involves*). Also, several artificial concepts were introduced to restrict the semantics of some most frequent relations. Namely, (positive) *type* was considered transitive and anti-symmetric, and *same as* was set transitive and symmetric. Also, *part of* was assumed transitive and inverse of *has part* for the current deployment. Note that the *has part* relation has rather general semantics within the extracted knowledge, i.e., its meaning is not strictly physically mereological, it can refer also to, e.g., conceptual parts or possession of entities.

The emergent quads were processed as follows:
(I) *addition* – The extracted quads were incrementally added into an emergent knowledge base K, using a fuzzy aggregation of the respective conceptual matrices. As a seed defining the basic domain semantics (i.e., synonymy and core taxonomy of K), we used the EMTREE and NCI thesauri.
(II) *closure* – After the addition of new facts into K, we computed its materialisation according to RDFS entailment rules [8] ported to the format specified in [2].
(III) *extension* – All the extracted concepts were analogically extended by means of similar stored knowledge.

We exposed the content of the eventual knowledge base via a query-answering module. It was returning answer statements sorted according to their relevance scores [2] and similarity to the query. Answers were provided by intersection of publication provenance sets corresponding to the respective statements' subject and object terms. The module supported queries in the following form: $t \mid s : (NOT\ )?p : o(\ AND\ s : (NOT\ )?p : o)^*$, where NOT and AND stands for negation and conjunction, respectively. s, o, p may be either variable—anything starting with the `?` character or even the `?` character alone—or a lexical expression.

t may be lexical expressions only. The ? and * wildcards mean zero or one and zero or more occurrences of the preceding symbols, respectively, | stands for or. Only one variable name is currently allowed to appear within a single query statement and across a statement conjunction.

Example queries and respective selected answers are as follows:

`QUERY: ? : type : breast cancer` $\rightsquigarrow$ `ANSWER: <cystosarcoma phylloides : TYPE : breast cancer>`^{1} `...`
`QUERY: rapid antigen testing : part of : ? AND ? : type : clinical study` $\rightsquigarrow$ `ANSWER: <dicom study : USE : protein info>`$^{0.8}$ `AND <initial study : INVOLVED : patients>`$^{0.9}$ `...`
`QUERY: acute granulocytic leukemia : NOT type : T-cell leukemia` $\rightsquigarrow$ `ANSWER: <acute granulocytic leukemia : TYPE : T-cell leukemia>`$^{-0.7}$ `...`

The sample answers above are presented in the statement syntax specified in [2] (with rounded degrees). In CORAAL itself, the statements are presented in more human readable way, very similarly to the query syntax. They are also provided by the following types of meta-information: (1) *source* provenance – articles relevant to the statement; (2) *context* provenance – sub-domain of life sciences the statement relates to (determined according to the main topic of the journal that contained the articles the statement was extracted from); (3) *certainty* – a real number meaning how certain the system is that the statement holds and is relevant to the query (values between 0 and 1; derived from the absolute value of the respective statement degree and from the actual similarity of the statement to the query); (4) *inferred* – a boolean value determining whether the statement was inferred or not (i.e., directly extracted).

More can be checked out at `http://coraal.deri.ie:8080/coraal` (points to an online interface of CORAAL deployed on the sample cancer research publication data).

3 Experiments and Evaluation

This section reports on a user-based applicability test of CORAAL and an experiment aimed at assessment of the exposed knowledge quality.

3.1 Applicability Tests with Experts

We prepared five tasks[1] to be worked out with both CORAAL and a base-line application (ScienceDirect or PubMed) by four sample users. Our hypothesis was that the users should perform better with CORAAL than with the base-line, since the tasks were focused rather on structured knowledge than than on a plain text-based search.

[1] E.g., find all authors who support the fact that the `acute granulocytic leukemia` and `T-cell leukemia` are different.

Using a questionnaire and additional structured interview, we evaluated three major features: (i) the degree to which the queries were considered realistic by the users; (ii) the number of successfully accomplished parts of particular tasks; (iii) the usability. The tasks were deemed rather realistic – the average value was above 4 on the scale from 1 to 6 (worst to best). The success rate of the task accomplishment was 60.7% and 10.7% when using CORAAL and the base-line application, respectively. This clearly confirms our hypothesis regarding improvement over the state-of-the-art. Still, users experienced a lot of frustration related to tasks they were not able to solve with CORAAL. Most sources of the frustration were eliminated by development of a new, better integrated and more intuitive user interface. Further improvements in the user performance were achieved after brief interactive educational sessions. In the beginning, users were just let to play, relying only on an online tutorial. For users given a short interactive lecture about the general features of the CORAAL user interface and query language, the performance was about 75% better and the frustration diminished accordingly.

3.2 Quality of the Exposed Knowledge

We evaluated quality of representative sample answers provided by CORAAL on the cancer research publication data-set. To do so, we picked 100 random concepts and generated 100 random statement queries based on the actually extracted content. We let a committee of domain experts vote on the relevance of respective concept and statement queries to their day-to-day work and used the following most relevant ones to evaluate the CORAAL answers:

Concept queries: `myelodysplastic syndrome`; `p53`; `BAC clones`; `primary cilia`; `colorectal cancer`

Statement queries: `? : type : breast cancer`; `? : part of : immunization`; `? : NOT type : chronic neutrophilic leukemia`; `rapid antigen testing : part of : ? AND ? : type : clinical study`; `? : as : complementary method AND ? : NOT type : polymerase chain reaction`

Table 1. Precision/recall results summary

Q. type/measure	P_s	R_s	F_s	P_d	R_d	F_d
concepts	0.474	0.143	0.183	0.496	0.154	0.234
concepts (base)	0.591	0.031	0.056	0.405	0.061	0.102
statements	0.719	0.583	0.586	0.704	0.489	0.541
statements (base)	0.169	0.053	0.067	0.216	0.145	0.171

We used the traditional notions of precision and recall for the answer quality evaluation, with average results summed up in Table 1.

For a base-line comparison, we employed state-of-the-art Semantic Web technologies – crisp RDFS inference [8] and SPARQL querying[2] on the same data as processed by CORAAL (setting degrees to 1.0 and omitting negative statements, though, since neither RDFS nor SPARQL support uncertainty and negation).

P, R, F in Table 1 columns stands for precision, recall and F-measure (computed as $\frac{2(P \cdot R)}{P+R}$), respectively. The s and d subscripts indicate retrieved *statement* and corresponding *provenance document* precision (or recall), respectively. Base-line results for concept and statement queries are given in the respective *base* lines. Particular precision/recall values were computed as follows. Let C be the corpus of the publications processed by CORAAL. $P_s = \frac{CSR}{ASR}$, $R_s = \frac{CSR}{CSA}$, where CSR, ASR is a number of correct and all answer statements returned by CORAAL, respectively. CSA is the number of all correct statements relevant to the query, as entailed by C data. $P_d = \frac{RDR}{ADR}$, $R_d = \frac{RDR}{RDA}$, where RDR, ADR is a number of relevant and all correct statement provenance publications returned, respectively. RDA is the number of all publications in C relevant to the query and its correct answers.

The degrees in the answer statements were taken into account in this way: if their absolute value was lower than 0.5, i.e., indicating substantial lack of heuristic confidence, the respective statement was deemed neither correct, nor incorrect, and was not considered in the precision/recall computation. Statements originating solely from the initial thesauri were discarded, too. First 400 results were only examined when more eligible answers were available. The results' relevance and numbers of the gold-standard statements and/or publications were determined by domain experts. They did so in a detailed analysis of the C article corpus via a full-text search. They examined both explicit and implicit knowledge in the paragraph contexts of the query and answer terms, as well as in the related NCI and EMTREE thesauri entries. Unequivocal agreement of evaluators was required at all times.

In terms of F-measure, CORAAL clearly outperformed the base-line. The difference was more than two and three-fold regarding F_s for concept and statement queries, respectively. Similarly, F_d was more than eight and three times higher. The base-line precision was higher for P_s and concept queries, though. This was caused by the absence of (partially incorrect) negative statements in the base-line results. On the other hand, recall of CORAAL was much higher due to approximate answer retrieval, and also due to the presence of negative and analogically inferred relations. CORAAL's precision for statement queries was higher due to the support for soft evaluation of both rules and queries – some incorrect crisp statements computed by the base-line were filtered out in CORAAL due to low certainty either in the intermediate, or in the eventual result. Generally better results for statement queries were caused by the fact that only statements directly related to the variable instances conforming to the query structure were taken into account. For concept-only queries, all resulting statements were considered.

The CORAAL results may still be considered rather poor when compared to the gold standard (i.e., F-measure for concept queries around 0.2). However, one

[2] Cf. `http://www.w3.org/TR/rdf-sparql-query/`

must realise that the construction of the gold standard took two working days of an expert committee only for the 10 sample queries. The CORAAL knowledge base was produced in about the same time for much larger amount of data. Using the faceted browsing provided by the CORAAL user interface, one can find relevant answers very quickly despite of some remaining noise in the purely automatically acquired knowledge. This is a reasonable and unprecedented trade-off according to our expert evaluators and potential users.

4 Related Work

Approaches tackling problems related to those addressed by the core technologies powering CORAAL are analysed in [2,3]. Here we offer an overview of systems targeting similar problems to those tackled by our framework. Figure 1 organises relevant applications in a plot with two axes – *effort* and *benefit* (the placement is only orientational, though, as it does not reflect any formal measure related to the particular systems). The *effort* axis indicates how much more or less manual effort must the creators and/or maintainers of a tool spend before it can perform sufficiently, or before it can be ported to a new domain. The *benefit* axis reflects how much benefit users get when searching for the knowledge hidden in publications with a tool.

The state-of-the-art applications like ScienceDirect or PubMed Central require almost no effort in order to expose arbitrary life science publications for search (therefore we used them as a base-line in the user-centric experiment). However, the benefit they provide is rather limited when compared to cutting-edge approaches aimed at utilising also the publication knowledge within the query construction and/or result visualisation. Such innovative solutions may require much more a priori effort in order to work properly, though.

FindUR [9], Melisa [10] and GoPubMed [11] are ontology-based front-ends to a traditional publication full-text search. They allow either for effective restriction and intelligent visualisation of the query results (GoPubMed), or for focusing the queries onto particular topics based on an ontology (FindUR and Melisa). FindUR and Melisa use a Description Logics [12] ontology built from

Fig. 1. Informative comparison of selected systems

scratch and a custom ontology based on MeSH (cf. `http://www.nlm.nih.gov/mesh/`), respectively. GoPubMed dynamically extracts parts of the Gene Ontology (cf. `http://www.geneontology.org/`) relevant to the query, which are then used for restriction and a sophisticated visualisation of the classical PubMed search results. None of the tools, nevertheless, offers querying for or browsing of arbitrary publication knowledge – terms and relations not present in the systems' rather static ontologies simply cannot be reflected in the search. On the other hand, CORAAL works on any domain and extracts arbitrary knowledge from publications automatically, although the offered benefits may not be that high due to possibly higher level of noisiness.

Textpresso [13] is quite similar to CORAAL concerning searching for relations between concepts in particular chunks of text. However, the underlying ontologies and their instance sets have to be provided manually, whereas CORAAL can operate with or even without any legacy ontology. Moreover, the system's scale regarding the number of publications' full-texts and concepts covered is much lower than for CORAAL.

From the overview of the related cutting-edge systems, it is obvious that the biggest challenge is a reliable automation of more expressive content acquisition. Contrary to CORAAL, none of the related systems addresses this problem appropriately, which makes them either poorly scalable, or difficult to port to a new domain. This is why we were not even able to use the related systems for a base-line comparison in our domain-specific application scenario – we simply could not adapt them so that they would be able to perform reasonably, both due to technical difficulties and lack of necessary human/time resources.

5 Discussion

In this paper, we have presented CORAAL – a unique combination of a publication repository enhanced by semantic links [3] and an engine for automated extraction, integration and exploitation of knowledge contained in the publication texts [2]. We have shown that the tool has promising results in real-world tasks related to biomedical literature search. Due to substantial automation, we are able to process large amounts of publications in more scalable and efficient way than possible with the state of the art. The potential of CORAAL has also recently been proven by the fact that it was selected as one of the four Elsevier Grand Challenge finalists (cf. `http://www.elseviergrandchallenge.com`).

Note that besides the presented application to literature search, CORAAL can directly be deployed in any use case involving the need for more efficient search in large amounts of textual data. For instance, one could deploy CORAAL in a hospital and feed it with patient records. Appropriate medical ontologies and/or diagnostic rules can be imported into CORAAL to support additional refinement and inference within the patient data. The knowledge scattered among large amounts of patient records can then be integrated and exposed in the same intelligent way as presented in this paper.

Despite of the promising results, there are still certain reserves. We plan to extend the current knowledge processing framework powering CORAAL to a

distributed solution, which will significantly improve scalability (from tens of thousands to millions of publications and beyond). In order to complement our automated approach by the wisdom of the crowds, we have to propose sound mechanisms for easy user involvement in the emergent knowledge (in)validation, updates, and general maintenance. Last but not least, we intend to continue in our cooperation with various groups of biomedical experts, who will help us to realise the CORAAL's promise in agile R&D settings.

Acknowledgments. This work has been supported by the 'Líon', 'Líon II' projects funded by SFI under Grants No. SFI/02/CE1/I131, SFI/08/CE/I1380, respectively. We acknowledge much appreciated help from Ioana Hulpus, who developed the initial user interface for CORAAL. Eventually, we are very grateful to our evaluators: Doug Foxvog, Peter Gréll, MD, Miloš Holánek, MD, Matthias Samwald, Holger Stenzhorn and Jiří Vyskočil, MD.

References

1. Bechhofer, S., et al.: Tackling the ontology acquisition bottleneck: An experiment in ontology re-engineering (2003), http://tinyurl.com/96w7ms (April 2008)
2. Nováček, V.: Towards an efficient knowledge-based publication data exploitation: An oncological literature search scenario. Technical Report DERI-TR-2009-03-23, DERI, NUIG (2009), http://tinyurl.com/csh3rf
3. Groza, T., Handschuh, S., Möller, K., Decker, S.: KonneXSALT: First steps towards a semantic claim federation infrastructure. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 80–94. Springer, Heidelberg (2008)
4. Manola, F., Miller, E.: RDF Primer (2004), http://www.w3.org/TR/rdf-primer/ (November 2008)
5. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. 5 (2001)
6. Maedche, A., Staab, S.: Discovering conceptual relations from text. In: Proceedings of ECAI 2000. IOS Press, Amsterdam (2000)
7. Voelker, J., Vrandecic, D., Sure, Y., Hotho, A.: Learning disjointness. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 175–189. Springer, Heidelberg (2007)
8. Brickley, D., Guha, R.V.: RDF Vocabulary Description Language 1.0: RDF Schema (2004), http://www.w3.org/TR/rdf-schema/ (Feburary 2006)
9. McGuinness, D.L.: Ontology-enhanced search for primary care medical literature. In: Proceedings of the Medical Concept Representation and Natural Language Processing Conference, pp. 16–19 (1999)
10. Abasolo, J.M., Gómez, M.: Melisa: An ontology-based agent for information retrieval in medicine. In: Proceedings of the First International Workshop on the Semantic Web (SemWeb 2000), pp. 73–82 (2000)
11. Dietze, H.: et al.: Gopubmed: Exploring pubmed with ontological background knowledge. In: Ontologies and Text Mining for Life Sciences, IBFI (2008)
12. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F.: The Decription Logic Handbook: Theory, implementation, and applications. Cambridge University Press, Cambridge (2003)
13. Müller, H.M., Kenny, E.E., Sternberg, P.W.: Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biology 2(11) (2004)

Detecting Intuitive Mentions of Diseases in Narrative Clinical Text

Stéphane M. Meystre

Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA
stephane.meystre@hsc.utah.edu

Abstract. A significant portion of the clinical information content of narrative text documents in the medical record is only mentioned intuitively, but automated information extraction systems typically focus on explicitly mentioned concepts only. To extend the extraction of clinical information to intuitively mentioned diseases, we have developed a natural language processing application based on MMTx and on context analysis algorithms, enhanced with the detection of disease-specific concepts (e.g. medications used only for this disease), and values of some specific biomarkers. This application was developed for the i2b2 obesity challenge, a competition focused on the detection of patients with obesity or common comorbidities.

1 Introduction

The automated extraction of information from clinical text documents has been the focus of several competitions organized by the i2b2 (Informatics for Integrating Biology and the Bedside) National Center for Biomedical Computing. These challenges took the recent application of Natural Language Processing (NLP) to clinical research a step further by providing a de-identified corpus of clinical narrative text documents and by stimulating new developments in this domain. The last challenge was organized in 2008 and focused on the identification of patients with obesity and/or some of its most common comorbidities. Textual mentions of these diseases were annotated, but also intuitive mentions of them, and we focused this development on the latter. For this challenge, we built a new NLP application based on a system that had been developed to automatically maintain the electronic problem list [1]. This system automatically extracted potential medical problems from all narrative text documents in the electronic medical record, and proposed them for inclusion in the problem list. The new NLP application called Textractor that evolved from this system, and its evaluation with the obesity challenge reference standard, are presented here.

2 Background

Information extraction methodologies are typically applied to extract information from biomedical text (i.e. scientific publications), and less frequently from

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 216–224, 2009.

clinical text, as described in Meystre et al. [2] A substantial part of the medical record is made of narrative clinical text documents that represent patient history and reports of therapeutic interventions or clinical progress [3]. Decision-support, research, the optimization of database operations, and improvement in medical administration create a need for coded data instead. As a possible answer to this issue, Natural Language Processing can convert free text into coded data. Techniques for automatically encoding textual documents from the medical record have been evaluated by several groups. Examples are the Linguistic String Project [4], and MedLEE (Medical Language Extraction and Encoding system) [5]. Other systems automatically mapping biomedical text concepts to standardized vocabularies have been reported, such as MetaMap [6]. MetaMap and its Java version called MMTx (MetaMap Transfer) were developed by the U.S. National Library of Medicine. MetaMap has been shown to identify most concepts present in MEDLINE titles [7] and has been used for Information Retrieval [8] and Information Extraction [9-11]. When extracting information from narrative clinical text documents, the context of the extracted concepts plays a critical role. Important contextual information includes negation (e.g. "denies any chest pain"), temporality (e.g. "...fracture of the tibia 2 years ago..."), and the event subject identification (e.g. "his mother has diabetes"). NLP systems such as the LSP [4] or MedLEE [5] include negation analysis in their processing, but research focused explicitly on negation detection started only a few years ago with algorithms like NegExpander [12] or NegEx [13]. Temporality analysis in clinical narrative text can be significantly more complex than negation analysis, and has been investigated by several teams including Zhou, Hripcsak et al.[14] and Bramsen et al.[15]. Finally, algorithms combining the analysis of the subject of the text (e.g., the patient) and other contextual features have recently been developed and evaluated. A good example is ConText, proposed by Chapman et al.[16]. This algorithm is an extension of NegEx cited above and determines the values of three contextual features: negation, temporality, and experiencer.

The i2b2 challenges started in 2006, with an automated de-identification challenge [17] and a smoking status detection challenge [18]. In 2008, the i2b2 Obesity Challenge focused on identifying obese patients and the 15 best-represented comorbidities they exhibit, based on narrative clinical text from their medical record. A corpus of 1237 clinical text documents from patients evaluated for obesity or diabetes has been semi-automatically de-identified and re-identified with realistic surrogates (i.e. identifying information was replaced by made up information that resembled the original information), and then split in a training corpus of 730 documents, and a test corpus of 507 documents. The reference standard was built by two obesity experts from the Massachusetts General Hospital who annotated all documents, and a third resident adjudicated their disagreements for the textual annotations (i.e. strictly based on text). Intuitive annotations (i.e. based on implicit information) experts disagreed on were removed from the corpus. The experts agreement was good (average Cohens κ coefficient of 0.8606 for textual annotations and 0.7642 for intuitive annotations).

3 Methods

For this experiment, a new NLP application called Textractor was developed. This application is based on the Automated Problem List system [1,19,20]. Preprocessing starts with sections and sentences detection, using regular expressions and a set of rules. Before passing sentences for concepts extraction, some disambiguation is required. We use MMTx, an application originally developed to analyze MEDLINE abstracts. Acronyms are less common in paper abstracts than in clinical documents, and are the principal source of ambiguity for our system. Examples of acronyms ambiguous to MMTx are Dr. (detected as diabetic retinopathy), Mr. (mitral regurgitation), M.D. (mental depression), PA (pernicious anemia), etc. In our case, disambiguation consists in expanding these acronyms. Another source of ambiguity is linked to a lack of local context and results in errors like detecting Depression in ST segment depression. The subsequent information extraction stage works in two steps: a first step using MMTx (version 2.4.C) to extract UMLS Metathesaurus concepts, and a second step to infer the context of each of those concepts (negation, temporality, and subject). The first step is based on the new MMTxAPILite class with the default dataset (complete 2006 UMLS Metathesaurus) and settings, as well as on a manually built mapping table linking obesity and its 15 comorbidities with all related subconcepts (e.g. Hypertriglyceridemia was mapped to primary hypertriglyceridemia, secondary hypertriglyceridemia, blood triglycerides increased, etc.). This table was built by adding all concepts with a CHD (Child), SIB (Sibling), RN (Narrower), and some concepts with a RL (Similar) and RQ (Related) relationships, and then performing a manual review to remove irrelevant concepts. Since MMTx lacks context analysis (e.g. Diabetes will be extracted in "No diabetes is reported in the patients family history"), the second context analysis step is required. Context analysis is based on ConText [16]. This algorithm uses regular expressions and lists of terms to analyze negation (a concept can be affirmed, negated, or questionable), temporality (recent, historical, or hypothetical), and the experiencer (patient or other). Finally, post-processing includes some disambiguation and the adjudication of the contextual analysis of each mentions of a same detected concept.

The intuitive annotations pose a new challenge: detecting diseases that are not clearly mentioned in text, that are only implicitly mentioned. For this purpose, we added two information extraction enhancements: one is based on a list of keywords that are specific to a certain disease (examples in Table 1), and the other one is based on biomarker values like the plasma triglycerides concentration to detect hypertriglyceridemia. The list of keywords was manually built from publicly available medical knowledge sources and include generic and commercial names of medications used only to treat a specific disease (e.g. allopurinol to treat gout), names of diagnostic or therapeutic procedures proper to a specific disease (e.g. fundoplication to treat a gastro-esophageal reflux), and other therapeutic means typically used with a specific disease (e.g. compression stockings in a case of venous insufficiency). The selection of keywords that are in general specific to

Table 1. Examples of keywords

Disease	*Keywords*
Asthma	theophylline, uniphyl, zyflo, peak flow
Coronary Artery Disease	coronary bypass, cabg
Chronic Heart Failure	digoxin, lanoxin, milrinone, natrecor, lvad
Depression	tricyclic, mao, ssri, anafranil, tofranil, prozac
Diabetes	insulin, humalog, lispro, glargine, glyburide, lantus
Gallstones	ursodesoxycholic, bile acid, stone dissolution
Gastro-esophageal Reflux Disease	fundoplication, nissen, toupet, gastroplication
Gout	allopurinol, zyloprim, colchicine, anturane
Hypercholesterolemia	fluvastatin, lescol, simvastatin, zocor
Hypertriglyceridemia	gemfibrozil, lopid, antara, lipofen, tricor
Osteoarthritis	diclofenac, voltaren, chondroitin, joint replacement
Obesity	orlistat, xenical, reductil, gastric bypass, bariatric
Obstructive Sleep Apnea	cpap, uppp, somnoplasty
Peripheral Vascular Disease	trental, cilostazol, fem-pop bypass, pletal
Venous insufficiency	compression stockings, stripping, venotonic

only one disease was based on the medical knowledge of the author, as well as on the training corpus for this i2b2 challenge.

The extraction of biomarker values is based on simplified diagnostic criteria derived from evidence-based clinical practice guidelines available at the U.S. National Guideline Clearinghouse [21]. Their extraction uses regular expressions and focuses on the following diseases:

- Obesity (body weight >=218 lbs or 99 kg; corresponds to a man with a BMI of 30 and a height on the 75th percentile).
- Hypertension (systolic blood pressure >140 mmHg).
- Chronic Heart Failure (ejection fraction <55%).
- Diabetes mellitus (blood glucose concentration >126 mg/dL).
- Hypertriglyceridemia (plasma triglycerides concentration >200mg/dL).

Four different versions of Textractor were developed, and three were used with the intuitive annotations:

1. Based on MMTx and ConText (used as a baseline with intuitive annotations).
2. Based on MMTx and ConText, with biomarker values.
3. Based on MMTx and ConText, with biomarker values and keywords.

4 Results

The testing corpus of 507 documents was made available for three days in June 2008, and each participating team could submit up to three runs for the textual

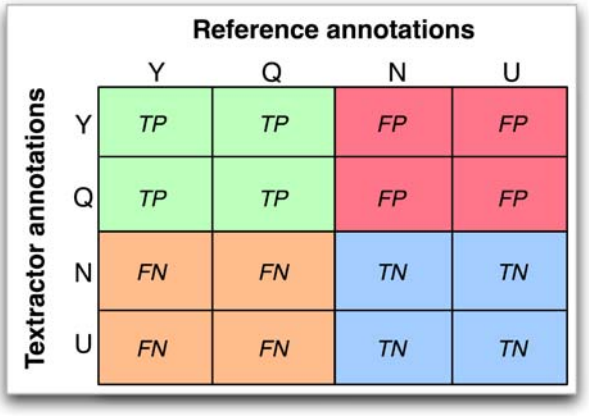


Fig. 1. Sensitivity-oriented annotations adjudication (Y=Yes, N=No, Q=Questionable, U=Unmentioned, TP=True Positive, FP=False Positive, FN=False Negative, TN=True Negative)

and for the intuitive annotation tasks. Evaluation metrics were recall (equivalent to sensitivity or true positive rate here), precision (equivalent to positive predictive value here), and the F1-measure (a harmonic mean of recall and precision [22]). Macro-averaged metrics were the primary metrics, and micro-averaged metrics were secondary metrics. Macro-averaged metrics use one contingency table for each category, are computed locally first, and then averaged across categories; they give equal weight to each category and tend to be more influenced by rare categories. Micro-averaged metrics use only one global contingency table that is the result of merging all local contingency tables; they give equal weight to each document and tend to be more influenced by the most common categories.

The official evaluation is based on methodologies described in Yang et al. [23], and the micro-averaged metrics end up being systematically equal. From the point of view of the researcher or clinician looking for patients with obesity and/or comorbidities, these micro-averaged results are difficult to comprehend. For the researcher, the main interest is to detect patients with an affirmed or questionable disease, and therefore to be more sensitive. We therefore propose a different sensitivity-oriented annotation adjudication method, as described in Fig. 1.

We submitted three runs for the intuitive annotation task (Textractor baseline version, with biomarkers, and with biomarkers and keywords). Table 2 lists all official results.

When using this alternate annotation adjudication method, and analyzing metrics at the disease level (i.e. each disease is a class), results for the best performing versions of Textractor for the intuitive annotation task (Baseline with keywords and biomarkers) are very different than the official results, as listed below in Table 3.

Table 2. Official results for all submissions for intuitive annotations (R=Recall, P=Precision, F1=F1-measure)

Measurement	*Baseline*	*Baseline +biomarkers*	*Baseline +biomarkers +keywords*	*Challenge average (all teams)*
Micro-averaged R	0.9528	0.9538	0.9566	0.90
Micro-averaged P	0.9528	0.9538	0.9566	0.90
Micro-averaged F1	0.9528	0.9538	0.9566	0.90
Macro-averaged R	0.6242	0.6266	0.6387	0.60
Macro-averaged P	0.6378	0.6367	0.6304	0.78
Macro-averaged F1	0.6304	0.6313	0.6343	0.60

Table 3. Results using the alternate annotations adjudication (averages and 95% confidence intervals; R=Recall, P=Precision, F1=F1-measure)

Measurement	*Sensitivity − oriented results*
Micro-averaged R	0.9552 (0.9452-0.9652)
Micro-averaged P	0.9134 (0.9008-0.9260)
Micro-averaged F1	0.9265 (0.9174-0.9356)
Macro-averaged R	0.9242 (0.8687-0.9796)
Macro-averaged P	0.8875 (0.8084-0.9666)
Macro-averaged F1	0.9014 (0.8330-0.9698)

5 Discussion

This evaluation showed that our NLP application performed satisfactorily, reaching the 6th rank (of 28 teams) for the best performing run submitted for the intuitive annotation task. The addition of keywords and biomarker values added only little to the overall performance: only the macro-averaged recall was not quite significantly improved (p-value = 0.06; analysis based on repeated measures analysis of variance). All other metrics were not significantly improved. This limited improvement could be related to several facts: biomarker values were only used with diseases that already had good recall and precision with the baseline system; and many keywords were already detected by the baseline system. Future improvements could include more biomarker values as well as functional diagnostic test results.

Using macro-averages and micro-averages gives a quite complete image of the system evaluated. Macro-averages give an equal weight to each class, and

are therefore more influenced by the performance on rare classes, such as the questionable class in the textual annotation task (represented only 39/11630 annotations, and had an average F1-measure of 0.16), or in the intuitive annotation task (represented only 26/10655 annotations, and had an average F1-measure of 0.01 !!!). Since the questionable class was very rare in this task, errors in this class had a huge impact on macro-averages. For example, only one case had a questionable Gastro-Esophageal Reflux Disease, and Textractor interpreted it as absent. Only two cases had a questionable Coronary Artery Disease, and Textractor interpreted the first mention of one as questionable, but another mention as present, and finally adjudicated it as present. These are examples of single errors with a very high impact.

Micro-averages give an equal weight to each document and tend to be more influenced by the performance on common classes, such as the unmentioned class in the textual annotation task (represented about 71% of the annotations, and had an average F1-measure of 0.94).

The method applied to calculate micro-averages always gives equal recall and precision results, as explained above. These results have little meaning for the researcher interested in detecting patients with a specific disease. For the latter researcher, the sensitivity of the detection method is often more important than the specificity. We therefore proposed an alternate sensitivity-oriented annotations adjudication method, and report a much higher performance, reaching recalls often above 0.95. These results might not correspond to some detailed performance measurements of the information extraction task itself, but they mean that more than 95% of the patients with a certain or possible specific disease were detected, and this is probably easier to understand for the clinical researcher.

The automated extraction of information from text is still a relatively new field of research in the biomedical domain, and the extraction of information from clinical text has received even less attention. The potential uses of information extraction from clinical text are numerous and far-reaching. In the same way the Message Understanding Conferences have fostered the development of information extraction in the general domain, similar competitive challenges for information extraction from clinical text, such as the i2b2 obesity challenge, will undoubtedly stimulate advances in the biomedical field.

Finally, this challenge gave us the opportunity to comparatively evaluate several methodologies for automated information extraction for the research infrastructure we are building at the University of Utah Health Sciences Center and at the Huntsman Cancer Institute.

Acknowledgments

I would like to thank the i2b2 challenge team for the development of the training and testing corpora and for the excellent organization of this challenge.

References

1. Meystre, S.M., Haug, P.: Randomized controlled trial of an automated problem list with improved sensitivity. Int. J. Med. Inform. 77(9), 602–612 (2008)
2. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med. Inform., 128–144 (2008)
3. Pratt, A.W.: Medicine, computers, and linguistics. Advanced Biomedical Engineering 3, 97–140 (1973)
4. Chi, E., Lyman, M., Sager, N., Friedman, C.: Database of computer-structured narrative: methods of computing complex relations. In: IEEE (ed.) SCAMC 1985, pp. 221–226 (1985)
5. Friedman, C., Johnson, S.B., Forman, B., Starren, J.: Architectural requirements for a multipurpose natural language processor in the clinical environment. In: Proc. Annu. Symp. Comput. Appl. Med. Care, pp. 347–351 (1995)
6. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proc. AMIA Symp., pp. 17–21 (2001)
7. Pratt, W., Yetisgen-Yildiz, M.: A Study of Biomedical Concept Identification: MetaMap vs. People. In: Proc. AMIA Symp., pp. 529–533 (2003)
8. Aronson, A.: Query expansion using the UMLS Metathesaurus. In: Proc. AMIA Symp. 1997, pp. 485–489 (1997)
9. Brennan, P.F., Aronson, A.R.: Towards linking patients and clinical information: detecting UMLS concepts in e-mail. J. Biomed. Inform. 36(4-5), 334–341 (2003)
10. Shadow, G., McDonald, C.: Extracting structured information from free text pathology reports. In: Proc. AMIA Symp., Washington, DC, pp. 584–588 (2003)
11. Weeber, M., Klein, H., Aronson, A.R., Mork, J.G., de Jong-van den Berg, L.T., Vos, R.: Text-based discovery in biomedicine: the architecture of the DAD-system. In: Proc. AMIA Symp., pp. 903–907 (2000)
12. Aronow, D.B., Fangfang, F., Croft, W.B.: Ad hoc classification of radiology reports. J. Am. Med. Inform. Assoc. 6(5), 393–411 (1999)
13. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. J. Biomed. Inform., 301–310 (2001)
14. Zhou, L., Melton, G.B., Parsons, S., Hripcsak, G.: A temporal constraint structure for extracting temporal information from clinical narrative. J. Biomed. Inform., 424–439 (2006)
15. Bramsen, P., Deshpande, P., Lee, Y.K., Barzilay, R.: Finding temporal order in discharge summaries. In: AMIA Annu. Symp. Proc., pp. 81–85 (2006)
16. Chapman, W., Chu, D., Dowling, J.N.: ConText: an algorithm for identifying contextual features from clinical text. In: BioNLP 2007: Biological, translational, and clinical language processing. Prague, CZ 2007 (2007)
17. Uzuner, O., Luo, Y., Szolovits, P.: Evaluating the state-of-the-art in automatic de-identification. J. Am. Med. Inform. Assoc. 14(5), 550–563 (2007)
18. Uzuner, O., Goldstein, I., Luo, Y., Kohane, I.: Identifying patient smoking status from medical discharge records. J. Am. Med. Inform. Assoc. 15(1), 14–24 (2008)
19. Meystre, S., Haug, P.J.: Automation of a problem list using natural language processing. BMC Med. Inform. Decis. Mak. 5, 30 (2005)

20. Meystre, S., Haug, P.J.: Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J. Biomed. Inform. 39(6), 589–599 (2006)
21. AHRQ. National Guideline Clearinghouse, http://www.guideline.gov
22. van Rijsbergen, C.J.: Information retrieval. Butterworth (1979)
23. Yang, Y.: An evaluation of statistical approaches to text categorization. Information Retrieval 1(1-2), 69–90 (1999)

Using Existing Biomedical Resources to Detect and Ground Terms in Biomedical Literature

Kaarel Kaljurand, Fabio Rinaldi, Thomas Kappeler, and Gerold Schneider

Institute of Computational Linguistics, University of Zurich
kalju@ifi.uzh.ch, rinaldi@ifi.uzh.ch, kappeler@bluewin.ch,
gschneid@ifi.uzh.ch

Abstract. We present an approach towards the automatic detection of names of proteins, genes, species, etc. in biomedical literature and their grounding to widely accepted identifiers. The annotation is based on a large term list that contains the common expression of the terms, a normalization step that matches the terms with their actual representation in the texts, and a disambiguation step that resolves the ambiguity of matched terms. We describe various characteristics of the terms found in existing term resources and of the terms that are used in biomedical texts. We evaluate our results against a corpus of manually annotated protein mentions and achieve a precision of 57% and recall of 72%.

1 Introduction

The complexity of biological organisms and the success of biological research in describing them, have resulted in a large body of biological entities (genes, proteins, species, etc.) to be indexed, named and analyzed. Proteins are among the most important entities. They are an essential part of an organism and participate in every process within cells. Most proteins function in collaboration with other proteins, and one of the research goals in molecular biology is to identify which proteins interact.

While the number of different proteins is large, the amount of their possible interactions and combinations is even larger. In order to record such interactions and represent them in a structured way, human curators who work for knowledge base projects, e.g. MINT[1] and IntAct[2] (see [5] for a detailed overview), carefully analyze published biomedical articles. As the body of articles is growing rapidly, there is a need for effective automatic tools to help curators in their work. Such tools must be able to detect mentions of biological entities in the text and tag them with identifiers that have been assigned by existing knowledge bases. As the names that are used to reference the proteins can be very ambiguous, there is a need for an effective ambiguity resolution.

[1] http://mint.bio.uniroma2.it
[2] http://www.ebi.ac.uk/intact

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 225–234, 2009.

In this paper, we describe the task of automatically detecting names of proteins, genes, species, experimental methods, and cell lines in biomedical literature and grounding them to widely accepted identifiers assigned by three different knowledge bases — UniProt Knowledgebase (UniProtKB)[3], National Center for Biotechnology Information (NCBI) Taxonomy[4], and Proteomics Standards Initiative (PSI) Molecular Interactions (MI) Ontology[5].

The term annotation uses a large term list that is compiled on the basis of the entity names extracted from the mentioned knowledge bases and from a list of cell line names. This resulting list covers the common expression of the terms. A term normalization step is used to match the terms with their actual representation in the texts. Finally, a disambiguation step resolves the ambiguity (i.e. multiple IDs proposed by the annotator) of the matched terms.

The work presented here is part of a larger effort undertaken in the OntoGene project[6] aimed at improving biomedical text mining through the usage of advanced natural language processing techniques. The results of the entity detection feed directly into the process of identification of protein interactions. Our approach relies upon information delivered by a pipeline of NLP tools, including sentence splitting, tokenization, part of speech tagging, noun and verb phrase chunking, and a dependency-based syntactic analysis of input sentences [7]. The syntactic parser takes into account constituent boundaries defined by previously identified multi-word entities. Therefore the richness of the entity annotation has a direct beneficial impact on the performance of the parser, and thus leads to better recognition of interactions.

2 Term Resources

As a result of the rapidly growing information in the field of biology, the research community has realized the need for consistently organizing the discovered information — assign identifiers to biological entities, enumerate the names by which the entities are referred to, interlink different resources (e.g. existing knowledge bases and literature), etc. This has resulted in large and ever-growing knowledge bases (lists, ontologies, taxonomies) of various biological entities (genes, proteins, species, etc.). Fortunately, many of these resources are also freely available and machine processable. These resources can be treated as linguistic resources and used as an input for the creation of large term lists. Such lists can be used to annotate existing biomedical publications in order to identify the entities mentioned in these publications. In the following we describe four resources: UniProtKB, NCBI Taxonomy, PSI-MI Ontology, and CLKB.

[3] http://www.uniprot.org
[4] http://www.ncbi.nlm.nih.gov/Taxonomy/
[5] http://psidev.sourceforge.net/mi/psi-mi.obo
[6] http://www.ontogene.org

2.1 UniProtKB

The UniProt Knowledgebase (UniProtKB)[7] assigns identifiers to 397,539 proteins and describes their amino-acid sequences. The identifiers come in two forms: numeric accession numbers (e.g. `P04637`), and mnemonic identifiers that make visible the species that the protein originates from (e.g. `P53_HUMAN`). In the following we always use the mnemonic identifiers for better readability.

In addition to enumerating proteins, UniProtKB lists their names that are commonly used in the literature. The set of names covers names with large lexical difference (e.g. both 'Orexin' and 'Hypocretin' can refer to protein `OREX_HUMAN`), but usually not names with minor spelling variations (e.g. using a space instead of a hyphen).

We extracted 626,180 (different) names from the UniProtKB XML file. The ambiguity of a name can be defined as the number of different UniProtKB entries that contain the name. UniProtKB names can be very ambiguous. This follows already from the naming guideline which states that "a recommended name should be, as far as possible, unique and attributed to all orthologs"[8]. Thus, a protein that is found in several species has one name but each of the species contributes a different ID. In UniProtKB, the average ambiguity is 2.61 IDs per name. If we discard the species labels, then the average ambiguity is 1.05 IDs. Ambiguous names (because the respective protein occurs in multiple species) are e.g. 'Cytochrome b' (1770 IDs), 'Ubiquinol-cytochrome-c reductase complex cytochrome b subunit' (1757), 'Cytochrome b-c1 complex subunit 3' (1757). Ambiguous names (without species labels) are e.g. 'Capsid protein' (103), 'ORF1' (97), 'CA' (88).

Table 1 shows the orthographic/morphological properties of the names in UniProtKB in terms of how much certain types of characters influence the ambiguity. Non alphanumeric characters or change of case, while increasing ambiguity, influence the ambiguity relatively little. But as seen from the last column, digits matter a lot semantically. These findings motivate the normalization that we describe in section 3.1. Table 1 also shows the main cause for ambiguity of the names — the same name can refer to proteins in multiple species. While these proteins are identical in some sense (similar function or structure), the UniProtKB identifies them as different proteins.

2.2 NCBI Taxonomy

The National Center for Biotechnology Information provides a resource called NCBI Taxonomy[9], describing all known species and listing the various forms of species names (e.g. scientific and common names). As explained in section 2.1, knowledge of these names is essential for disambiguation of protein names.

[7] We use the manually annotated and reviewed Swiss-Prot section of UniProtKB version 14, in its XML representation.

[8] `http://www.uniprot.org/docs/nameprot`

[9] `http://www.ncbi.nlm.nih.gov/Taxonomy/`

Table 1. Ambiguity of UniProtKB terms. `ID_ORG` stands for the actual identifiers, which include the species ID. `ID` stands for artificially created identifiers where the qualification to the species has been dropped. "Unchanged" = no change done to the terms; "No whitespace" = all whitespace is removed; "Alphanumeric" = only alphanumeric characters are preserved; "Lowercase" = all characters are preserved but lowercased; "Alpha" = only letters are preserved.

	Unchanged	No whitespace	Alphanumeric	Lowercase	Alpha
`ID_ORG`	2.609	2.611	2.624	2.753	10.616
`ID`	1.049	1.050	1.053	1.058	4.145

We compiled a term list on the basis of the taxonomy names list[10], but kept only names whose ID mapped to a UniProtKB species "mnemonic code" (such as `ARATH`)[11]. The final list contains 31,733 entries where the species name is mapped to the UniProtKB mnemonic code. To this list, 8877 entries were added where the genus name is abbreviated to its initial (e.g. 'C. elegans') as names in such form were not included in the source data. These entries can be ambiguous in general (e.g. 'C. elegans' can refer to four different species), but are needed to account for such frequently occurring abbreviation in biomedical texts. Furthermore, six frequently occurring names that consist only of the genus name were added. In these cases, the name was mapped to a unique identifier (e.g. 'Arabidopsis' was mapped to `ARATH`), as it is expected that e.g. 'Arabidopsis' alone is always used to refer to *Arabidopsis thaliana*, and never to e.g. *Arabidopsis lyrata*.

2.3 PSI-MI Ontology

The Proteomics Standards Initiative (PSI) Molecular Interactions (MI) Ontology[12] contains 2207 terms (referring to 2163 PSI-MI IDs) related to molecular interaction and methods of detecting such interactions (e.g. 'western blot', 'pull down'). There is almost no ambiguity in these names in the ontology itself. Several reasons motivate including the PSI-MI names in our term list. First, names of experimental methods are very frequent in biomedical texts. It is thus important to annotate such names as single units in order to make the syntactic analysis of the text more accurate. Second, in some cases a PSI-MI name contains a substring which happens to be a protein name (e.g. 'western blot' contains a UniProtKB term 'blot'). If the annotation program is not aware of this, then some tokens would be mistagged as protein names. Third, some PSI-MI terms overlap with UniProt terms, meaning that the corresponding proteins play an important function in protein interaction detection, but are not the subject of the actual interaction. An example of this is 'GFP' (PSI-MI ID `0367`, UniProtKB ID `GFP_AEQVI`), which occurs in sentences like "interaction between Pop2p and

[10] `ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz` (file `names.dmp`)
[11] `http://www.uniprot.org/help/taxonomy`
[12] `http://psidev.sourceforge.net/mi/psi-mi.obo`

GFP-Cdc18p was detected" where the reported interaction is between POP2 and CDC18, and GFP only "highlights" this interaction.

2.4 Cell Line Names

Cell line names occur frequently in biomedical articles, and one has to be aware of these names in order to avoid tagging them as e.g. protein names. Secondly, almost every cell line comes from one species (although also "chimeric" cell lines are sometimes used), thus the mention of a cell line in a sentence can give a hint of which species the given sentence is about.

We extracted 8741 cell line names from the Cell Line Knowledgebase (CLKB)[13] which is a compilation of data (names, identifiers, cell line organisms, etc.) from various cell line resources (HyperCLDB, ATCC, MeSH) [8]. The data is provided in the standard RDF format. The cell line names in CLKB contain very little ambiguity and synonymy.

CLKB does not map the cell line organism labels to NCBI IDs. This is not directly possible because the organism label often points to a strain, breed, or race of a particular organism (e.g. 'human, Caucasian', 'mouse, BALB/c'), but NCBI does not assign IDs with such granularity. In total, there are 257 organism labels, the most frequent of which we map to the UniProtKB species mnemonic codes (e.g. HUMAN, MOUSE) and the rest to a dummy identifier.

2.5 Compiled Term List

We compiled a term list of 1,688,224 terms based on the terms extracted from UniProtKB, NCBI, PSI-MI, and CLKB, listing the term name, the term ID, and the term type in each entry. The type corresponds roughly to the resource the term originates from. For UniProtKB, there are two types, PROT and GEN. For

Table 2. Frequency distribution of types in the compiled term list, together with the source of the IDs that are assigned to the terms

Frequency	Type	ID	Description
884,641	PROT	UniProt	UniProtKB protein name
752,019	GEN	UniProt	UniProtKB gene name
16,979	ocs	NCBI	NCBI common name, species or below
8877	oss	NCBI	NCBI scientific name, species or below
8877	ogs2	NCBI	oss name, genus abbreviated (e.g. 'A. thaliana')
8741	CLKB	NCBI	CLKB cell line name
3316	oca	NCBI	NCBI common name, above species
2561	osa	NCBI	NCBI scientific name, above species
2207	MI	PSI-MI	PSI-MI term
6	ogs1	NCBI	NCBI selected genus name (e.g. 'Arabidopsis')

[13] http://stateslab.org/data/CellLineOntology/

NCBI, there are six types, distinguishing between common and scientific names, and the rank of the name in the taxonomy. For the PSI-MI Ontology terms and CLKB cell line names there is one type — MI or CLKB, respectively. The frequency distribution of types is listed in table 2. There is relatively little type ambiguity — three terms ('P22', 'LI', 'D2') can belong to three different types, 300 terms to two different types. In the latter case, the ambiguity is between `PROT/GEN` and `CLKB` in 209 cases, and between `PROT/GEN` and `MI` in 69 cases.

3 Automatic Annotation of Terms

Using the described term list, we can annotate biomedical texts in a straightforward way. First, the sentences and tokens are detected in the input text. We use the LingPipe[14] tokenizer and sentence splitter which have been trained on biomedical corpora. The tokenizer produces a granular set of tokens, e.g. words that contain a hyphen (such as 'Pop2p-Cdc18p') are split into several tokens, revealing the inner structure of such constructs which would e.g. allow to discover the interaction mention in "Pop2p-Cdc18p interaction". The processing then annotates the longest possible and non-overlapping sequences of tokens in each sentence, and in the case of success, assigns all the possible IDs (as found in the term list) to the annotated sequence. The annotator ignores certain common English function words (we use a list of ∼50 stop words), as well as figure and table references (e.g. 'Fig. 3a' and 'Table IV').

3.1 Normalization

In order to account for possible orthographic differences between the terms in the term list and the token sequences in the text, a normalization step is included in the annotation procedure. The same normalization is applied to the term list terms in the beginning of the annotation when the term list is read into memory, and to the tokens in the input text. In case the normalized strings match exactly, the input sequence is annotated with the IDs of the term list term. Our normalization rules are similar to the rules reported in [1,10], e.g.

- Remove all characters that are not alphanumeric or space
- Remove lowercase-uppercase distinction
- Normalize Greek letters and Roman numerals, e.g. 'alpha' → 'a', 'IV' → '4'
- Remove the final 'p' if it follows a number, e.g. 'Pan1p' → 'Pan1'
- Remove certain species-indicating prefixes (e.g. 'h' for human, 'At' for *Arabidopsis thaliana*), but in this case, admit only IDs of the given species

In general, these rules increase the recall of term detection, but can lower the precision. For example, sometimes case distinction is used to denote the same protein in different species (e.g. according to UniProtKB, the gene name 'HOXB4' refers to `HXB4_HUMAN`, 'Hoxb4' to `HXB4_MOUSE`, and 'hoxb4' to `HXB4_XENLA`). The gain in recall, however, seems to outweigh the loss of precision.

[14] `http://alias-i.com/lingpipe/`

3.2 Disambiguation

A marked up term can be ambiguous for two reasons. First, the term can be assigned an ID from different term types, e.g. a UniProtKB ID and a PSI-MI Ontology ID. This situation does not occur often and usually happens with terms that are probably not interesting as protein mentions (such as 'GFP' discussed in section 2.3). We disambiguate such terms by removing all the UniProtKB IDs. (Similar filtering is performed in [9].) Second, the term can be assigned several IDs from a single type. This usually happens with UniProtKB terms and is typically due to the fact that the same protein occurs in many different species. Such protein names can be disambiguated in various ways. We have experimented with two different methods: (1) remove all the IDs that do not reference a species ID specified in a given list of species IDs; (2) remove all IDs that do not "agree" with the IDs of the other protein names in the same textual span (e.g. sentence, or paragraph) with respect to the species IDs.

For the first method, the required species ID list can be constructed in various ways, either automatically, on the basis of the text, e.g. by including species mentioned in the context of the protein mention, or by reusing external annotations of the article. We present in [2] an approach to the detection of species names mentioned in the article. The species mentions are used to create a ranked list, which is then used to disambiguate other entities (e.g. protein mentions) in the text.

The second method is motivated by the fact that according to the IntAct database, interacting proteins are usually from the same species: less than 2% of the listed interactions have different interacting species. Assuming that proteins that are mentioned in close proximity often constitute a mention of interaction, we can implement a simple disambiguation method: for every protein mention, the disambiguator removes every UniProtKB ID that references a species that is not among the species referenced by the IDs of the neighboring protein mentions.

In general, the disambiguation result is not a single ID, but a reduced set of IDs which must be further reduced by a possible subsequent processing step.

4 Evaluation

We evaluated the accuracy of our automatic protein name detection and grounding method on a corpus provided by the IntAct project[15]. This corpus contains a set of 6198 short textual snippets (of 1 to about 3 sentences), where each snippet is mapped to a PubMed identifier (of the article the snippet originates from), and an IntAct interaction identifier (of the interaction that the snippet describes). In other words, each snippet is a "textual evidence" that has allowed the curator to record a new interaction in the IntAct knowledge base. By resolving an interaction ID, we can generate a set of IDs of interacting proteins and a set of species involved in the interaction, for the given snippet. Using the PubMed identifiers, we can generate the same information for each mentioned article. By

[15] ftp://ftp.ebi.ac.uk/pub/databases/intact/current/various/data-mining/

Table 3. Results obtained on the IntAct snippets, with various forms of disambiguation, measured against PubMed IDs. The evaluation was performed on the complete IntAct data (*all*), and on a 5 times smaller fragment of IntAct (*subset*) for which we automatically extracted the species information. Three forms of disambiguation were applied: IntAct = species lists from IntAct data; TX = species lists from our automatic species detection; span = the species of neighboring protein mentions must match. Additionally, combinations of these were tested: e.g. IntAct & span = IntAct disambiguation followed by span disambiguation. The best result in each category is in boldface.

Disamb. method	Corpus	Precision	Recall	F-Score	True pos.	False pos.	False neg.
No disamb.	all	0.03	0.73	0.05	2237	81,662	848
IntAct	all	0.56	0.73	0.63	2183	1713	804
span	all	0.03	0.71	0.06	2186	68,026	899
IntAct & span	all	0.57	0.72	**0.64**	2147	1599	840
span & IntAct	all	0.57	0.72	0.64	2142	1631	821
No disamb.	subset	0.02	0.69	0.04	424	20,344	188
IntAct	subset	0.51	0.71	0.59	414	397	170
span	subset	0.02	0.67	0.05	407	16,319	205
IntAct & span	subset	0.53	0.69	**0.60**	404	363	180
span & IntAct	subset	0.52	0.69	0.59	399	369	177
TX	subset	0.42	0.59	0.49	340	478	241
TX & span	subset	0.43	0.57	**0.49**	332	445	249
span & TX	subset	0.42	0.57	0.48	329	457	244

comparing the sets of protein IDs reported by the IntAct corpus providers, and the sets of protein IDs proposed by our tool, we can calculate the precision and recall values.

We annotated the complete IntAct corpus by marking up with an entry in the term list the token sequences that the normalization step matched. Each resulting annotation includes a set of IDs which was further reduced by the two disambiguation methods described in 3.2, i.e. some or all of the IDs were removed. Results before and after disambiguation are presented in table 3. The results show a relatively high recall which decreases after the disambiguation. This change is small however, compared to the gain in precision. False negatives are typically caused by missing names in UniProtKB, or sometimes because the normalization step fails to detect a spelling variation. A certain amount of false positives cannot be avoided due to the setup of task — the tool is designed to annotate all proteins contained in the sentences, but not all of them necessarily participate in interactions, and thus are not reported in the IntAct corpus.

5 Related Work

There is a large body of work in named entity recognition in biomedical texts. Mostly this work does not cover grounding the detected named entities to existing knowledge base identifiers. Recently, however, as a result of the BioCreative

workshop, more approaches are extending from just detecting entity mentions to "normalizing" of the terms. In general, such normalization handles gene names (by grounding them to EntrezGene[16] identifiers). [6] gives an overview of the BioCreative II gene normalization task.

A method of protein name grounding is described in [10]. It uses a rule-based approach that integrates a machine-learning based species tagger to disambiguate protein IDs. The reported results are similar to ours. In the BioCreative Meta Server (BCMS)[17] [3], 2 out of 13 gene/protein taggers annotate using UniProtKB protein identifiers. The Whatizit[18] webservice annotates input texts with UniProtKB, Gene Ontology[19], and NCBI terms. A preliminary comparison showed that our approach gives results of similar quality.

Several linguistic resources have been compiled from existing biomedical databases. BioThesaurus[20] is a thesaurus of gene and protein names (and their synonyms and textual variants) where each name is mapped to a UniProtKB identifier [4]. The latest version 5.0 of BioThesaurus contains more than 9 million names, extracted from 35 different databases. The biggest contributer, however, is UniProtKB, mainly its TrEMBL section.

ProMiner[21] is a closed source dictionary-based named entity tagger that uses an entity name database compiled from a wide variety of sources for gene, protein, disease, tissue, drug, cell line, and other names. Detailed information about this resource has not been published.

6 Conclusions and Future Work

The main goal of the work described in this paper is to reliably identify protein mentions in order to identify protein-protein interactions in a subsequent processing step. We use a large term list compiled from various sources, and a set of normalization rules that match the token sequences in the input sentences against the term list. Each matched term is assigned all the IDs that are possible for this term. The following disambiguation step removes most of the IDs on the basis of the term context and knowledge about the species that the article discusses. For the evaluation, we have used the freely available IntAct corpus of snippets of textual evidence for protein-protein interactions. To our knowledge, this corpus has not been used in a similar evaluation before.

In the future, we would like to include more terminological resources in the annotation process. While the described four resources (UniProtKB, NCBI Taxonomy, PSI-MI Ontology, CLKB cell line names) contain the most important names used in biomedical texts, there exist other names that are frequently used but that are not covered by these resources, e.g. names of certain chemical compounds, diseases, drugs, tissues, etc.

[16] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
[17] http://bcms.bioinfo.cnio.es
[18] http://www.ebi.ac.uk/webservices/whatizit/
[19] http://www.geneontology.org
[20] http://pir.georgetown.edu/iprolink/biothesaurus/
[21] http://www.scai.fraunhofer.de/prominer.html

Acknowledgments

This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, NITAS, Text Mining Services, CH-4002, Basel, Switzerland. The authors would like to thank the three anonymous reviewers of AIME'09 for their valuable feedback.

References

1. Hakenberg, J.: What's in a gene name? Automated refinement of gene name dictionaries. In: Proceedings of BioNLP 2007: Biological, Translational, and Clinical Language Processing; Prague, Czech Republic (2007)
2. Kappeler, T., Kaljurand, K., Rinaldi, F.: TX Task: Automatic Detection of Focus Organisms in Biomedical Publications. In: BioNLP 2009, NAACL/HLT, Boulder, Colorado, June 4–5 (2009)
3. Leitner, F., Krallinger, M., Rodriguez-Penagos, C., Hakenberg, J., Plake, C., Kuo, C.-J., Hsu, C.-N., Tsai, R.T.-H., Hung, H.-C., Lau, W.W., Johnson, C.A., Saetre, R., Yoshida, K., Chen, Y.H., Kim, S., Shin, S.-Y., Zhang, B.-T., Baumgartner, W.A., Hunter, L., Haddow, B., Matthews, M., Wang, X., Ruch, P., Ehrler, F., Ozgur, A., Erkan, G., Radev, D.R., Krauthammer, M., Luong, T., Hoffmann, R.: Introducing meta-services for biomedical information extraction. Genome Biology 9(suppl. 2), S6 (2008)
4. Liu, H., Hu, Z.-Z., Zhang, J., Wu, C.: BioThesaurus: a web-based thesaurus of protein and gene names. Bioinformatics 22(1), 103–105 (2006)
5. Mathivanan, S., Periaswamy, B., Gandhi, T.K.B., Kandasamy, K., Suresh, S., Mohmood, R., Ramachandra, Y.L., Pandey, A.: An evaluation of human protein-protein interaction data in the public domain. BMC Bioinformatics 7(suppl. 5), 19 (2006)
6. Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C.: Overview of BioCreative II gene normalization. Genome Biology 9(suppl. 2), S3 (2008)
7. Rinaldi, F., Kappeler, T., Kaljurand, K., Schneider, G., Klenner, M., Clematide, S., Hess, M., von Allmen, J.-M., Parisot, P., Romacker, M., Vachon, T.: OntoGene in BioCreative II. Genome Biology 9(suppl. 2), S13 (2008)
8. Sarntivijai, S., Ade, A.S., Athey, B.D., States, D.J.: A bioinformatics analysis of the cell line nomenclature. Bioinformatics 24(23), 2760–2766 (2008)
9. Tanabe, L., John Wilbur, W.: Tagging gene and protein names in biomedical text. Bioinformatics 18(8), 1124–1132 (2002)
10. Wang, X., Matthews, M.: Distinguishing the species of biomedical named entities for term identification. BMC Bioinformatics 9(suppl. 11), S6 (2008)

An Ontology for the Care of the Elder at Home

David Riaño[1], Francis Real[1], Fabio Campana[2], Sara Ercolani[3], and Roberta Annicchiarico[4]

[1] Research Group on Artificial Intelligence, Rovira i Virgili University, Tarragona, Spain
[2] CAD RM B, Rome, Italy
[3] Department of Geriatrics, University of Perugia, Perugia, Italy
[4] IRCCS S. Lucia, Rome, Italy
{david.riano, francis.real}@urv.net, saraercolani@libero.it, fcampana@tiscali.it, r.annicchiarico@hsantalucia.it

Abstract. The health-care of the elder at home is highly demanded in modern societies. It is based on the difficult task of coordinating multiple professionals and procedures acting on the same patient. K4CARE is a project aiming at implementing and testing a technology-based incremental and adaptable model to assist health care systems in home care. One of the key components of this model is the Case Profile Ontology (CPO) that is used to support the activities in the life-cycle of home care. These activities define a path that goes from assessing the problem to deploying a care plan. Along this path several CPO-based tools have been implemented to ease the assessment step, to manage care plans as State-Decision-Action diagrams, to combine care plans for comorbid patients, and to personalize care plans. The use of these tools significantly reduces the complexity of dealing with patients at home.

1 Introduction

K4CARE (www.k4care.net) is the join effort of thirteen European institutional partners to construct a technology-based model for the care of the elder at home that could not only be deployed in European Health-Care Systems, but also be adapted to other Health-Care Systems worldwide. This construction is divided into three consecutive steps: propose an adaptable model, develop the technologies and computer-based tools to implement the model, and validate the model in the health care system of the town of Pollenza, Italy.

The health-care model [1] is defined to have two dimensions: human resources and services. This model is formalized in what is called the Agent Profile Ontology (APO) [2,3].

Besides the APO that is devoted to formalize the management issues of home care in a health-care system, K4CARE provides a Case Profile Ontology (CPO) that gathers, provides structure to, and relates the concepts required to assess, to diagnose, and to treat patients at home. This new ontology is concerned with the clinical, medical and social levels of the treatment. In K4CARE, the CPO is finally used to support home care decisions, and also to personalize the treatment of patients.

In this paper, we describe the Case Profile Ontology and how this ontology is used to provide support to health-care professionals (i.e., physicians, nurses, social workers, etc.) in the care of patients at home.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 235–239, 2009.

2 The Case Profile Ontology

The average home care patient is an elderly patient, with co-morbid conditions and diseases, cognitive and/or physical impairment, functional loss from multiple disabilities, and impaired self-dependency [1].

The care of this sort of patient requires complex health-care management policies to be integrated with expert supervision and online adaptation of the care plan to the patient evolving condition. In this sense, the American Medical Association (AMA) indicates that the management of home care patients is a function of the physician's skills in optimizing the patient's independence while utilizing medical and social resources to minimize the effects of illness and disability in the patient's daily life [6], according to an evolving care plan.

The *care plans*, together with the *assessment tools*, are the final components of the life-cycle in the management of home care patients. This *life-cycle* starts with the admission of the patient to the home care service. Then the patient condition is assessed in order to propose a care plan. This plan must be adapted to the medical and social particularities of that patient before it is performed and the results evaluated. Depending on the evaluation, the care plan can be adjusted or a new care plan proposed and the process repeated.

The Case Profile Ontology (CPO) whose root concepts are shown in figure 1 was conceived to support professional decisions in the life-cycle of home care treatments. It is based on the peripheral concepts of Problem Assessment (i.e., assessment tools) and Intervention (i.e., actions of the care plans), and how these two concepts are related through intermediate (but relevant) concepts as Signs & Symptoms and Care Problems as Social Issues, Syndromes, or Diseases.

Social Issues in the CPO comprise the concepts of lack of family support, low income, lack of social network, bad environment, and insanity. The concept Syndrome represents a complex health situation in which a combination of Signs and Symptoms co-occur more frequently than would be expected on the basis of chance alone, generating a functional decline. The CPO includes the syndromes of cognitive impairment [7,8] and immobility [8,9], and a subset of diseases as explicit concepts.

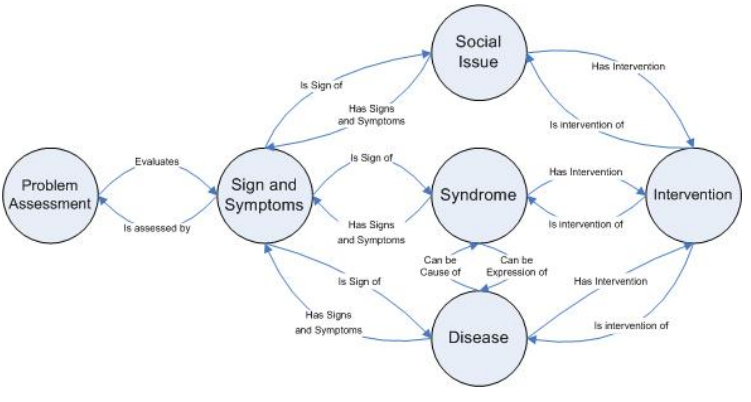


Fig. 1. Case Profile Ontology: main classes and properties

The number of concepts in the CPO and their codification system are provided in table 1. Moreover, the ontology has other complementary concepts: 28 routes of administration and 177 ATC [10] codes for the pharmacological interventions, and 270 ICD10 [11] codes for the diseases.

Table 1. Number of classes of the CPO

	CODE	CPO v3
Disease	DIxx.x	21
Intervention	INxx.x.x	174
ProblemAssessment	several	214
SignAndSymptom	SSxx.xx	317
SocialIssue	SIxx	5
Syndrome	SYx.x	2
Remaining concepts	-	475
Total number of classes	-	**1,208**

We have developed this ontology in OWL [12] using Protégé [13].

Figure 1 also shows as arrows the properties that relate the concepts in the CPO. In the life-cycle of home care, these properties can be used to support several sorts of health-care reasoning: *Forward reasoning* is used to provide the assessments of some signs and symptoms of a patient, determine what are the feasible social issues, syndromes, and diseases (*isSignOf* property) of this patient, the set of interventions he or she requires (*hasIntervention* property), and the way they are applied (*hasFIP* property). *Backward reasoning* is used to know the interventions a patient is receiving, determine the social issues, syndromes, and diseases of this patient (*isInterventionOf* property) and the signs and symptoms this patient should present (*hasSignsAndSymptoms* property), and finally get some recommendations on the appropriate assessments of the health and social conditions of this patient (*IsAssessedBy* property). *Inward reasoning* is used to observe the diseases of a patient, foresee the possible syndromes the patient can develop (*CanBeCauseOf* property) and suggest proper interventions (*hasIntervention* property).

A relevant aspect of social issues, syndromes, and diseases is their datatype property *hasFIP* that is used to store a recommended Formal Intervention Plan to deal with these problems. A *Formal Intervention Plan* (FIP) is a computer-interpretable structure describing a care plan that relates signs, symptoms, social conditions, and secondary diseases with interventions.

There are several languages that can be used to represent FIPs: EON [14], GLIF [15], Prodigy [16], Proforma [17], SDA [4], etc. In K4CARE we use the SDA language.

SDA stands for State-Decision-Action. Terminology in states is restricted to the instances and classes of the sort Signs and Symptoms, Social Issue, Syndrome, and Disease of the CPO in figure 1. Terminology in decisions is restricted to CPO instances and classes of the sort Signs and Symptoms. In actions, terminology is taken from the set of instances and classes of Problem Assessment, and Intervention in the CPO.

3 CPO-Based Problem Assessment

In the CPO, problem assessment comprises some aspects that assess the condition of the patient during the first encounter and whenever a re-evaluation is required. We distinguish between: *Comprehensive assessment* which is devoted to detect the whole series of the patient diseases, conditions, and difficulties, from both the medical and the social perspectives. It comprises Multi-Dimensional Evaluation, Clinical Assessment, Physical Examination and Social needs and network assessment. *Consultation* which is a referral to a specialist physician. E.g., neurologist or endocrinologist. *Diagnostic Examination* which is a process by which physicians evaluate an area of the subject's body that is not externally visible, seeking to diagnose. E.g., hearing test or EEG. *Laboratory Analysis* which is an examination of several parameters in patient's fluids as blood, urine, etc. E.g., glucose tolerance test or INR.

4 Care Plan Configuration

Care plans (or FIPs) are represented as SDA [4] diagrams converted to XML notation.

SDA diagrams contain states, decisions and actions interconnected. These are described with state, decision, and action terms, respectively. All these terms must appear in the CPO as instances or subclasses of the different concepts represented in figure 1.

When a social issue, a syndrome or a disease is detected in a patient, then the related SDA in the CPO (*hasFIP* property) is activated as the current care plan. Unfortunately, the average home care patient is a comorbid case in which more than one social issue, syndrome, or disease co-occur. In this case, several SDAs are simultaneously activated for the patient. We consider that having several simultaneous care plans for the same patient is counterintuitive and source of medical errors. Therefore, we have started to develop merging techniques to combine several SDAs to form a single action plan [18].

Before a care plan is applied to a patient, it is transformed into an *individual care plan* (i.e., only valid for this patient) by simplifying all the decisions whose terms are known for the target patient.

The CPO helps us to detect *unjustified assessment orders*, *test omissions*, *unjustified interventions* and *useless information* in individual care plans.

5 Conclusion

We have proposed an OWL ontology for the care of patients at home which is based on the concepts of assessment, sign and symptom, social issue, syndrome, disease, and intervention. This ontology rules not only the way that the processes of problem assessment and care plan proposal are carried out in the life-cycle of home care, but also the sorts of health-care reasoning that goes from patient assessment to care plan proposal (i.e., problem-to-solution view or *forward reasoning*) and from the sorts of interventions a patient is receiving to the assessment of signs and symptoms (i.e., solution-to-problem view or *backward reasoning*). The first view helps the physician in the task of clinical decision, whereas the second view provides a way of detecting and reducing medical errors in the treatment of patients at home. The ontology is complemented with several

tools to edit, merge, and personalize care plans. The tool to edit formal intervention plans is called *SDA Lab* and it is described in [5].

Acknowledgement

This work was funded in part by grants IST-2004-026968 from the 6^{th} Framework Program of the European Commission and TIN2006-15453 from the Spanish Ministry of Science and Education.

References

1. Campana, F., Annicchiarico, R., Riaño, D., et al.: The K4CARE model (2006), http://www.k4care.net/fileadmin/k4care/public_website/downloads/K4C_Model_D01.rar
2. Gibert, K., Valls, A., Casals, J.: Enlarging a medical actor profile ontology with new care units. In: Riaño, D. (ed.) K4CARE 2007. LNCS (LNAI), vol. 4924, pp. 101–116. Springer, Heidelberg (2008)
3. Gibert, K., Valls, A., Riaño, D.: Knowledge engineering as a support for building an actor profile ontology for integrating home-care systems. In: Proc. MIE, Göteborg, Sweden (2008)
4. Riaño, D.: The SDA* model: a set theory approach. In: Proc. 20th CBMS, Slovenia (2007)
5. López-V, J.A., Riaño, D.: SDA lab v1.3 (2007), http://banzai-deim.urv.net/repositories/repositories.html
6. Ramsdell, J.W. (ed.): Medical management of the home care patient: guidelines for physicians. American Medical Association (2007)
7. Guideline on medicinal products for the treatment of Alzheimer's disease and other dementias. In: EMEA/CHMP (2007), http://www.emea.europa.eu/pdfs/human/ewp/055395endraft.pdf
8. Ercolani, S., Federici, A., et al.: Formal Intervention Plans II. K4CARE project D06.2 deliverable (2008), Available from david.riano@urv.net
9. Campana, F., Riaño, D., et al.: Formal Intervention Plans III. K4CARE project D06.3 deliverable (2008), Available from david.riano@urv.net
10. Anatomical therapeutic chemical classification system (2008), http://www.whocc.no/atcddd/
11. International classification of diseases (ICD), http://www.who.int/classifications/icd/en/
12. OWL Web Ontology Language Guide (2004), http://www.w3.org/TR/owl-guide/
13. Gennari, J.H., Musen, M.A., Fergerson, R.W., et al.: The evolution of Protégé: an environment for knowledge-based systems development. Int. J. Hum. Comput. Stud. 58(1), 89–123 (2003)
14. Musen, M., Tu, S., Das, A., Shahar, Y.: EON: a component-based approach to automation of protocol-directed therapy. JAMIA 3, 367–388 (1996)
15. Boxwala, A.A., Peleg, M., Tu, S., et al.: GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. J. Biomed. Inform. 37(3), 147–161 (2004)
16. Johnson, P.D., Tu, S., Booth, N., et al.: Using scenarios in chronic disease management guidelines for primary care. In: Proc. AMIA (2000)
17. Sutton, D.R., Fox, J.: The syntax and semantics of the PROforma guideline modelling language. JAMIA 10(5), 433–443 (2003)
18. Real, F., Riaño, D.: Automatic combination of formal intervention plans using SDA* representation model. In: Riaño, D. (ed.) K4CARE 2007. LNCS (LNAI), vol. 4924, pp. 75–86. Springer, Heidelberg (2008)

Ontology-Based Personalization and Modulation of Computerized Cognitive Exercises

Silvana Quaglini[1], Silvia Panzarasa[2], Tiziana Giorgiani[1], Chiara Zucchella[3], Michelangelo Bartolo[3,4], Elena Sinforiani[3], and Giorgio Sandrini[3]

[1] Department of Computer Science and Systems, University of Pavia
[2] CBIM, Pavia
[3] IRCCS Foundation Hospital "C. Mondino", Pavia, Italy
[4] Dept. of Neurorehabilitation II, NEUROMED Institute IRCCS, Pozzilli, Italy
Silvana.Quaglini@unipv.it

Abstract. Cognitive rehabilitation may benefit from computer-based approaches that, with respect to paper-based ones, allow managing big amounts of stimuli (images, sounds, written texts) and combining them to create ever-new exercises. Moreover, they allow storing and analysing patients' performance, that may vary in time, thus increasing/decreasing difficulty of the exercises accordingly. An ontological organisation of the stimuli may help to automatically generate patient-tailored exercises, accounting for patients' performance, skills and preferences.

Keywords: Cognitive rehabilitation, ontology, exercise adaptation.

1 Introduction

Cognitive rehabilitation is designed to reduce and/or compensate the impact on daily living of cognitive dysfunction in patients suffering from brain damage [1,2]. Rehabilitative strategies are focused to improve attention, memory, space and time orientation, logical abilities and abstract reasoning and to stimulate speech production and comprehension. These strategies, traditionally carried out by paper-based approach during face-to-face encounters with neuropsychologists and speech therapists, can include the use of computer programs. The shift of stimuli from the paper to the screen may facilitate the issue of proposing new stimuli to patients, by means of big corpora stored in databases [3,4,5]. We propose a new approach, based on the stimuli domain ontology and on patient's profile, that allows modulating the difficulty of exercises and, more generally, personalizing the exercises.

Since some years [6], the medical partner of this study is using *E-Prime*®[1], that offers the typical facilities to generate and run cognitive exercises. It also allows programming new functionalities and connecting to external databases. We started by increasing this commercial tool with a more user-friendly interface. Then we integrated the system with a wide database of stimuli, their relationships and hierarchic

[1] Science Plus Group, Zernikelaan 6, 9747 AA Groningen (The Netherlands).

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 240–244, 2009.

classification (representing the system ontology). To tailor the exercises on the basis of users' preferences and skill, we enriched the patient database with a patient's profile (education, hobbies, etc.), and performances in terms of correct answers and execution time. The exercises can be carried out by the patient himself or with the supervision of even unskilled personnel, both at hospital and patient's home.

2 The Solution Proposed

Our repository, implemented by MS Access and named *Trials-DB*, contains about 5,000 nouns, classified in different semantic categories (foods, animals, sports, etc), and about 1,000 verbs; 2,000 stimuli are associated with sound and half of them also to an image.

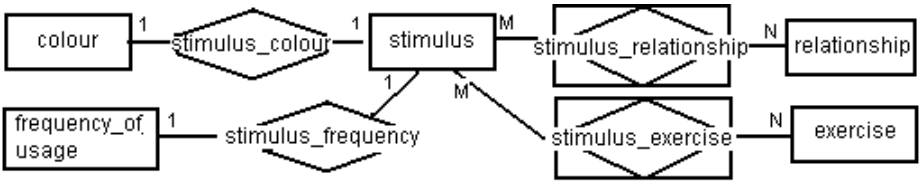


Fig. 1. The E-R Diagram of the stimuli repository

Fig. 1 represents the E-R diagram: the core tables are *Stimulus*, *Stimulus_Exercise* and *Stimulus_Relationship,* while the other ones represent encoding tables. Each row in the *Stimulus* table describes a stimulus characterized by the following attributes (keys are underlined): ***ID***, stimulus id; ***Description***; ***Length***; ***Frequency of usage***, word's frequency of usage in the spoken Italian language; ***Colour***, the leading colour for stimulus associated to an image; ***ImageFile***, path to the image associated to the stimulus; ***Soundfile***, path to the associated audio file. The stimulus' ID is a hierarchic code: for example the "IS-A" relationship between carnivorous and mammal is represented by the two codes 19.1.7 and 19.1 respectively. The *Stimulus_Relationship* table contains the relationships among stimuli as a combination of three fields: ***StimulusID1*** , ***StimulusID2*** (foreign keys to table *Stimulus*) and ***RelationshipID*** (foreign key to table *Relationship*). Taking into account the "COMPOSED_BY" relationship, "magazine" is in association with "paper". As another example, "dark" is in association with "light" in the "OPPOSITE" relationship. The *Stimulus_Exercise* table specifies which stimuli can be used for each exercise. A stimulus is suitable for a particular exercise on the basis of its properties like the presence of the corresponding image, sound, etc.

2.1 The Patient Database

Usually, difficulty of exercises is related to both patient's clinical status and pre-disease patient's capabilities, skills, scholarship, etc. Moreover, difficulty of exercises depends on the periodical controls assessing the patient recovery trend. To this aim, it is fundamental to store patient's status, preferences and progresses into a database accessible by *E-Prime*®. Here the core tables are: "*Personal_Data*" containing in

particular education degree, gender, date of birth and the pathology requiring rehabilitation; "*Patient_Preferences*" storing patient's hobbies and interests, encoded as in the "*Stimulus*" table (e.g. sports, movies, nature, etc); "*Session_Logs*" storing the exercise type, date, execution time, a binary value for the answer (correct 1, incorrect 0) and the answer itself; "*Session_Stimuli*" storing the stimuli shown to the patients in every session, that can be useful for further statistics.

2.2 The User Interface

As mentioned, a new user interface to *E-Prime*® has been implemented. The therapist can easily create personalized exercises (Fig. 2), by choosing them from a pre-defined list (lower part of the figure), and by tuning the parameters shown in the upper part. The general layout of an exercise type must be previously designed with E-Studio: it includes instructions for patients, the graphical arrangement of the stimuli and the feedback page (e.g. smilings for correct, incorrect or no response).

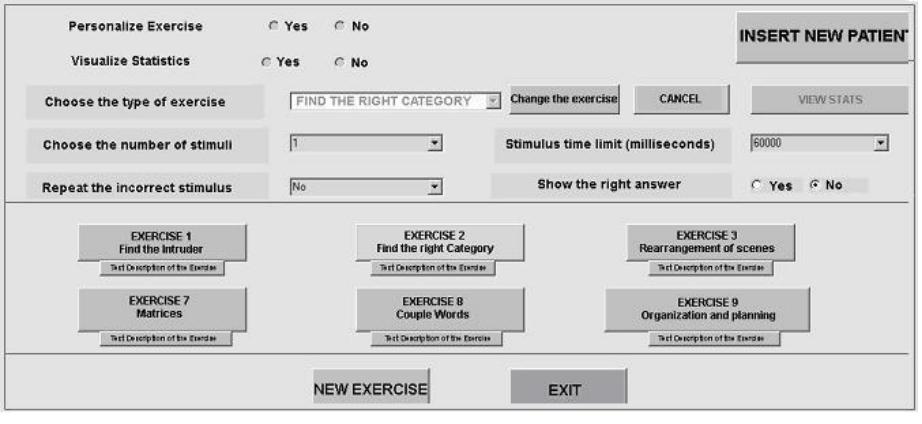


Fig. 2. The therapist's interface (only a subset of exercises are reported)

Once the therapist has designed the exercises for a patient, they can be run using the same interface,in such a way that final users (therapists and patients) will only deal with the *E-Prime*® runtime component, while the other components, much more difficult to use, are transparent to them. This process is illustrated in Fig. 3.

2.3 The Ontology-Based Engine

The following examples illustrate the usefulness of a stimuli domain ontology for generating and tailoring some of the most common exercises.
"Find the right category" displays three (or more) images and the patient must choose the correct category among some possibilities written below (see Fig. 3, right part). In general, in our ontology it is easier for a patient to differentiate objects of more general categories than of more specific ones (e.g. distinguishing between

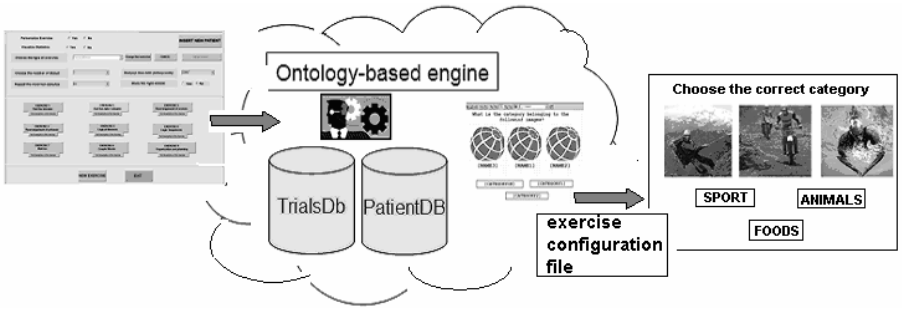


Fig. 3. The process of exercise generation, from the therapist's interface to the patient's one

mammals and birds is easier than between marsupials and primates). Thus, "IS-A" relationships can be efficiently exploited to modulate the difficulty level in this exercise.
"Couple words" lists a number of words that the patient must couple according to a specific relation. The exercise is complete when all the words are correctly associated. Mentioning or not the underlying relation makes the exercise less or more difficult. According to the "*Stimulus_Relationship*" table, a simple exercise could be created using couples of stimuli from the same relationship. In this case considering the "LIVING" relationship the generated list could be "Person; Dog; Horse; House; Kennel; Stable". A more difficult exercise could be generated using couples of stimuli from two or more different relationships.

An interesting observation, from the above describe exercise, is that the association "Dog" with "Horse" is incorrect, with respect to the LIVING relationship. However, they are both mammals. Therefore this association, from the ontological point of view, is not completely senseless as could be "Person" with "Kennel", and the patient could, at a first glance, try this association. This could be taken into account to rate the severity of a patient's mistakes.

Personalised exercises -As mentioned, the system stores the patient's performance in terms of number of correct answers and time spent. This information is used for modulating the exercise difficulty. Actually the performance indicator is the weekly percentage of correct answers. If the patient's performance is higher than a pre-defined cut-off (e.g. 80%), the difficulty is automatically increased. About patient-tailoring, knowing patients' personal preferences (e.g. favourite sports, teams, hobbies, etc.), the system extracts stimuli of his major interest with a highest probability: this can be useful to increase the system acceptability and patient's compliance. Other aspects of patient's profile, such as scholarship, are used to extract words more or less frequently used in the Italian language.

3 Results

The system is in its early evaluation phase. A questionnaire is administered to every patient, after some computer sessions, to collect information about his satisfaction with the new rehabilitation modality. The questionnaire is in graphical format, as shown in Fig. 4, to facilitate comprehension for aphasic patients. While data necessary to

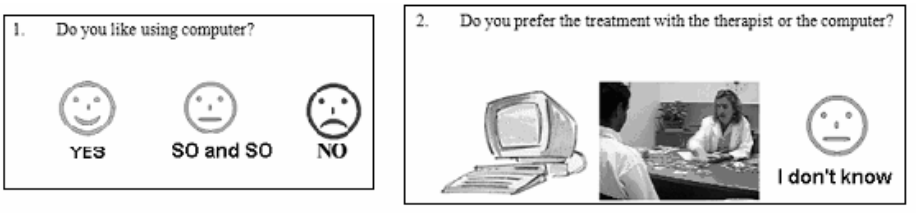


Fig. 4. A scratch of the questionnaire administered to the patients

produce reliable statistics will be available only in some months, preliminary results on the first seven patients are encouraging, and half of them declared to prefer using the computer rather than the traditional paper-based approach.

4 Conclusions

The current trend in several countries is to move the care delivery, when possible, from hospital to patients' home. Rehabilitation is one of the medical tasks that can benefit from this trend, and telemedicine and tele-homecare may represent an appropriate framework. We illustrated a system that, also thanks to an ontology-based approach, is able to automatically adapt to a patient's needs and capabilities, and for this reason it can be efficiently inserted within a tele-homecare service.

Acknowledgments. The authors thank C. Pistarini and B. Cattani (IRCCS Foundation "S. Maugeri", Pavia) for their collaboration in previous experiments with *E-Prime*®.

References

1. Mazzucchi, A.: La riabilitazione neuropsicologica. Masson ed (1999)
2. Christensen, A., Uzzel, B.P.: International Handbook of neuropsychological rehabilitation. Plenum Press (1999)
3. Schuhfried, G.: Rehacom Computer-aided cognitive rehabilitation. EMS Bologna (1986)
4. Grawemeyer, B., Cox, R., Lum, C.: AUDIX:a knowledge-based system for speech-therapeutic auditory discrimination exercises. Stud Health Technol Inform 77, 568–572 (2000)
5. Bruce, C., Edmundson, A., Coleman, M.: Writing with voice: an investigation of the use of a voice recognition system as a writing aid for a man with aphasia. Int. J. Lang Commun. Disord. 38(2), 131–148 (2003)
6. Albanesi, M.G., Panzarasa, S., Cattani, B., Dezza, S., Maggi, M., Quaglini, S.: Segmentation Techniques for Automatic Region Extraction: An Application to Aphasia Rehabilitation. In: Bellazzi, R., Abu-Hanna, A., Hunter, J. (eds.) AIME 2007. LNCS (LNAI), vol. 4594, pp. 367–377. Springer, Heidelberg (2007)

HomeNL: Homecare Assistance in Natural Language. An Intelligent Conversational Agent for Hypertensive Patients Management

Lina Maria Rojas-Barahona, Silvana Quaglini, and Mario Stefanelli

Department of Computer Science and Systems, University of Pavia,
Via Ferrata 1, 27100 Pavia, Italia
{linamaria.rojas,silvana.quaglini,mario.stefanelli}@unipv.it
http://www.labmedinfo.org/

Abstract. The prospective home-care management will probably offer intelligent conversational assistants for supporting patients at home through natural language interfaces. Homecare assistance in natural language, HomeNL, is a proof-of-concept dialogue system for the management of patients with hypertension. It follows up a conversation with a patient in which the patient is able to take the initiative. HomeNL processes natural language, makes an internal representation of the patients' contributions, interprets sentences by reasoning about their meaning on the basis of a medical-knowledge representation and responds appropriately. HomeNL's aim is to provide a laboratory for studying natural language processing (NLP) and intelligent dialogues in clinical domains.

Keywords: NLP, Dialogue Systems, Telemedicine, Hypertension.

1 Introduction

A range of approaches for modeling Dialogue Systems (DSs) has been proposed in the literature from simple pattern-matching techniques to structured architectures. The most sophisticated approaches consider the cognitive state of the conversational agent and provide methods for modeling intelligent dialogues by enabling mixed-initiative and complex discourse phenomena [1]. Despite these advances, most of the dialogue-based interfaces implemented in the medical domain are devoted to simple approaches to dialogue management [2], offering system-driven applications that limit the variety of expressions users might use [1]. The adoption of more elaborated methods in the medical domain is an open research area [2]. Indeed, emergent prototypes have been deployed, of which the most salient is Chester, a personal medication advisor for elders [3]. Although the resulting prototypes are still far from being fully operative in real settings and building them requires complicated and costly solutions, these efforts highlighted promising lines of research in the field.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 245–249, 2009.

In this paper we present HomeNL a proof of concept for intelligent conversational agents in the management of hypertensive patients. Thereby, we have explored striking formalisms of computational linguistics and NLP. Moreover, we have adopted theoretical-based frameworks that simplify the programming burden. Despite being in a preliminary stage, HomeNL stores its cognitive state, understands and generates natural language, supports reasoning by accessing a logic-based knowledge representation (KR) and maintains a coherent conversation with a patient at the same time it enables mixed-initiative.

2 The Architecture

Several components were arranged into an extensible distributed multiagent architecture (Fig. 1, right): the open agent architecture (OAA). These components are in charge of a specific linguistic task. For instance, the language understanding and generation components were both modeled in the linguistic-driven formalism of multimodal-categorial grammars (MMCCG) [4]. DIPPER [5], which is the dialogue manager, follows the notion of *information state* (IS). Moreover, the KR was implemented in LISP and RACER was used as inference engine [6]. Further processing is carried out by the Interpreter, the Behavioral and the Generation agents that are part of the core of HomeNL whilst Festival was used for speech synthesis[1].

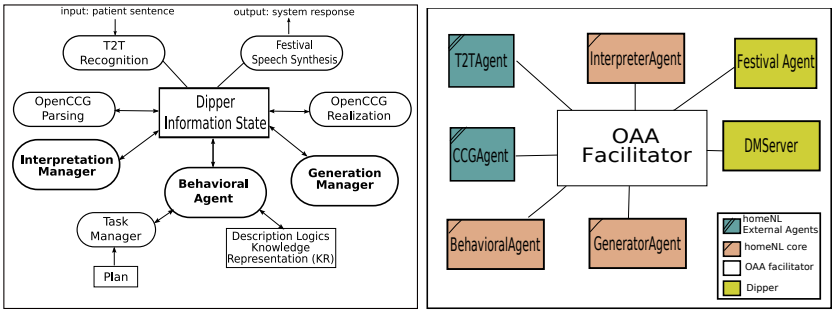


Fig. 1. Left: HomeNL architecture. Right: OAA-agents arrangement.

In its primary scope, HomeNL is concerned with the process of comprehending and using language once the words are recognized rather than the process of recognizing spoken sentences. Despite incorporating existing automatic speech recognition (ASR) inside the architecture is quite straightforward, much work remains to be done to support medical terminology in ASR's language models [3]. Therefore, the input is an Italian text sentence and the output is a coherent spoken response (Fig. 1, left). The *OpenCCG parsing* parses sentences and returns the corresponding semantic representation (SR), while the *Interpretation*

[1] http://www.cstr.ed.ac.uk/projects/festival/

Manager performs contextual interpretation. The *behavioral agent* reads the dialogue plan, goals and mandatory information before starting the dialogue. It also queries the KR and prepares the system response during the conversation. The *Generation Manager* transforms the system response into a semantic formula while *the OpenCCG realization* realizes that formula and returns the corresponding sentence in natural language. In turn, *Festival* transforms the sentence into voice.

2.1 Language Understanding and Generation

An Italian grammar was developed by using OpenCCG[2] for building MMCCG grammars. MMCCG is based on the lexicalized grammar formalism of combinatory categorial grammars (CCG), in which a grammar is defined in terms of *categories* in the *lexicon*. Categories define the part-of-the-speech of words and their relation with adjacent words in sentences. The lexicon is a complete dictionary of words that describes their lexical categories, linguistic (e.g., number, gender, etc.) and semantic features. Categories in a sentence are derived by applying combinatory rules that license the parsing and generation of several linguistic phenomena while the SR is built compositionally via unification. MMCCG has a mild context-sensitive generative power, thus it can handle a variety of constructions from a number of languages [4].

The grammar was implemented not only for parsing users' utterances but also for generating possible system responses. It is made up of a lexicon enriched with the morphology of words and a medical ontology. This ontology is used to build a SR that references ontological concepts. Nevertheless, the SR provided by OpenCCG cannot be used for reasoning [7]. The resulting grammar contains around 300 words including their inflected forms, grouped into 84 categories. These are further arranged in 65 lexical families. In spite of being a short grammar, it supports the Italian constructions commonly used by patients and doctors as collected by the Homey Hypertension Management project [8], used to either claim or inquire about health conditions, measurements, side effects, health status and habits. We are currently working on extending the CCG grammar in compliance with the Italian guideline for the management of hypertensive patients considering risk management and further constructions regarding symptoms and habits.

2.2 Description Logics Knowledge Representation

Description Logics (DLs) was selected to represent the medical knowledge [9]. It bears information about the patients' body, illness and therapy. In particular, it contains concepts related to active principles, medicines, symptoms and measurements e.g., blood pressure, weight and heartbeat. Concepts are defined as presented in Example 1, symptoms can be localized, if they affect a specific body-part, or unlocalized e.g., cough or fever. All the information about active principles and symptoms described in Table 1 has been formalized in a similar

[2] http://openccg.sourceforge.net/

fashion. Therefore, it is possible to perform inference tasks in RACER regarding the patient's side-effects and condition e.g., normal, low risk or serious.

$$
\begin{array}{l}
\textit{TBox} \\
\text{symptom} \equiv \quad (\text{local-symptom} \sqcup \text{non-local-symptom}) \\
\text{irregular-intestine} \equiv \quad \exists \text{affects-locally}.\, \textit{intestine} \sqcap \\
\qquad\qquad\qquad\quad \exists \text{has-property}.\, \textit{irregular} \\
\text{calcium-antagonist} \equiv \quad \exists \text{produces-secondary-eff.}\, (\text{swelling-leg} \sqcup \\
\qquad\qquad\qquad\quad \text{irregular-intestine} \sqcup \text{tachycardia}) \\
\ldots \\
\equiv \text{ equivalence } \sqcap \text{ conjunction } \sqcup \text{ disjunction } \exists \text{ existential quantifier}
\end{array}
\tag{1}
$$

Table 1. Knowledge about drug-types, medicines and possible side-effects for the management of patients with hypertension implemented in the knowledge base

Active Principle	Symptoms	Medicines
Aceinhibitor	cough	ramiprile, captoprile, enalaprile
Calcium-antagonist	swelling legs, irregular intestine, tachycardia	lacidipina
Diuretic	tiredness	furosemide, candesartan-hct
Betablocker	impotence	atenololo, atenololo-clortalidone, nebivololo
Alpha1bp	tachycardia	doxazosin
Alpha2ac	driesout mouth, driesout eyes, blush face, drowsy	clonidina

2.3 The Information State

The information state theory proposes a blackboard structure (i.e. the IS), a set of update rules and a set of dialogue moves to model dialogues [10]. In DIPPER the IS and the update rules are declared in the OAA-logical language, namely the interagent communication language (ICL). On the one hand, the IS keeps track of the relevant information of the discourse e.g., *cognitive state*, *goals*, *mandatory information*, *plan*, *input*, *SR*, *unsolved goals*, etc. In addition, the IS contains the information to be grounded (to be clarified) and *the common ground* with the information that has already been clarified, that is to say, the information shared by both HomeNL and the patient. On the other hand, the update rules implemented are *initialisation*, *recognise*, *parsing*, *interpreting*, *behavior*, *generating*, *realizing* and *synthesise*. The dialogue moves, also called dialogue-acts, deployed are *assertions*, *information requests* and *grounding* acts. Information requests are usually performed by HomeNL while assertions and grounding-acts are performed by both the system and the patient. Further details about HomeNL are given in [7].

3 Discussion and Future Work

Several levels of linguistic analysis (e.g., syntactic, semantics, pragmatics and discourse) were studied and implemented in HomeNL. The selection of MMCCG formalism for language understanding and generation pursued the improvement of home-care automated-dialogues by making the interaction less restrictive than those widely adopted systems that merely exploit the generative power of context

free grammars (CFG). Unlike similar prototypes in health that enable mixed-initiative through extensions to CFG, like augmented-CFG [3], here we developed and evaluated a categorial grammar formalism for language understanding and generation. The understanding capability was evaluated on the basis of the concept accuracy metric [11] giving an accuracy of 91.32%, whereas the generation capability has been tested in the realization of 74 sentences, giving an accuracy of 89%. We plan to evaluate HomeNL on the basis of usability test and we expect to improve patients' satisfaction of the system in comparison with HOMEY [8], the precedent dialogue for the management of patients with hypertension.

HomeNL can be improved by integrating an ASR and by filling up the KR with electronic health records (EHR) in order to support patient-tailored and telephone-linked dialogues. Furthermore, a planning agent can be incorporated in the architecture for solving dynamically changing goals at each interaction. Exploring risk reasoning in the KR is another appealing direction for future research.

References

1. Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L.: Towards conversational human-computer interaction (2001)
2. Bickmore, T., Giorgino, T.: Health dialog systems for patients and consumers. J. of Biomedical Informatics 39(5), 556–571 (2006)
3. Allen, J., Ferguson, G., Blaylock, N., Byron, D., Chambers, N., Dzikovska, M., Galescu, L., Swift, M.: Chester: towards a personal medication advisor. J. of Biomedical Informatics 39(5), 500–513 (2006)
4. Baldridge, J.: Lexically Specified Derivational Control in Combinatory Categorial Grammar. PhD thesis, School of Informatics. University of Edinburgh (2002)
5. Bos, J., Klein, E., Lemon, O., Oka, T.: Dipper: Description and formalisation of an information-state update dialogue system architecture (2003)
6. Haarslev, V., Moller, R.: RACE system description. In: Description Logics (1999)
7. Rojas-Barahona, L.M.: Health Care Dialogue Systems: Practical and Theoretical Approaches to Dialogue Management. PhD thesis, University of Pavia, Pavia, Italy (2009)
8. Giorgino, T., Azzini, I., Rognoni, C., Quaglini, S., Stefanelli, M., Gretter, R., Falavigna, D.: Automated spoken dialogue system for hypertensive patient home management. International Journal of Medical Informatics 74(1386-5056), 159–167 (2004)
9. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, Cambridge (2003)
10. Larsson, S., Traum, D.: Information state and dialogue management in the trindi dialogue move engine toolkit. Natural Language Engineering 6, 323–340 (2000)
11. Boros, M., Eckert, W., Gallwitz, F., Hanrieder, G., Niemann, H.: Towards understanding spontaneous speech: Word accuracy vs. concept accuracy. In: Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP 1996), pp. 1009–1012 (1996)

Explaining Anomalous Responses to Treatment in the Intensive Care Unit

Laura Moss[1], Derek Sleeman[1],
Malcolm Booth[2], Malcolm Daniel[2], Lyndsay Donaldson[2], Charlotte Gilhooly[2],
Martin Hughes[2], Malcolm Sim[2], and John Kinsella[2]

[1] Department of Computing Science, University of Aberdeen, Aberdeen, AB24 3UE
[2] University Section of Anaesthesia, Pain & Critical Care Medicine,
University of Glasgow, Glasgow, G31 2ER

Abstract. The Intensive Care Unit (ICU) provides treatment to critically ill patients. When a patient does not respond as expected to such treatment it can be challenging for clinicians, especially junior clinicians, as they may not have the relevant experience to understand the patient's anomalous response. Datasets for 10 patients from Glasgow Royal Infirmary's ICU have been made available to us. We asked several ICU clinicians to review these datasets and to suggest sequences which include anomalous or unusual reactions to treatment. Further, we then asked two ICU clinicians if they agreed with their colleagues' assessments, and if they did to provide possible explanations for these anomalous sequences. Subsequently we have developed a system which is able to replicate the clinicians' explanations based on the knowledge contained in its several ontologies; further the system can suggest additional explanations which will be evaluated by the senior consultant.

1 Introduction

Intensive Care Units (ICUs) provide treatment to patients who are often critically ill and possibly rapidly deteriorating. Occasionally a patient may not respond as expected to treatment; this can be considered as anomalous. An anomaly can be defined as 'a counterexample to a previous model of the domain'[10]. For example, based on knowledge of the ICU domain, it may be reasonable to expect that when a patient is administered the drug noradrenaline, it should *increase* a patient's blood pressure. However, if a *decrease* in a patient's blood pressure is observed, this would be a counterexample and considered anomalous. Such scenarios can be challenging for a clinician, especially as a similar event may not have been experienced previously. The focus of this study is the analysis of explanations given by two ICU consultants of patients' anomalous behaviour. Based on these analyses we are in the process of implementing a tool to replicate these explanations. The rest of the paper is structured as follows: section 2 provides a literature review, section 3 presents explanations for anomalous patient behaviour in the ICU and section 4 outlines an ontology-based tool which suggests explanations for anomalous scenarios.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 250–254, 2009.

2 Related Work

It is recognized that the ICU is a challenging domain in which to perform decision making[9]. Several intelligent data analysis systems have been developed to aid decision making in the ICU, e.g. RÉSUMÉ[11] and VIE-VENT[8]. Some systems have been implemented 'live' in the ICU, such as those developed by the Pythia/MIMIC[1] project; others use data 'offline' for example, ICONS[2], a case based reasoning system. Despite the wide variety of decision-support systems implemented in the ICU, none have focused on providing support to clinicians when faced with anomalous patient behaviour.

The generation of medical hypotheses from data has also been discussed widely in the literature, of most relevance to this work is Blum et al[7] which created hypotheses from a knowledge base and then verified these using statistical methods applied to patient data.

From a cognitive science perspective, it is widely acknowledged that anomalous scenarios provide a key role in knowledge discovery; an anomaly can indicate to an expert that their understanding of a domain may require further refinement which in turn may lead to the discovery of new (clinical) knowledge[4]. It is also known that experts can differ in their strategies when faced with anomalous data[5] [3].

3 Identifying and Explaining Anomalous Responses to Treatment

A senior consultant at Glasgow Royal Infirmary's ICU selected 10 patients from their repository and confirmed that a sizeable number of these records contained some anomalous sequences. Physiological data for these patients' complete stay in the ICU were made available to us from the unit's patient management system. A group of five further clinicians examined these datasets for sequences they thought involved anomalous behaviour. The clinicians were asked to 'talk-aloud' as they completed the task[6]. Protocol analysis[5] was performed on the transcripts by two analysts and yielded the following categories:

- **A** Anticipated patient responses to treatment, possibly with minor relapses (default if clinician does not provide any other classification)
- **B** Anticipated patient responses to treatment, with significant relapses e.g., additional bouts of sepsis, cardiac or respiratory failure
- **C** Patient not responding as expected to treatment
- **D** Odd / unusual set of physiological parameters (or unusual rate of change)
- **E** Odd / unusual treatment

In total, 65 anomalies (categories C-E) were identified by the clinicians. Figure 1 describes an anomalous response to treatment.As a further phase of this analysis, sequences which had been identified as including anomalous responses to treatment were presented to two further ICU clinicians, who were asked to provide as many explanations as possible for these sequences. A wide range of hypotheses were proposed which were organised as the following broad categories: 1) *clinical*

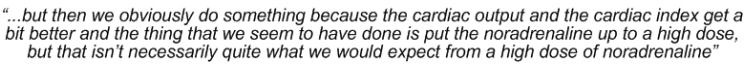


Fig. 1. An anomalous response to treatment as detailed in clinician 2's transcript

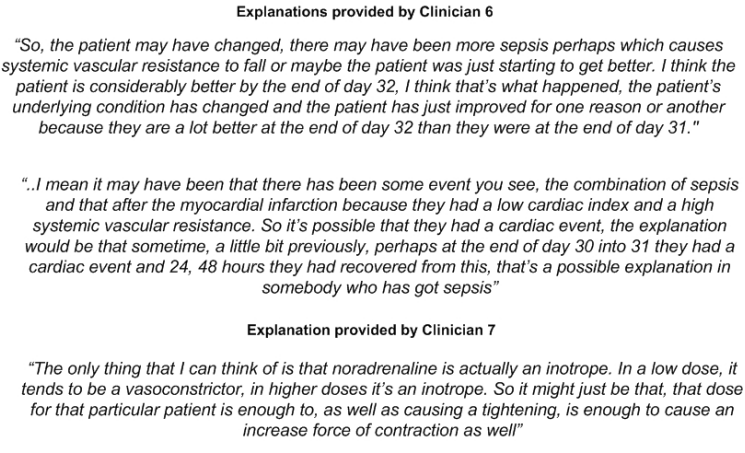


Fig. 2. Explanations given by Clinicians 6 and 7

conditions, 2) *hormone regulation*, 3) *progress of the patient's condition*, 4) *treatment*, 5) *organ functioning* and 6) *errors in recordings*. For example, in response to the anomaly detailed in Figure 1, the first clinician suggested sepsis (*clinical conditions*), an improvement in the patient's condition (*progress of the patient's condition*) and a combination of sepsis and myocardial infarction (*clinical conditions*) as potential explanations (Figure 2)[1].

These interviews were analysed further and a method of information selection and hypothesis generation used by the clinicians was proposed. Figure 3 illustrates this general model of hypothesis generation. Beginning with an anomaly, for example, *noradrenaline increased cardiac output and cardiac index*, it can be broken down into the treatment, '*noradrenaline*' and the effect '*increase cardiac output and cardiac index*'. The clinician then proceeds to explain any combination of the anomalous treatment and effects through the various routes shown. The clinician appeared to use domain knowledge about treatment, medical conditions and the desired physiological state of the patient to explain the treatment or effect. Further, the domain knowledge can also be applied whilst examining the data to determine facts; for example, the patient is suffering from a myocardial infarction. In addition, the patient's data can be used to eliminate hypotheses. For example, one of the explanations for the anomaly detailed in Figure 1 was that the patient may be getting better, if the data does not show this, the hypothesis could be eliminated. After suggesting a hypothesis, the clinician repeats the process until they are satisfied that all viable hypotheses have been proposed.

[1] Both clinicians also identified that the patient had an abnormally low systemic vascular resistance (SVR).

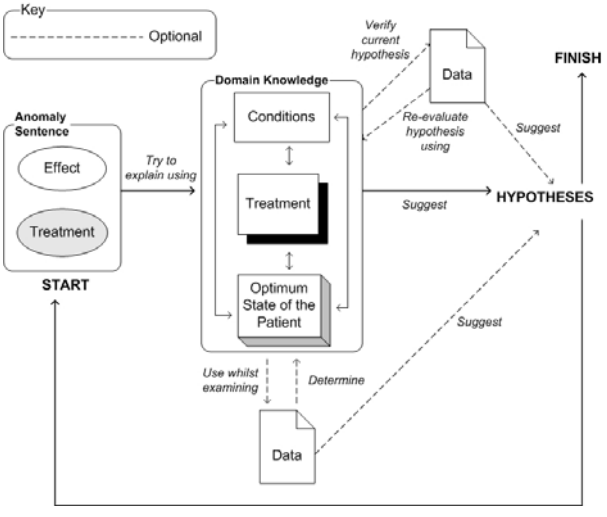


Fig. 3. General Model of Hypothesis Generation

4 Ontology-Based Explanations of Anomalous Responses to Treatment

The model of hypothesis generation (Figure 3) forms the basis for an ontology-based hypothesis generation tool. In the initial stage, various methods (Figure 3) of querying the knowledge base and the patient data are used to generate a list of potential hypotheses for a given anomaly. The knowledge base comprises a set of ontologies coded in OWL containing the following concepts a) *Treatments* b) *Disorders* c) *Acceptable Parameters* and d) *Physiological Data*. The suggested hypotheses will subsequently be evaluated by an ICU clinician for clinical relevance. Building on this initial stage the work will be extended to explore the domain knowledge further. For example,

- Suppose: It had *not* been noted in the ontology that noradrenaline can, in high doses, increase a patient's cardiac output
- Observed Anomaly: The patient's cardiac output increased when the patient was on high doses of noradrenaline (as described in Figure 1)
- Known facts from knowledge base: *1)* Inotropes (a class of drugs) increase cardiac output *2)* Noradrenaline is a vasoconstrictor
- Conclusion: In this circumstance (high dose), noradrenaline is acting as an inotrope

5 Conclusions and Further Work

In this paper we have suggested a classification for the types of anomalies identified in the ICU domain and subsequently the types of explanations for such anomalies provided during interviews with domain experts. An initial system has

been outlined to replicate the generation of these explanations. Planned future work involves a systematic evaluation of this system and enhancements namely a) the system could be extended to automatically detect anomalous scenarios in the patient data rather than rely on them being highlighted by a clinician and b) the system could explore more extensively both the data and the ontologies to suggest new hypotheses not currently contained in the knowledge base, for example, a new side effect of a drug not currently recorded in the treatments ontology.

Acknowledgments

- Kathryn Henderson and Jennifer McCallum (CareVue Project, Glasgow Royal Infirmary) & the staff and patients of the ICU Unit, Glasgow Royal Infirmary.
- This work was an extension of the routine audit process in Glasgow Royal Infirmary's ICU; requirements for further Ethical Committee Approval has been waved.
- This work has been supported under the ESPRCs grant number GR/N15764.

References

1. Pythia Project, http://mimic.mit.edu/index.html
2. Heindl, B., Schmidt, R., Schmid, G., Haller, M., Pfaller, P., Gierl, L., Pollwein, B.: A case-based consiliarius for therapy recommendation(ICONS):computer-based advice for calculated antibiotic therapy in intensive care medicine. Computer Methods and Programs in Biomedicine 52, 117–127 (1997)
3. Chinn, C.A., Brewer, W.F.: An Empirical Test of a Taxonomy of Responses to Anomalous Data in Science. Journal of Research in Science Teaching 35, 623–654 (1998)
4. Kuhn, D.: The structure of scientific revolutions. University of Chicago Press (1962)
5. Alberdi, E., Sleeman, D., Korpi, M.: Accommodating Surprise in Taxonomic Tasks: The Role of Expertise. Cognitive Science 24, 53–91 (2000)
6. Anders Ericsson, K., Simon, H.A.: Protocol analysis: verbal reports as data. MIT Press, Cambridge (1993)
7. Walker, M.G., Wiederhold, G.: Acquisition and Validation of Knowledge from Data. Intelligent Systems, 415–428 (1990)
8. Miksch, S., Horn, W., Popow, C., Paky, F.: VIE-VENT: knowledge-based monitoring and therapy planning of the artificial ventilation of newborn infants. Artificial Intelligence in Medicine 10, 218–229 (1993)
9. Patel, V.L., Zhang, J., Yoskowitz, N.A., Green, R., Sayan, O.R.: Translational cognition for decision support in critical care environments: A review. Journal of Biomedical Informatics 41, 413–431 (2008)
10. Bridewell, W., Buchanan, B.G.: Extracting Plausible Explanations of Anomalous Data. Technical report
11. Shahar, Y., Musen, M.A.: Knowledge-Based Temporal Abstraction in Clinical Domains. Artificial Intelligence in Medicine 8, 267–298 (1996)

Multiple Terminologies in a Health Portal: Automatic Indexing and Information Retrieval

Stéfan J. Darmoni[1], Suzanne Pereira[1,2,3], Saoussen Sakji[1], Tayeb Merabti[1], Élise Prieur[1], Michel Joubert[2], and Benoit Thirion[1]

[1] CISMeF, LITIS EA 4108, University of Rouen, Normandy, France
[2] LERTIM, Marseille Medical University, France
[3] VIDAL, Issy les Moulineaux, France

Abstract. Background: In the specific context of developing quality-controlled health gateways, several standards must be respected (e.g. Dublin Core for metadata element set; thesaurus MeSH as the controlled vocabulary to index Internet resources; HON code to accredit quality of health Web sites). These standards were applied to create the CISMeF Web site (French acronym for Catalog & Index of Health Internet resources in French). Objective: In this work, the strategic shift of the CISMeF team is intended to index and retrieve French resources not anymore with a single terminology (MeSH thesaurus) but with the main health terminologies available in French (ICD 10, SNOMED International, CCAM, ATC). Methods & Results: Since 2005, we have developed the French Multi-Terminology Indexer (F-MTI), using a multi-terminology approach and mappings between health terminologies. This tool is used for automatic indexing and information retrieval. Conclusion: Since the last quarter of 2008, F-MTI is daily used in the CISMeF production environment and is connected to a French Health Multi-Terminology Server.

1 Introduction

Regardless of their web experience and general information retrieval skills, users have difficulties in seeking health information on the Internet [1]. In this context, several quality-controlled health gateways have been developed [2]. Quality-controlled subject gateways (or portals) were defined by Koch [2] as Internet services which apply a comprehensive set of quality measures to support systematic resource discovery. Among several quality-controlled health gateways, CISMeF ([French] acronym for Catalog and Index of French Language Health Resources on the Internet) was designed to catalog and index the most important and quality-controlled sources of institutional health information in French in order to help health professionals, patients and students to find electronic medical information available on the Internet quickly and precisely [3]. To respect quality standards, CISMeF is accredited by the Health on the Net Foundation since 1998 [4]. In the catalog, all the resources are described with 11[1] out of

[1] Inclusion: title, creator, subject, description, publisher, date, type, format, identifier, source and language. Exclusion: relation, coverage, rights and contributor.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 255–259, 2009.

15 Dublin Core (DC) metadata set [5] including the title and resource types, and indexed with a set of indexing terms to describe the information content. Since 1995 (creation of CISMeF in February), the indexing terms are descriptor/qualifier pairs or descriptors from the MeSH® thesaurus (Medical Subject Headings), the U.S. National Library of Medicine's (NLM's) controlled vocabulary used to index articles from the biomedical literature. From 1995 to 2002, CISMeF was exclusively manually indexed by a team of four indexers, which are medical librarians and systematically checked by the chief information scientist[2]. The objective of this paper is to describe the strategic shift to use several health terminologies for the automatic indexing and the information retrieval in the CISMeF quality-controlled health portal vs. the previous use of only one medical terminology (the MeSH thesaurus).

2 Methods

2.1 Automatic Indexing

Since 2002, faced with the growing amount of online resources to be indexed and included in the catalog, the CISMeF team consistently evaluated advanced automatic MeSH indexing techniques. The automatic indexing tools used primarily natural language processing (NLP) and K-nearest neighbours (KNN) methods [6], followed by a simpler bag of words algorithm [7]. The latter was successfully evaluated in the context of teaching resources. In August 2006, the CISMeF team decided to use this algorithm in the daily practice for most of the Internet resources rated as "low priority" resources (except guidelines which are still manually indexed because this type of resources rated as "high priority" need in-depth indexing). These "low priority" resources are teaching resources or resources belonging to a topic substantively covered in the catalog that do not require in-depth indexing.

Since 2005, the CISMeF team and the Vidal company developed the F-MTI tool [8]; the goal of the CISMeF team was to use a new automated indexing tool to index health resources in CISMeF; from a bag-of words algorithm based on a mono-terminology approach, we choose to use the F-MTI tool based on several health terminologies. In 2006, besides the MeSH, four health terminologies were included in F-MTI: ICD-10 (International Classification of Diseases) and SNOMED 3.5 (Systematized Nomenclature of Medicine) which are included in the UMLS, CCAM (the French equivalent of US CPT) and TUV (a French terminology for therapeutic and clinical notions for the use of drugs), which are not included in the UMLS. These four terminologies are mapped to the French MeSH. Several formal evaluations were performed with each of these terminologies [8], including the latest with the MeSH thesaurus [9]. During 2008, four new terminologies or classifications were added: ATC classification (N=5,514), drug names with international non-proprietary names (INN) and brand names

[2] Its URLs are http://www.chu-rouen.fr/cismef or http://www.cismef.org

(N=22,662), Orphanet thesaurus for rare diseases (N=7,421), MeSH Supplementary Concepts translated in French by the CISMeF team (N=6,004 out of over 180,000). Currently, after the formal evaluation of the F-MTI to index health resource [9], this tool F-MTI has enabled the automatic indexing of 33,951 resources in the CISMeF catalog and the semiautomatic (or supervised) indexing[3] of another 12,440 resources based on resources titles; overall, 65,242 resources are included in this gateway. Three levels of indexing were defined in the CISMeF catalogue:

- Level 1 or Core-CISMeF (N=18,851) which is totally manually indexed resources (e.g. guidelines).
- Level 2 or supervised resources (N=12,440): these resources are rated by the CISMeF editorial board as less important than level 1. These resources do not need in-depth indexing (e.g. technical reports, teaching resources designed at the national level, document for patients from medical specialties).
- Level 3 or automatically indexed resources (N=33,951). The CISMeF editorial board has rated these resources as less important than level 1 and level 2 (e.g. teaching resources designed at the medical school level, patient association Web sites).

Since CISMeF achieved this milestone in automatic indexing, it strived to improve the automatic indexing algorithm and make it on par with manual indexing. One of the challenges that the CISMeF automatic indexing algorithm needs to address is identifying all the different forms a term can take in natural language, specifically with respect to lexical and grammatical variations. Most terminologies such as MeSH provide synonyms and variants for the terms but this information is usually insufficient to describe all the forms that can be encountered for a given term in a document.

To allow automatic indexing using multiple terminologies to be used in the CISMeF catalog, the CISMeF team in collaboration with eight 4th year students of the INSA of Rouen engineering school has to integrate the previously listed health terminologies in the CISMeF back-office. To do so, for each terminology, a UML model and a parser were developed. A generic model was also developed to allow inter-terminology interoperability. Each specific terminology is integrated in an OWL format into a health multi-terminology server (French acronym SMTS) using the ITM® model (Mondeca) for implementation. From this SMTS, all the health terminologies were uploaded in the CISMeF backoffice. Then, F-MTI automatic indexing results are able to be easily integrated in the CISMeF backoffice and allowing multi-terminology information retrieval.

2.2 Information Retrieval

Since the overall CISMeF structure has evolved from a mono-terminological world (based on the MeSH thesaurus) to a multi-terminological universe, similarly the information retrieval algorithm has been modified. The formulation

[3] Supervision means that these resources are primarily indexed automatically, and then this indexing is reviewed by a CISMeF human indexer, who is a medical librarian.

of the requests consists in re-writing a user query in order to conceive queries closer to the expected needs. The Doc'CISMeF search engine carries out a comparison between character strings. Currently, the task of query reformulation is carried out by listing all the possible combinations of the bag of words query terms (terms obtained following the initial user query treatment by eliminating the blank words and by stemming the terms) in order to find the maximum of possible correspondences with the documents descriptors- the terms considered as the most significant. Indeed, pairing resource-query is performed by a comparison of what could exist as correspondence between the query terms and the resources descriptors. Several operations are done during this process such as: natural language treatment techniques for the multi terms query, phonemisation, terms adjacency. To match as much as possible queries with the CISMeF corpus, we have implemented a three-step heuristics. The process consists in recognizing a user query expression.

- Step 1. The reserved terms or the document's title: If the user query expression matches CISMeF terminology terms or the document's title, the process stops, and the answer to the query is the union of the resources that are indexed by the query terms, and those that are indexed by the terms they subsume, directly or indirectly, in all the hierarchies they belong to. This step is modified since the implementation of the multi-terminology. This process is generalized to match to any descriptor belonging to multiple health terminology, and the generalization includes also the list of terms.
- Step 2. The CISMeF metadata: The search is performed over all the other fields of the CISMeF metadata (abstract, author, publisher, identifier ...).
- Step 3. Adjacency of words in the text: A full text search over the document with adjacency of n words with n= 10 × (number of words of the query - 1) is realised.

The steps 2 & 3 are similar in the multi-terminology context than previously in the mono-terminology context. By default and for each of the three steps, CISMeF displays the query results starting with the most recent document and displays the resources indexed with the MeSH Major headings, which express the main topic and then the resources indexed with the MeSH minor headings, which produce a complementary information about the indexing. The resources that were indexed automatically in the catalogue are displayed after those that were indexed manually.

3 Conclusion and Perspectives

As far as we know, the F-MTI automatic indexing tool is the first attempt to use multiple health terminologies besides the English (specially the MTI initiative of the US National Library of Medicine). In the near future, a formal evaluation of F-MTI will be performed on French scientific articles included in the MEDLINE database: a French/English comparison is feasible on this MEDLINE subset.

Information retrieval using multiple terminologies will also be compared to the previous information retrieval based only on the MeSH thesaurus. The information retrieval using multiple terminologies will be adapted on a new context: to retrieve reports from the electronic health record of patients.

Acknowledgments

This research was partially supported by the European Union funded project FP7-ICT-1-5.2-Risk Assessment and Patient Safety (PSPIP) (n°216-130), and the ANR-funded project ANR-07-TECSAN-010. The authors would like to thank CISMeF indexers for their help in the study design and result analysis.

References

[1] Keselman, A., Browne, A.C., Kaufman, D.R.: Consumer Health Information Seeking as Hypothesis Testing. J. Am. Med. Inform. Assoc. 15(4), 484–495 (2008)
[2] Douyère, M., Soualmia, L.F., Névéol, A., Rogozan, A., Dahamna, B., Leroy, J.P., Thirion, B., Darmoni, S.J.: Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Info. Libr. J. 21(4), 253–261 (2004)
[3] Koch, T.: Quality-controlled subject gateways: definitions, typologies, empirical overview, Subject gateways. Online Information Review 24(1), 24–34 (2000)
[4] Boyer, C., Gaudinat, A., Baujard, V., Geissbühler, A.: Health on the Net Foundation: assessing the quality of health web pages all over the world. Stud Health Technol Inform. 129(Pt 2), 1017–1021 (2007)
[5] Dekkers, M., Weibel, S.: State of the Dublin Core Metadata Initiative. D-Lib Magazine 9(40) (2003)
[6] Névéol, A., Rogozan, A., Darmoni, S.J.: Automatic indexing of online health resources for a French quality controlled gateway. Information Management & Processing 1, 695–709 (2006)
[7] Névéol, A., Pereira, S., Kerdelhué, G., Dahamna, B., Joubert, M., Darmoni, S.J.: Evaluation of a simple method for the automatic assignment of MeSH descriptors to health resources in a French online catalogue. Medinfo., 407–411 (2007)
[8] Pereira, S.: Multi-terminology indexing of concepts in health. [Indexation multi-terminologique de concepts en santé]. PhD Thesis, University of Rouen, Normandy, France
[9] Pereira, S., Neveol, A., Kerdelhué, G., Serrot, E., Joubert, M., Darmoni, S.: Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue. In: AMIA Annu. Symp. Proc., November 6, pp. 586–590 (2008)
[10] Joubert, M., Dahamna, B., Delahousse, B., Fieschi, M., Darmoni, S.: SMTS®: Un Serveur Multi-Terminologies de Santé. In: Informatique & Santé, Journées Francophones d'Informatique Médicales (in press)
[11] Soualmia, L., Dahamna, B., Thirion, B., Darmoni, S.J.: Strategies for health information retrieval. Stud Health Technol. Inform. 124, 595–600 (2006)

CodeSlinger: An Interactive Biomedical Ontology Browser

Jeffery L. Painter and Natalie L. Flowers

GlaxoSmithKline, Research Triangle Park, NC 27709, USA

Abstract. CodeSlinger is a highly interactive and semi-intelligent application designed to support the search and navigation of large biomedical coding schemes, thesauri, and ontologies. We discuss how CodeSlinger is used by epidemiologist/physicians in the creation of coding sets for data extraction and analysis, the exploratory nature of the application, and finally, the issues facing our knowledge-representation model and extension of the UMLS.

1 Introduction

There are several issues facing medical informatics today especially with regard to the identification and classification of medical "concepts" in controlled vocabularies, dictionaries, thesauri and ontologies [1].

Coping with the plethora of information available is a matter of utmost importance for both the efficacy with which we are able to make use of medical data, as well as insuring that the results we produce can be viewed with the confidence that reported analysis are accurate and inclusive of the appropriate populations. In response, we developed a multi-user client/server based application called *CodeSlinger* to assist the informaticists with at least one dilemma they face in the process of data extraction. *CodeSlinger* empowers the user to promptly and accurately identify the appropriate set of medical codes relevant to their studies. It also gives them a higher degree of confidence that they are no longer "missing" data which might be relevant to their investigations.

2 Background

In almost every medical records database, one or more "coding schemes" are employed to represent the medical concepts within it. The medical concepts can include drugs, devices, symptoms, conditions, procedures etc. The broad scope of a medical concept is just one of many difficulties when dealing with medical informatics [2].

CodeSlinger helps the user to aggregate sets of codes for the purpose of data extraction and analysis. Until now, the state of the art for this task involved (1) using large medical coding books, (2) relying on one's medical expertise from use or experience with a particular coding scheme, (3) studying the literature to see what sets of codes have been used in a study before and of course (4) using Google and other search engines to locate medical codes.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 260–264, 2009.

2.1 Code Selection

Our goal was to provide an application where the user could simply focus on the concepts of interest, and by making use of the highly interactive visual display of codes and their relations, quickly and with confidence develop a set of codes which properly characterizes a medical concept in the database(s) under review.

The databases the informaticists must deal with are also disparate with regard to their format and content; and comparison and analysis of data across such disparate sources requires some way of translating among the coding scheme representations (or "normalizing" the references) so that references to the "same disease", "same condition", "same procedure", or "same drug" may be identified. For example, an epidemiologist may want to extract a cohort of patient data for a study on "heart failure" from two databases that have been coded using ICD-9 [3] in one and MedDRA [4] in the other. This leads to the question of how does one then find the corresponding code maps between coding schemes which would allow the researcher to ensure that the selected populations are comparable? That is, the selected cohorts are representative of the *same* concept or condition.

This challenge led to the development of *CodeSlinger* as a semi-automated code mapping and navigation system. While there are other electronic resources available for mapping codes from one scheme to another, including the UMLS Knowledge Server [5] and TermWorks [6], we focused on developing an application that would be highly interactive and provide visual queues to assist the user in code selection while allowing for the exercise of the user's medical expertise.

In our example (see Fig. 1), the user runs a search for "heart failure", and a set of results is returned in which they will find several codes under both ICD-9 and MedDRA. Findings under ICD-9 include the code "428.1 - Left heart failure". The user may then reveal the possibly related codes by highlighting the code in the user interface. One relation then displayed is the MedDRA code "10024119 - Left ventricular failure". *CodeSlinger* supports several concept mappings that we have exploited from the UMLS system noted in a previous paper [7]. This allows the user to more easily explore the inter-relations among source vocabularies.

3 Interface

As illustrated in Fig. 1, the interface of CodeSlinger is made up of a search box, a results box, coding scheme browser(s), and a final list box (used to compile the code sets that will be used to perform data extraction). The search box at the top of Fig. 1 allows the user to select which coding schemes are of interest for a particular search.

After the search results are displayed, the user can explore each entry by clicking on the code or term to see its related concept maps in each of the sources chosen at the beginning of the search. And, if any alternate codes are associated with that particular code, those codes are made clear in the "Alternate Terms or Codes" window pane.

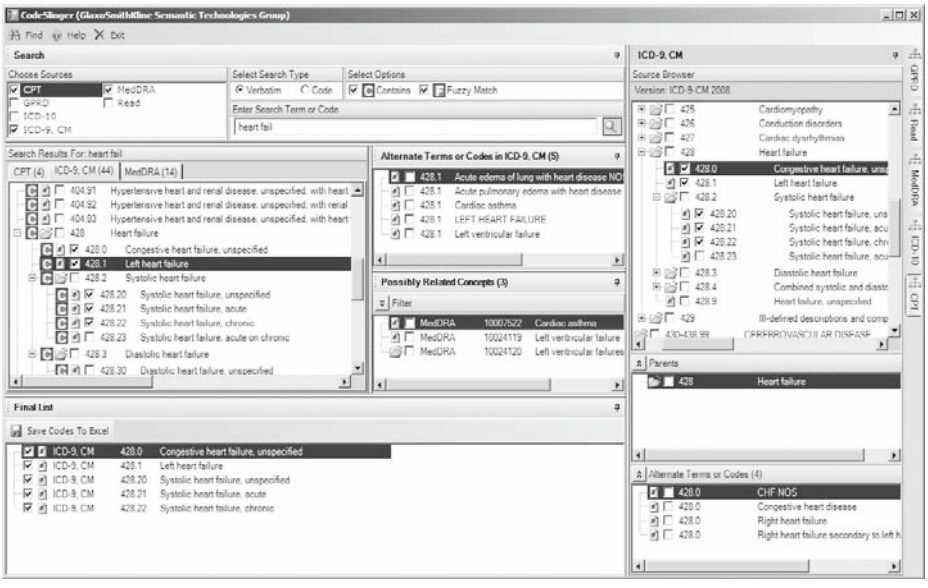


Fig. 1. CodeSlinger Component Layout View

Additionally, a term can be double clicked and the application will open the coding scheme browser (found on the right side of Fig. 1), displaying the code in its hierarchical context, allowing for further visual search and navigation.

At any time, the user can click the check box next to a code to add or remove it from their final list found at the bottom of Fig. 1. Once the code set is collected, the user can export the list as a simple spreadsheet by clicking the icon to save the codes to Excel.

4 Knowledge Representation

Much of the benefit our users experience with *CodeSlinger* is based on the fundamental knowledge representation created to facilitate the search, mapping and exploration of the "concepts" extracted from the UMLS Metathesaurus [8]. Where in most applications based on the UMLS, developers strictly take an unmodified database view of the UMLS content, we have created our own set of tools to extract and manipulate relations directly from the RRF files. We also construct our own custom database, which is comprised of a relatively minimal number of tables used to store source hierarchies, terms and relations of the concept model. The *CodeSlinger* server loads an object model of the hierarchical structures along with active concept nodes which are self-aware of both their attributes (atom identifiers, terms, alternate terms, etc) and relationships to other concept nodes within the terminology server.

The hierarchical trees are indexed to enhance the lookup and retrieval time based on key elements of the concept attributes, including *source, term* and

concept identifiers which are tied to the concept nodes in a cost conscious (in terms of both memory and speed) manner. We preserve the atom, concept and string identifiers found in the UMLS to enable quick retrieval of the information requested by the search interface.

5 Constructing a Search

The search interface allows the user several options to refine and constrain the search results. The user begins by choosing which vocabularies are of interest for the search query. Currently ICD-9, ICD-10, Read (Clinical Terms Version 3), Current Procedural Terminology (CPT), MedDRA, and a customized representation of the OXMIS coding scheme are supported. Next, the user has the option to select whether the search is for a code or term. In some cases the may know a particular code and want to see what other codes may have relevance. Finally, the user may choose from one of two search algorithms or both: (1) a simple *contains* search – indicating that the results must contain either the exact code or search string entered, or (2) a *fuzzy* match.

5.1 Contains Search

Our *contains* search ignores word order and case distinctions. This allows us to match "heart failure" to "Heart Failure", "failure of the heart" and "failure, heart" since each of the terms contains both the individual words "heart" and "failure". While this may return more than the number of items the user may like to see, the stance we chose to take in many aspects of the application was to "cast a wide net" and let the user employ medical expertise to evaluate and discriminate the output.

5.2 Fuzzy Match

The *fuzzy* match generates a Bayesian model to approximate the probability of a lexical match based on the user's search string. The *fuzzy* match is capable of dealing with misspellings and minor word variations. For example, the terms "color" versus "colour" would be matched on either spelling using our *fuzzy* match. The strings are first normalized in the following manner.

1. Tokenize and case fold the terms
2. Remove contractions and parenthetical plurals
3. Apply standard stemming algorithms
4. Remove stop words (customized for our domain)

The normalized string is converted to a sorted bi-gram sequence [9] and stored in our database. When the user submits a new search, the *fuzzy* match first normalizes the search string and converts it to a bi-gram sequence as well. A probability value is then assigned between the search sequence and each of the pre-computed sequences. If the probability of match is high enough, the source

term is matched to a set of possible "concepts" and the corresponding codes are then included in the results. We can tune the output by changing the lower bound required for a match, but this is not a user configurable option. Our best user experience results when the lower bound is set within: $p \in [0.55 - 0.60]$.

6 Results

It should be noted that we are investigating several avenues of releasing the *CodeSlinger* application to the medical informatics community. It is our goal to make this application available to those who would most benefit from its development.

A public demonstration of *CodeSlinger* was given to the UMLS user community in 2008. The application was well received and many favorable comments and feedback were given which will be incorporated in future versions of the application. Also, after giving internal demonstrations, we identified a major initiative where we were able to assist in improving the level of concept coverage for a new internal medical conditions dictionary within GlaxoSmithKline.

Acknowledgements

The authors would like to thank Drs. Kathleen Beach, MD, MPH and Hoa Le, MD for help in defining the requirements for *CodeSlinger* and providing valuable feedback during its development.

References

1. Hasman, A.: Challenges for medical informatics in the 21st century. International Journal of Medical Informatics 44, 1–7 (1997)
2. Cimino, J.J.: Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. Methods Inf. Med. 37(4-5), 394–403 (1998)
3. ICD-9 refers to ICD-9, CM the International Classification of Diseases, 9th Revision, Clinical Modification
4. MedDRA (Medical Dictionary for Regulatory Activities) is a registered trademark of the International Federation of Pharmaceutical Manufacturers
5. UMLS Knowledge Source Server is a project of the (US) National Library of Medicine, http://umlsks.nlm.nih.gov/
6. TermWorks is copyrighted by Apelon, Inc., http://www.apelon.com/products/termworks.htm
7. Kleiner, K., Merrill, G.H., Painter, J.L.: Inter-translation of Biomedical Coding Schemes Using UMLS. In: Proceedings of the 2006 AAAI Fall Symposium on Semantic Technologies (2006)
8. UMLS Metathesaurus is a project of the (US) National Library of Medicine, Department of Health and Human Services, http://www.nlm.nih.org/research/umls/
9. El-Nasan, A., Veeramachaneni, S., Nagy, G.: Word Discrimination Based on Bigram Co-Occurrences. In: Sixth International Conference on Document Analysis and Recognition (ICDAR 2001), p. 0149 (2001)

Subgroup Discovery in Data Sets with Multi–dimensional Responses: A Method and a Case Study in Traumatology

Lan Umek[1], Blaž Zupan[1,2], Marko Toplak[1], Annie Morin[3], Jean-Hugues Chauchat[4], Gregor Makovec[5], and Dragica Smrke[5]

[1] Faculty of Computer and Information Sciences, University of Ljubljana, Slovenia
[2] Dept. of Human and Mol. Genetics, Baylor College of Medicine, Houston, USA
[3] IRISA, Universite de Rennes 1, Rennes cedex 35042, France
[4] Universite de Lyon, ERIC-Lyon 2, 69676 Bron Cedex, France
[5] Dept. of Traumatology, University Clinical Centre, Ljubljana, Slovenia

Abstract. Biomedical experimental data sets may often include many features both at input (description of cases, treatments, or experimental parameters) and output (outcome description). State-of-the-art data mining techniques can deal with such data, but would consider only one output feature at the time, disregarding any dependencies among them. In the paper, we propose the technique that can treat many output features simultaneously, aiming at finding subgroups of cases that are similar both in input and output space. The method is based on k-medoids clustering and analysis of contingency tables, and reports on case subgroups with significant dependency in input and output space. We have used this technique in explorative analysis of clinical data on femoral neck fractures. The subgroups discovered in our study were considered meaningful by the participating domain expert, and sparked a number of ideas for hypothesis to be further experimentally tested.

Keywords: subgroup discovery, multi–label prediction, k-medoids clustering, χ^2 statistics, femoral neck fracture.

1 Introduction

According to Hand [1], data mining tasks can belong to either of two types: (1) construction of models describing global relations between independent variables (attributes) and a selected dependent variable (class), like classification or regression, and (2) the search for the local patterns in the attribute space where the class behaves unusually. The two tasks are also often referred to as predictive and descriptive data mining [2]. While both approaches may present their results in an interpretable way, it is really the locality of the discovered patterns that distinguishes between them. In biomedical research, any patterns that relate descriptions of the experiments with their outcomes may be valuable and may lead to discovery of new knowledge, regardless whether they apply to the entire population or only to the subgroup of the examined cases.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 265–274, 2009.

The interest in finding the patterns that refer to a subset of examined cases gave rise to a field of *subgroup discovery*, for which researchers have recently adapted a number of known data mining techniques, including CN2 [3, 4] and APRIORI [5, 6]. These approaches are mainly concerned with data sets that include a single class variable, that is, a single feature that describes the outcome of the experiments. This is much in line with the above Hand's definition (local patterns describing a single class). In biomedicine, there are many cases, however, where the outcome of the experiment is described by many features. For instance, in high–throughput genomics the experimental studies may observe responses of many (single-mutant) strains to a defined stimuli. Similarly, in chemical genomics, the effect of a drug may be recorded across different strains. Gene expression studies do often record the responses under many different experimental conditions. And in clinical medicine, the health of the patients that underwent a certain medical procedure could be assessed through a number of different clinical observations.

Clearly, there is a need for a tool that uses subgroup discovery for analysis of data sets with multi-dimensional responses. The input of such tool would therefore be an attribute-valued description of the inputs (*e.g.* patient's pre-treatment status and description of the underwent treatment) and the attribute-valued description of the outcomes (*e.g.* incorporating features describing patient's status after the treatment).

Perhaps the precursor of the set of methods that could address such problems were Blockeel *et al.*'s clustering trees [7]. Their technique could be applied to above problems such that the distances between cases would be measured on outcome features, while the attributes for the inference of the tree would be those from the input set. Notice, however, that just like classification trees, their better-known predecessor, the clustering trees model the entire data set. This is also the case for the approach called clustering rules [8], but mitigated since the rules, unlike the leaves in clustering tree, can overlap, thanks to the weighted coverage approach borrowed from CN2-SD [4]. While its authors have designed it to consider only single-variate outcomes, APRIORI-SD [5] could also be extended to treat multi-dimensional responses. As a last resort, multi-dimensional response problems could be considered by "one-response-feature-at-a-time", but that would at the start prevent the discovery of any relations between outcome features and increase the risk of false discoveries (overfitting of the data).

In the paper, we propose an original technique for subgroup discovery which is designed to deal with the data sets that include many input and output features. The method's principal idea is to separately find clusters of cases in input and output feature space and examine their overlap. A proposed procedure then reports on case subgroups, that is, cases that are statistically similar both in input and output features. With the aim to provide a concise and readable description of the subgroups, we also apply a set of heuristics to merge similar subgroups and describe them only with the most representative features.

We have applied the proposed technique in the explorative analysis study of the clinical data from traumatology. As provided, the data already included interesting groups of features that we split to those for the input and those describing the outcome. Admittedly, this is precisely the type of the data that inspired the development of our approach. We present some of the discovered subgroups in the paper, and rely on the domain-expert's enthusiasm while commenting on the potential utility of the discovered patterns when reporting on the success of the application and the proposed approach.

2 Methods

We have developed a method which tries to find local patterns where an outcome is a realization of a high–dimensional vector with the components of mixed–type (nominal or continuous). We will assume that our data is a random sample of n realizations of independent identically distributed random vectors X (attributes) and Y (classes). Realizations $(x_1, y_1), \ldots, (x_n, y_n)$ will be identified with instances (cases) $e_1, \ldots, e_n$ with relation $(X, Y)(e_i) = (x_i, y_i)$. With this notation a *subgroup* G is a subset of $\{e_1, \ldots, e_n\}$.

The aim of the proposed algorithm is the identification of statistically exceptional case subgroups of sufficient (user-specified) size. The method's main idea is to reduce the complexity of the possible links between the two sets of variables X and Y in a two–step process:

1. summarize the variability among the variables X (respectively Y) by inferring a cluster system L_x (respectively L_y),
2. study the relation between X and Y by studying the cross–table $L_x \times L_y$.

Knowing which variables belong to X and which to Y we define that a subgroup G is interesting if:

1. it is sufficiently large (a user-defined criteria),
2. the instances belonging to G are similar to each other based on values of features from X and Y, and
3. features in X and Y are conditionally dependent given the subgroup G.

2.1 Clustering in Input and Output Feature Space

We have used k–medoids clustering [9] to separately find instance clusters in input and output space with distances defined using Manhattan metric. The continuous variables were normalized by their ranges and all possible values of categorical variables were considered equally dissimilar. This partitioned the entire set of instances to a predefined number of disjoint clusters by minimizing the sum of distances between members of the clusters and their centres. The k–medoids approach uses a medoid (an instance of the cluster whose average dissimilarity to all other instances in the clusters is minimal) as the cluster's centre.

A common problem in clustering is the determination of appropriate number of clusters. Several different methods have been proposed, among them Tibshirani's gap statistics [10], silhouette method [11] and INCA [12]. As they estimate the number of clusters in the entire data set they are not necessarily optimal for local patterns. Therefore, we used an alternative approach. To avoid missing any of the interesting groups due to "wrong" choice of k, our algorithm tries all k in a given range. Since, consequently, probability of overfitting is increased, the number of groups was reduced in the post-processing (section 2.3).

2.2 The Search Algorithm

The input to the search algorithm is:

- a data set with features in input X and output Y feature set,
- parameters k_{1Max}, k_{2Max} representing upper bounds for number of clusters in X–space and Y–space,
- minimal subgroup size min_{size}, and
- FDR threshold for estimated proportion of falsely rejected null hypothesis [13].

The output of the algorithm is a set of potentially interesting subgroups $\mathcal{G}$. For every pair of clustering parameters $(k_1, k_2) \in \{2, \ldots, k_{1Max}\} \times \{2, \ldots, k_{2Max}\}$ **,** the algorithm executes the following steps:

1. *Clustering.* Perform k_1–medoids clustering in X–space and k_2–medoids clustering in Y–space. Label each instance with corresponding cluster index $L_x \in \{1_x, \ldots, k_1\}^n$ and $L_y \in \{1_y, \ldots, k_2\}^n$.
2. *Inference of relationship between variables L_x and L_y.* Present the results of clustering in a contingency table for L_x and L_y:

		L_y				
		1_y	2_y	$\cdots$	k_2	
L_x	1_x	n_{11}	n_{12}	$\cdots$	n_{1k_2}	n_{1+}
	2_x	n_{21}	n_{22}	$\cdots$	n_{2k_2}	n_{2+}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	k_1	$n_{k_1 1}$	n_{k_1}	$\cdots$	$n_{k_1 k_2}$	n_{k_1+}
		n_{+1}	n_{+2}	$\cdots$	n_{+k_2}	n

   where $n_{ij} = |\{e : L_x(e) = i, L_y(e) = j\}|$. Let marginal frequencies be $n_{i+} = \sum_{j=1}^{k_2} n_{ij}$ and $n_{+j} = \sum_{i=1}^{k_1} n_{ij}$. Every cell in contingency table represents a subgroup of instances. Only cells with sufficient number of instances ($n_{ij} > min_{size}$) are considered for subsequent analysis.
3. *Significance estimation.* We form the following null hypothesis:

$$H_0 : \textit{variables } L_x \textit{ and } L_y \textit{ are independent}$$

Under the null hypothesis each cell approximately follows χ^2 distribution with one degree of freedom. For each cell we compute χ^2 statistics c_{ij}:

$$c_{ij} = \frac{(n_{ij} - \frac{n_{i+}n_{+j}}{n})^2}{\frac{n_{i+}n_{+j}}{n}}. \quad (1)$$

and estimate the corresponding p-value $p_{ij} = P(\chi_0^2 > c_{ij})$ where $\chi_0^2 \sim \chi^2(1)$.

The algorithm tests a potentially large number of hypotheses: if every cell would contain more than min_{size} instances, the number of tested hypothesis would be equal to $\big((k_{1Max}-1)(k_{2Max}-1)\big)^2$. Then, it estimates each subgroup's false discovery rate (FDR) using a procedure as proposed by Benjamini [13]. Only the subgroups with FDR below the user-defined threshold are analyzed further. The proposed correction neglects the fact that the clusters themselves are based on the data and is therefore too optimistic. The adjusted p–values are however merely used for ranking of the subgroups and the false discoveries are diminished with post–processing.

2.3 Subgroup Post–Processing

Even with FDR correction, the number of discovered subgroups $|\mathcal{G}|$ can be large. Discovered subgroups may share a significant number of cases, as the clusters we use are coming from a set of k-medoid clusterings with varying k and as such necessarily overlap. To minimize the number of resulting subgroups, we measure their similarity and report only the most representative ones. Our post–processing first measures the similarity between each pair G_i, G_j of subgroups with Jaccard coefficient (2):

$$J(G_i, G_j) = \frac{|G_i \cap G_j|}{|G_i \cup G_j|}. \quad (2)$$

We then create a network where each node represents a discovered subgroup; two nodes in the network representing G_i and G_j are connected if $J(G_i, G_j) > J_t$. Based on visual inspection of the resulting networks, we have used a threshold value of $J_t = \frac{1}{2}$ in our studies. The subgroup with the largest value of χ^2 statistics is chosen as a representative subgroup of each connected component, with the remaining subgroups in the component removed for further analysis. The original set of subgroups $\mathcal{G}$ is in this way reduced to a set $\mathcal{G}'$.

2.4 Subgroup Description

For expert's interpretation of the set $\mathcal{G}'$ it is necessary to obtain a suitable description. The subgroups are presented to the users (domain experts) through its most representative member (medoid) by listing only the most informative

features, *i.e.*, features that best distinguish subgroup and non-subgroup instances. Feature scores are computed by ReliefF [14], which unlike popular univariate measures (*e.g.* information gain, gini index, ...) also considers interactions between other features. To make the ReliefF measures more understandable for domain experts, we computed the significance of the resulting ReliefF scores by approximating their null distribution through permuting the subgroup membership labels and observing ReliefF scores on permuted data.

3 The Data

Our study considers a data set of 1,267 patients with femoral neck fractures admitted and operatively treated at the Department of Traumatology of University Clinical Centre in Ljubljana that was collected from December 1986 to December 2000. For each patient, a study recorded a set of features at the time of operation and immediately after the operation. The long-term clinical status was assessed through a follow-up on average in 435 days after the operation. According to their clinical use and meaning the features were divided into six feature sets (Table 1). This division was introduced by physicians that collected the data, that is, was defined prior to the study.

Table 1. Feature sets, the number of features included (n_{var}) and a list of names for a selection of representative features from the set

feature set	n_{var}	representative features
basic	6	sex, age, primary treatment, ...
health status	8	diseases: cardiovascular, dementia, diabetes, ...
complications	17	plate protrusion, fracture of diaphysis, death, ...
postoperative state	5	sitting, standing up, walking, ...
general health after operation	7	mobility, pain, general state, ...
cause of secondary treatment	9	femoral head necrosis, infection, broken screw, ...

In total, the data set included 58 features, of which 54 were nominal (discrete) and 4 were continuous (real-valued). The data set included a number of missing values, mainly due to semantics. For instance, if there was no secondary treatment, the values of features describing its cause were missing.

Our subgroup discovery algorithm requires a data set which consists of input (X) and output feature set (Y). Using feature sets from Table 1, we considered 11 clinically meaningful models (see the following section for the list) where all the features from any feature set were all incorporated either in X or Y. For example, a data set was analyzed where $X = \{x_i : x_i \in \text{basic} \cup \text{health status}\}$ and $Y = \{y_i : y_i \in \text{complications}\}$. In this model we used $6 + 8 = 14$ input and 17 output features.

4 Experiments and Results

Due to a large number of missing values (due to semantics, as we reported above), the entire data set is not appropriate for the study of all 11 models. Depending on the model different suitable subsets of instances were selected for the analysis. We have set the parameters $min_{size}, k_{1Max}, k_{2Max}$ according to the size of the working data n'. Following parameters of our proposed method were used:

- $min_{size} = \lceil 0.05 \cdot n' \rceil$: a subgroup must contain at least 5% of instances from the working data set,
- $k_{1Max}, k_{2Max} = \lceil \sqrt{0.05 \cdot n'} \rceil + 1$: the expected number of instances in each cell should be larger than 5%,
- $FDR = 0.1$: we tolerate 10% of falsely rejected hypotheses.

Among 11 examined models only five of them resulted in a description with at least one interesting discovered subgroup (Table 2).

Table 2. Overview of results: n' is the size of the working data set, and $|\mathcal{G}|$ and $|\mathcal{G}'|$ the number of discovered subgroups before and after post–processing

input set (X)	output set (Y)	n'	$\lvert\mathcal{G}\rvert$	$\lvert\mathcal{G}'\rvert$
basic, health status, complications	postoperative state	328	53	10
basic, health status, complications (yes/no)	postoperative state	1267	345	23
basic, health status	postoperative state	1267	306	28
basic, health status, complications	cause of secondary treatment	138	9	3
basic, health status, complications (yes/no)	general health after operation	1267	431	42
postoperative state				
basic, health status	complications	1267	0	0
basic, health status	postoperative state	138	0	0
cause of secondary treatment				
basic, health status, complications	general health after operation	138	0	0
cause of secondary treatment				
basic, health status, complications	general health after operation	138	0	0
postoperative state				
basic, health status, complications (yes/no)	cause of secondary treatment	138	0	0
basic, health status, complications (yes/no)	general health after operation	1267	0	0
postoperative				

The domain expert (physician) commented on the relevance of the rules to clinical practice and their usefulness in forming hypotheses further research. Subgroups were presented with increasing p-values. The comments were purely qualitative at this stage. We report six subgroups that were selected by the expert as the most interesting. A brief description (name, size) with medoid representation and expert's explanation is listed. All the presented subgroups have their associated p-values less than 0.002, and are described with the medoid in attribute space (X) in the left and medoid in classes space (Y) in right column using the features with ReliefF p-values below 0.01.

1. a subgroup of 95 patients

   hearth disease = yes
   plate protrusion = no
   peripheral neurological deficit = no
   cardiovascular complications = no
   death = after 3 day
   pulmonary embolism = no
   age = 93.3
   patient collaborates = no

   A high-risk group (elderly patients with hearth disease), patients did not collaborate, the decision to operate on such patients is questionable (death after the operation).

2. a subgroup of 30 patients

   hearth disease = no
   peripheral neurological deficit = no
   aseptic nacrosis = no
   sitting = perfect
   standing up = perfect
   patient collaborates = yes

   Patients without major complications, rehabilitation is successful and fast, patients do cooperate.

3. a subgroup of 283 patients

   age = 81
   hearth disease = yes
   sex = female
   sitting = good
   standing up = belated
   patient collaborates = no

   Elderly patients with major complications, engaged in earlier testing (sitting) but then had problems with standing and later failed to collaborate.

4. a subgroup of 470 patients

   primary procedure plating ostheosynthesis = plate with 1 screw + 1 extra screw
   hearth disease = no
   malign diseases = no
   dementia = no
   hearth disease = no
   sitting = perfect
   patient collaborates = yes

   This procedure is the prevailing type of the operation used, and shows the success for the patients with no major complications.

5. a subgroup of 348 patients

   hearth disease = yes
   dementia = no
   patient collaborates = yes

   A hypothesis here is that even with some complications (hearth disease), patient without dementia would typically collaborate in the treatment procedure.

6. a subgroup of 143 patients

hearth disease = no	patient collaborates = yes
dementia = no	mobility = walking without support
complications = no	fracture healed = yes
patient collaborates = yes	general condition = good
age = 62	general condition belated = no

Healthy younger patient, the outcome of treatment is good.

5 Discussion and Conclusions

Computational evaluation of the results of explorative data analysis is hard, especially for researchers that are used to well-defined scoring techniques available for supervised data mining. While we crafted our subgroup discovery algorithm and scoring of subgroups to expose only the subgroups of statistical significance, and corrected the results for multiple testing, we are aware that this does not form a computational test of robustness of the resulting patterns. The proposed method does not aim at construction of predictive models; it should instead be used to ease data exploration. And in this we found our approach successful.

The data we have considered comes from the major country's clinical centre, where Traumatology department performs majority of state-wide femoral neck fracture operations. The data set we have used records 14 years of operative treatments by the same surgeon (DS), and can of course be regarded as a major resource for assessment of factors that do and do not influence the outcome of the treatment. The subgroups we have discovered and that were presented to participating domain experts (DS, GM) pointed to some obvious patterns (*e.g.*, better clinical outcomes for younger patients and worse for elderly), but also uncovered some less obvious factors that impact the outcome, like exposed influence of collaboration of the patient, or effects of accompanying diseases, despite being unrelated to the injury in general.

For data miners, the success of explorative data analysis lies in the engagement of the user in terms of *enlightenment* during interpretation of the results and in the ease by which the results of the analysis are understood. Since we are not able to quantify these, it should be clear that the particular technique we have proposed can foster both: the subgroups we have found are, at least for this non-trivial domain, few enough and easy to interpret. The choice of medoids for subgroup representation was intentional and proved itself well: clinicians most often think of specific cases rather than of population descriptors. Feature scoring and filtering (ReliefF accompanied with p-value assignment) proved beneficial, too. Our early attempts that did not include the filtering let to page-long descriptions of subgroups that were at best confusing to physicians, if not misleading due to too much irrelevant information.

By and large, the subgroups we have discovered provide for a compact summary across the entire data set, pointing out patterns that should be considered

by the domain experts. In our case, they were either the source of confirmation of the present state of knowledge, or outlined new hypotheses that have yet to be tested in future studies.

References

[1] Hand, D.J.: Handbook of data mining and knowledge discovery. Oxford University Press, Inc., New York (2002)
[2] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The kdd process for extracting useful knowledge from volumes of data. Commun. ACM 39(11), 27–34 (1996)
[3] Lavrač, N., Flach, P., Kavšek, B., Todorovski, L.: Adapting classification rule induction to subgroup discovery. In: Proceedings of IEEE International Conference on Data Mining, pp. 266–273 (2002)
[4] Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research 5, 153–188 (2004)
[5] Kavšek, B., Lavrač, N., Jovanoski, V.: APRIORI-SD: Adapting association rule learning to subgroup discovery. In: R. Berthold, M., Lenz, H.-J., Bradley, E., Kruse, R., Borgelt, C. (eds.) IDA 2003. LNCS, vol. 2810, pp. 230–241. Springer, Heidelberg (2003)
[6] Kavšek, B., Lavrač, N.: APRIORI-SD: Adapting association rule learning to subgroup discovery. Applied Artificial Intelligence 20(7), 543–583 (2006)
[7] Blockeel, H., De Raedt, L., Ramon, J.: Top-down induction of clustering trees. In: Proceedings of the 15th International Conference on Machine Learning, pp. 55–63. Morgan Kaufmann, San Francisco (1998)
[8] Ženko, B., Struyf, J.: Learning predictive clustering rules. In: Bonchi, F., Boulicaut, J.-F. (eds.) KDID 2005. LNCS, vol. 3933, pp. 234–250. Springer, Heidelberg (2006)
[9] Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley, Chichester (1990)
[10] Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a dataset via the gap statistic (2000)
[11] Rousseeuw, P.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20(1), 53–65 (1987)
[12] Irigoien, I., Arenas, C.: INCA: new statistic for estimating the number of clusters and identifying atypical units. Statistics in Medicine 27(15), 2948–2973 (2008)
[13] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57(1), 289–300 (1995)
[14] Kononenko, I.: Estimating attributes: Analysis and extensions of relief. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994)

A Framework for Multi-class Learning in Micro-array Data Analysis

Nicoletta Dessì and Barbara Pes

Università degli Studi di Cagliari, Dipartimento di Matematica e Informatica,
Via Ospedale 72, 09124 Cagliari, Italy
{dessi,pes}@unica.it

Abstract. A large pool of techniques have already been developed for analyzing micro-array datasets but less attention has been paid on multi-class classification problems. In this context, selecting features and quantify classifiers may be hard since only few training examples are available in each single class. This paper demonstrates a framework for multi-class learning that considers learning a classifier within each class independently and grouping all relevant features in a single dataset. Next step, that dataset is presented as input to a classification algorithm that learns a global classifier across the classes. We analyze two micro-array datasets using the proposed framework. Results demonstrate that our approach is capable of identifying a small number of influential genes within each class while the global classifier across the classes performs better than existing multi-class learning methods.

Keywords: Micro-array data analysis, Multi-class learning, Feature Selection.

1 Introduction

During the last decade, the advent of micro-array technology has enabled biologists to quantify the expression levels of thousands of genes in a single experiment. Analysis of micro-array data presents unprecedented opportunities and challenges for data mining in areas such as gene clustering, sample clustering and class discovery, sample classification [1].

A micro-array dataset is provided as a training set of labeled records, the task being to learn a classifier that is expected not only to produce the correct label on the training data but to predict correctly the label (i.e. diseases or phenotypes) of novel unlabeled samples. Since a typical micro-array dataset may contain thousands of genes but only a small number of samples (often less than 200), the obvious need for dimension reduction techniques was realized [2]. A number of studies have shown that selecting a small number of discriminative genes is essential for successful sample classification. In addition to reducing noise and improving the accuracy of classification, the gene selection process may have important biological interpretation. Genetic markers, in fact, can be useful in further investigation of the disease and in future therapies.

Classification of micro-array data has been extensively studied over the past few years and most research efforts focused on datasets involving only two classes (binary

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 275–284, 2009.

classification problems) [3][4]. Less attention has been paid on the classification case [5] where there are at least three class labels (multi-class classification problems). In this last context, classification techniques can be organized into two categories depending on how they build the classifier.

One category extends binary classification algorithms to handle multi-class problems [6]. These techniques focus on learning a single classifier that discriminates several classes simultaneously and can be advantageous, but this comes at a certain price since the problem dimensions introduce a computational complexity that often makes the problem intractable. Feature selection can help to reduce that complexity by selecting and keeping as input a subset of the original genes, while irrelevant and redundant features are removed. Selecting features and quantify classifiers may be hard since only few training examples may be available in only single class. As well, difficulties arise when the micro-array dataset exhibits one class (i.e. a tumor type) that is much more clearly characterized than others. In this situation, traditional feature selection methods result in a gene set that might be most representative of only one (or some, not all) class value.

The other category groups algorithms that decompose the original multi-class problem into binary ones [7][8]. These approaches focus on learning multiple "related" binary classifiers separately and are proven quite successful in practice; however, theoretical justification for this success has remained elusive. The goal of these algorithms is to estimate separate predictive models for several classes. They assume predictive models to be sufficiently different that learning a specific model for each class results in improved performance, but neglect that models can share some common underlying representation that should make a single classifier beneficial.

In this paper we present a framework for multi-class learning that is a natural extensions of the above algorithms and aims at overcoming their limits. The proposed framework focuses on the scenario where specialized binary classifiers select optimal sets of relevant features that are grouped to learn a single classifier in a large common space. Instead of choosing a particular feature selection method and accepting its outcome as the final subset, we consider to build a specific classifier for each class with the aim of selecting an optimal subset of features. Then, we build a common classifier across the classes that benefits from that optimal subset resulting from learning a classifier for each class independently.

Involved methods are organized in a single framework that carries on data classification through two steps. The first step consists of independently learning classifiers within each class. Specifically, we first consider breaking the original multi-target problem into a set of binary sub-problems (one for each class). To this end, the original dataset is decomposed into a set of binary datasets each separating the instances of a given class from the rest of the classes. A feature selection algorithm is then applied to discriminate the genes most strongly correlated to each class. The second step consists of grouping all relevant features in a single dataset that is presented as input to a classification algorithm that learns a global classifier across the classes.

We report experiments which demonstrate that the proposed framework learns a common classifier across the classes while also improving the performance relative to learning each class independently. Moreover, the experiments show that the proposed method performs better than existing multi-class learning methods.

The paper is organized as follows. Section 2 summarizes some related works. Section 3 illustrates the proposed framework, whose validation is asserted by experiments presented in Section 4. Finally, conclusions as well as future works are outlined in Section 5.

2 Related Work

There are two main steps in the classification of tumors using gene expression data: *feature (gene) selection* and *classification*. The first step aims at identifying a subset of predictive genes that characterize the different tumor classes. Once this subset is constructed by gene selection, the second step is to identify samples into known classes using predictive genes as properties of those samples.

The problem of feature selection received a thorough treatment in machine learning and pattern recognition. Most of the feature selection algorithms approach the task as a search problem, where each state in the search specifies a distinct subset of the possible features [9]. The search problem is combined with a criterion in order to evaluate the merit of each candidate feature subset. There are a lot of possible combinations between each search procedure and each feature evaluation measure [10].

Based on the evaluation measure, feature selection algorithms can broadly fall into the *filter* model and the *wrapper* model [11]. The filter model relies on general characteristics of the training data to select predictive features (i.e. features highly correlated to the target class) without involving any mining algorithm. Conversely, the wrapper model uses the predictive accuracy of a predetermined mining algorithm to state the goodness of a selected feature subset, generally producing features better suited to the classification task at hand. However, it is computationally expensive for high-dimensional data [9][11]. As a consequence, in gene selection the filter model is often preferred due to its computational efficiency. Hybrid and more sophisticated feature selection techniques are explored by recent micro-array research efforts [2].

As regards the classification task, different learning algorithms have been explored for cancer sub-type discrimination and outcome prediction. However, it is not clear from the literature which classifier, if any, performs best among the many available alternatives [12]. Indeed, the success of a particular classifier depends on the peculiarities of the data, the goal of the practitioner, and very strongly on the sample size. It is also currently poorly understood what are the best combinations of classification and gene selection algorithms across most micro-array datasets. Selecting simple classifiers that need minimal parameter tuning seems to be the appropriate approach, independently of data complexity and especially for small sample sizes [12].

Further difficulties arise when more than two targets (i.e. cancer sub-types) are involved in the classification problem, with the accuracy rapidly degrading as the number of classes increases [5]. To address this issue, multi-target problems are usually approached by training and combining a family of binary classifiers each of ones is devoted to a specific sub-classification problem. Majority voting or winner-takes-all is then applied to combine the outputs of binary classifiers, but it often causes problems to consider tie-breaks and tune the weights of individual classifiers. There have been many studies of aggregating binary classifiers to construct a multi-target classifier based on one-versus-the-rest (1-r), one-versus-one (1-1), or other coding strategies [13]. However, the studies found that the best coding depends on each situation.

All the above crucial issues still remain an active focus of micro-array research, along with a range of other questions compactly summarized in [14]. As a major concern, the "curse of dataset sparsity" can render any classification result statistically suspect, and imposes to carefully investigate the biological relevance of selected genes.

3 The Proposed Framework

As previous mentioned, multi-classification problems are intrinsically more complex than binary ones and proper learning strategies and heuristics must be devised to successfully address them. Our proposal is depicted in Fig. 1 that illustrates a framework involving two separated learning levels.

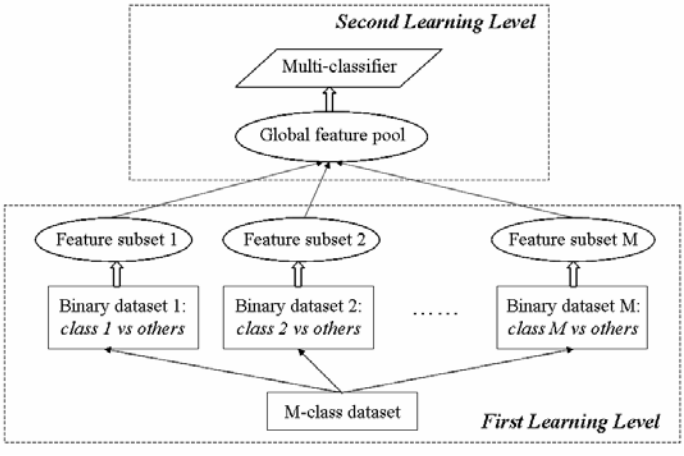


Fig. 1. Architecture of the proposed framework

At first level, the framework considers breaking the original M-class problem into a set of binary sub-problems (one for each class). Towards this end, the original M-class dataset is decomposed into M binary datasets, each of ones separates the instances of a given class from the rest of the classes and provides a clear genetic characterization of each cancer sub-type. As well, this decomposition focuses on the correlation of each gene with only one class at time and allows a separate feature selection to be performed for each binary sub-problem in order to identify the most relevant features for each cancer sub-type.

However, the notion of feature relevance has not yet been rigorously defined in common agreement [15]. Moreover, there is no guarantee that just because a feature is relevant, it will necessarily be useful to a classification algorithm (or vice-versa). Our concept of relevance is based on the definition given by [16]:

Given a sample of data D, a learning algorithm L, and a feature subset F, the feature x_i is incrementally useful to L with respect to F if the accuracy of the hypothesis that L produces using the group of features $\{x_i\} \cup F$ is better than the accuracy achieved using just the subset of features F.

The above definition suggests to select a subset of relevant features according to the classification performance while building that subset in incremental way. Following this approach, our basic idea is to select features step by step from an ordered list of ranked features. Specifically, the first level of the proposed framework considers to apply a scoring function computed from the values of each gene and the class label (on the basis of some statistical or entropic criterion). An high score is indicative of a valuable gene. Then, the learning process sorts genes in decreasing order of this score and carries on a sequential forward selection that starts with a subset that contains the first gene of the ordered list. The evaluation of this subset is obtained by training and testing a binary classifier. Following the ordered list, the genes are added one by one to the subset while such inclusion improves the classifier accuracy.

Being independent from both the classification algorithm and the ranking criterion, this incremental approach is modular in nature and can be implemented using different classification and ranking methods. Its feasibility has been already proved in a number of previous studies [17][18]. In this work, we further validate this approach in the context of a more general framework.

The first learning level of the framework takes advantage from binary decomposition for identifying the genes that are most strongly correlated to each class value: the incremental gene selection is performed separately for each cancer sub-type (class value), in order to determine which genes are responsible of each single pathology. We observe that the number of discriminative genes can be different for different cancer sub-types.

The second learning level considers to join all the selected gene subsets in a single global feature pool which is presented as input to a classification algorithm. The classification task is not performed on binary sub-problems, but a single classifier is learnt that discriminates several classes simultaneously while reducing problem dimensions and computational complexity. This approach overcomes the need of combining the outcome of binary classifiers (often resulting in a doubtful global class assignment) and is not affected by the intrinsic weakness of traditional multi-target classifiers which don't benefit of a sub-type specific gene selection.

4 Experimental Analysis

The Acute Lymphoblastic Leukaemia (ALL) dataset and the Lung Cancer (LC) dataset have been used as a test-bed for the experimental study presented in this section. The ALL dataset [19] consists of 327 samples (215 for training and 112 for test), each one described by the expression level of 12558 genes. 7 classes are involved in total, i.e. all known ALL sub-types (T-ALL, E2A-PBX1, TEL-AML1, BCR-ABL, MLL, Hyperdip > 50) and a generic class OTHERS, that groups all samples not belonging to any of the previous sub-types. The LC dataset [20] consists of 203 samples (136 for training and 67 for test). 5 classes are involved in total, i.e. ADEN, COID, SQUA, SCLC and NORMAL. Each sample is described by the expression level of 12600 genes. Details on class distributions are provided in Table 1.

To validate the proposed approach, we applied the χ^2 statistics as a ranking criterion and three well known classification algorithms: Naïve Bayes (NB), k-Nearest Neighbor (k-NN) and Support Vector Machines (SVM). We are not interested in

Table 1. Class distributions

Dataset	*Training records*	*Test records*
ALL	T-ALL: 28, E2A-PBX1: 18, TEL-AML1: 52, BCR-ABL: 9, MLL: 14, Hyperdip > 50: 42, OTHERS: 52	T-ALL: 15, E2A-PBX1: 9, TEL-AML1: 27, BCR-ABL: 6, MLL: 6, Hyperdip > 50: 22, OTHERS: 27
LC	ADEN: 94, COID: 13, SQUA: 14, SCLC: 4, NORMAL: 11	ADEN: 45, COID: 7, SQUA: 7, SCLC: 2, NORMAL: 6

obtaining the best performance of one special method on a given dataset; instead, we are interested in demonstrating the effectiveness of the proposed framework. All the experiments have been carried out using the Weka machine learning software package [21].

According to the proposed framework, the original training set was decomposed into M binary sub-datasets, where M is the number of cancer sub-types, by separating the instances of each single sub-type from the other classes (whose instances were considered negative examples). As a next step, the ranking procedure was applied separately to each sub-dataset in order to evaluate, for every single feature, the strength of its correlation to a given cancer sub-type. This resulted in M lists of genes ordered by their rank within a specific class. Each ranked list was then provided as input to the incremental feature selection algorithm, in order to identify the subset of genes capable of best discriminating a given cancer sub-type.

Table 2 and 3 show, for ALL and LC dataset respectively, the cardinality of the selected subsets, as well as the predictive accuracy of these subsets when used to train NB, k-NN and SVM binary classifiers specialized on a specific cancer sub-type. As

Table 2. Cardinality of the gene subset selected for each ALL sub-type and accuracy of NB, k-NN and SVM binary classifiers built on this subset (in brackets)

	T-ALL	*E2A-PBX1*	*TEL-AML1*	*BCR-ABL*	*MLL*	*Hyperdip>50*
NB	1 *(100 %)*	1 *(100 %)*	14 *(100 %)*	10 *(97,3 %)*	16 *(99,1 %)*	12 *(96,4 %)*
k-NN	1 *(100 %)*	1 *(100 %)*	12 *(100 %)*	3 *(97,3%)*	3 *(100 %)*	9 *(95,5 %)*
SVM	1 *(100 %)*	1 *(100 %)*	12 *(100 %)*	9 *(99,1 %)*	3 *(100 %)*	5 *(96,4 %)*

Table 3. Cardinality of the gene subset selected for each LC sub-type (including "normal") and accuracy of NB, k-NN and SVM binary classifiers built on this subset (in brackets)

	ADEN	*COID*	*NORMAL*	*SCLC*	*SQUA*
NB	9 (89,6 %)	3 (100 %)	2 (95,5 %)	1 (100 %)	1 (97,0 %)
k-NN	4 (92,5 %)	1 (100 %)	1 (98,5 %)	4 (100 %)	2 (97,0 %)
SVM	20 (91,0 %)	1 (100 %)	12 (98,5 %)	4 (100 %)	2 (97,0 %)

we can see, the number of selected genes is different for different classifiers, since the feature selection is performed according to a wrapper approach which is tuned to the specific classification algorithm.

Results in Table 2 and 3 show that certain cancer sub-types can be perfectly discriminated using very few genes. As regards ALL, in particular, only one gene is sufficient to perfectly classify T-ALL and E2A-PBX1 sub-types, irrespective of the adopted learning algorithm. No error occurs in TEL-AML1 classification too, even if a higher number of features (12-14) is required. Most misclassifications occur in discriminating BCR-ABL and Hyperdyp>50, suggesting a less sharp genetic characterization of these sub-types. Analogously, some LC sub-types (COID, SCLC) can be perfectly classified with very few genes, while other sub-types (ADEN) exhibit a less sharp genetic characterization.

By implementing the second learning level, the M selected gene subsets, one for each cancer sub-type, were joined into a "global" subset involving all (and only) the genetic information which is really necessary to handle the multi-classification problem. That is, the resulting subset was used to train a single classifier capable of discriminating between multiple class values, in that providing a global description of the underlying biological domain. Results obtained on ALL and LC datasets are summarized in Table 4, in terms of both feature cardinality and classification accuracy.

Table 4. Cardinality of the "global" feature subset and classification accuracy of the resulting NB, k-NN and SVM multi-target classifiers

	ALL dataset		***LC dataset***	
	Num. of features	***Accuracy***	***Num. of features***	***Accuracy***
NB	54	86,6 %	16	92,5 %
k-NN	29	93,8 %	12	98,5 %
SVM	31	96,4 %	39	97,0 %

To demonstrate the effectiveness of the proposed approach, we compared the performance of the above multi-classifiers with the one of traditional multi-classifiers that extend binary classification algorithms to handle multi-class problems by simultaneously measuring the correlation of each gene with all class values. We observe that this traditional classification process makes it impossible to determine which genes are responsible of a single pathology (class value) since it doesn't benefit of a sub-type specific gene selection. Indeed, the above standard approach to multi-classification is especially unsuitable when one class is much more clearly characterized than others and it may require many features to be included in the final subset in order to adequately represent all class values (and not just the ones with the strongest characterization) [17][18].

Specifically, we applied a ranking procedure on the original multi-target datasets (both ALL and LC) by measuring the correlation of each gene with the multi-value class label. Then, we studied the performance of NB, k-NN and SVM multi-target classifiers as a function of the number of top-ranked features used to train them. This

analysis showed that, irrespective of the adopted learning algorithm, a large number of genes (in the order of hundreds) is necessary to reach classification accuracies comparable with results in Table 4, while the accuracy measured on small gene subsets is quite poor, in agreement with findings of recent literature [5][22].

This traditional approach to multi-classification is significantly outperformed by the learning framework proposed in this paper, as clearly emerges when the resulting classification performances are compared using the same number of features. As far as ALL, the accuracy of a standard NB multi-classifier is only 55,4% with 50 features, while our approach achieves 86,6% with 54 features, as reported in Table 4. Analogously, the performance of k-NN and SVM multi-classifiers greatly improves when the proposed framework is adopted. Specifically, k-NN achieves 82,1% with 30 features selected in the standard way, while its accuracy is 93,8% when a global set of 29 features is selected according to the proposed approach. Similarly, the accuracy of a standard SVM multi-classifier is 80,4% with 30 features, while our learning framework achieves 96,4% with 31 features.

The experiments on the LC dataset confirm the effectiveness of the proposed approach in selecting a small-size set of predictive genes. Specifically, as Table 4 shows, 16 genes are selected for NB (92,5% of accuracy), 12 for k-NN (98,5% of accuracy), and 39 for SVM (97,0% of accuracy), while standard NB, k-NN and SVM multi-classifiers trained on a comparable number of features achieve only 71,6%, 79,1% and 92,5% respectively. A more explicit comparison is provided by Table 5 which shows the confusion matrix of a k-NN multi-classifier built in the standard way (left panel) and according to the proposed framework (right panel).

Table 5. Confusion matrix of a k-NN multi-classifier built in the standard way (left panel) and according to the proposed framework (right panel). Results refer to the LC dataset.

actual class	predicted class					
	a	**b**	**c**	**d**	**e**
a	38	5	0	2	0
b	2	4	0	0	1
c	0	0	6	1	0
d	0	1	0	1	0
e	2	0	0	0	4

10 features, accuracy ***= 79,1 %***

actual class	predicted class					
	a	**b**	**c**	**d**	**e**
a	44	0	0	0	1
b	0	7	0	0	0
c	0	0	7	0	0
d	0	0	0	2	0
e	0	0	0	0	6

12 features, accuracy = ***98,5%***

a = ADEN, **b** = SQUA, **c** = COID, **d** = SCLC, **e** = NORMAL

The above analysis on both ALL and LC datasets demonstrates the effectiveness of the proposed framework. In terms of accuracy, our results are comparable to the ones provided by other approaches in recent literature [23][24][25], but the number of genes we selected is lower. In particular, [23] explores different heuristics to rank genes within each ALL sub-type; the 20 top-ranked genes are then selected for each sub-type, being 20 an arbitrary threshold which lacks a theoretical or experimental

justification. As regards LC, a t-test is applied in [24] in order to measure the correlation of each gene with the multi-value class label, and an arbitrary threshold of 50 is used to cut-off top ranked features. Both ALL and LC datasets are analyzed in [25] where a variant of RFE (recursive feature elimination) is employed to search for optimal gene subsets, requiring at least 40 genes to reach best classification performance on ALL dataset and at least 100 genes on LC dataset.

5 Concluding Remarks

The proposed framework is a mechanism to integrate the information coming from multiple independent binary classifiers that are similar to specialized local experts. Specifically, each binary classifier is employed just for feature selection, i.e. for achieving knowledge about genes featuring each single class. The multi-classification task is instead performed by a single classifier learnt on a "globally optimal" subset of features that results by joining all the subsets selected by the binary classifiers. Because only the genes really relevant to each pathology are involved in the multi-classification task, this approach allows overcoming the intrinsic weakness of existing multi-learning methods which don't benefit of a class specific feature selection.

[26] is the only work where a similar framework is adopted. However, in [26] the global subset of features is built by taking an equal number of genes for each class, hence requiring more genes than in our approach. Indeed, as witnessed by our analysis, the optimal number of features can be different for different class values, suggesting a different strength in genetic characterization of cancer sub-types. Our approach is tailored to capture genetic sub-type specificity, enabling to sensibly reduce the total number of features involved in multi-classification and returning only the features that are really relevant. As future extension, we plan to further validate our framework on different multi-cancer datasets, to obtain more insights on genetic mechanisms underlying cancer sub-type differentiation.

References

1. Piatetsky-Shapiro, G., Tamayo, P.: Microarray Data Mining: Facing the Challenges. ACM SIGKDD Explorations 5(2) (2003)
2. Saeys, Y., Inza, I., Larranaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)
3. Golub, T.R., et al.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
4. Guyon, I., Weston, J., Barnill, S.: Gene Selection for Cancer Classification Using Support Vector Machines. Machine Learning 46, 389–422 (2002)
5. Li, T., Zhang, C., Ogihara, M.: A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20(15), 2429–2437 (2004)
6. Hastie, T., Tibshirani, R., Friedman, J.: The elements of Statistical Learning: Data Mining, Inference, Prediction. Springer, Heidelberg (2001)
7. Weston, J., Watkins, C.: Multi-class support vector machines. Technical Report, Department of Computer Science, Holloway, University of London, Egham, UK (1998)

8. Lee, Y., Lee, C.K.: Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics 19, 1132–1139 (2003)
9. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. Artif. Intell. 97(1–2), 245–271 (1997)
10. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. 17(3), 1–12 (2005)
11. Kohavi, R., John, G.: Wrappers for feature subset selection. Artificial Intelligence 97(1-2), 273–324 (1997)
12. Pranckeviciene, E., Somorjai, R.: On Classification Models of Gene Expression Microarrays: The Simpler the Better. International Joint Conference on Neural Networks (2006)
13. Yukinawa, N., et al.: Optimal aggregation of binary classifiers for multi-class cancer diagnosis using gene expression profiles. IEEE/ACM Transactions on Computational Biology and Bioinformatics (preprint) (2008)
14. Simon, H.: Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). SIGKDD Explorations 5(2), 31–36 (2003)
15. Bell, D., Wang, H.: A formalism for relevance and its application in feature subset selection. Mach. Learning 41(2), 175–195 (2000)
16. Caruana, R., Freitag, D.: How useful is relevance? In: Working Notes of the AAAI Fall Symposium on Relevance. AAAI Press, N. Orleans (1994)
17. Bosin, A., Dessì, N., Pes, B.: A Cost-Sensitive Approach to Feature Selection in Micro-Array Data Classification. In: Masulli, F., Mitra, S., Pasi, G. (eds.) WILF 2007. LNCS, vol. 4578, pp. 571–579. Springer, Heidelberg (2007)
18. Bosin, A., Dessì, N., Pes, B.: Capturing Heuristics and Intelligent Methods for Improving Micro-array Data Classification. In: Yin, H., Tino, P., Corchado, E., Byrne, W., Yao, X. (eds.) IDEAL 2007. LNCS, vol. 4881, pp. 790–799. Springer, Heidelberg (2007)
19. Yeoh, E.J., et al.: Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1, 133–143 (2002)
20. Bhattacharjee, A., Richards, W.G., et al.: Classification of human lung carcinomas by mrna expression profiling reveals distinct adenoma subclasses. PNAS 98, 13790–13795 (2001)
21. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Elsevier, Amsterdam (2005)
22. Statnikov, A., Aliferis, C.F., et al.: A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21(5) (2005)
23. Liu, H., et al.: A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns. Genome informatics 13, 51–60 (2002)
24. Ling, N.E., Hasan, Y.A.: Classification on microarray data. In: IMT-GT Regional Conference on Mathematics, Statistics and Applications, Malaysia (2006)
25. Ding, Y., Wilkins, D.: Improving the Performance of SVM-RFE to Select Genes in Microarray Data. BMC Bioinformatics 7(suppl. 2), S12 (2006)
26. Piatetsky-Shapiro, G., et al.: Capturing Best Practice for Microarray Gene Expression Data Analysis. In: SIGKDD 2003, Washington, USA (2003)

Mining Safety Signals in Spontaneous Reports Database Using Concept Analysis

Mohamed Rouane-Hacene[1], Yannick Toussaint[2], and Petko Valtchev[1]

[1] Dépt. Informatique, UQÀM, CP 8888, succ. CV, Montréal, Canada, H3C 3P8
[2] LORIA, B.P. 239, F-54506 Vandœuvre-lès-Nancy, France

Abstract. In pharmacovigilance, linking the adverse reactions by patients to drugs they took is a key activity typically based on the analysis of patient reports. Yet generating potentially interesting pairs (drug, reaction) from a record database is a complex task, especially when many drugs are involved. To limit the generation effort, we exploit the frequently occurring patterns in the database and form *association rules* on top of them. Moreover, only rules of minimal premise are considered as output by concept analysis tools, which are then filtered through standard measures for statistical significance. We illustrate the process on a small database of anti-HIV drugs involved in the HAART therapy while larger-scope validation within the database of the French Medicines Agency is also reported.

1 Introduction

Pharmacovigilance (PV) aims at, first, studying and, then, preventing the adverse reactions to drugs (ADR) based on the data collected by spontaneous reporting systems (SRS) and stored in case report databases (DB). SRS DB comprises a collection of reports each capturing the patient characteristics including demographic data (age, race, gender, etc.), the suspected drugs and a description of the observed ADR. Table 1 depicts a set of case reports on AIDS patients and antiretroviral drugs, i.e., treating infection by retroviruses such as HIV.

In PV, drug-reaction combinations, known as *safety signals*, help devising a drug therapy, hence the importance of their detection. For instance, in HIV treatment, the caregivers are interested in the response of various classes of patients to the HAART therapy in order to adapt the overall anti-HIV therapy. Their prime target is an appropriate combination of antiretroviral drugs that, while effective, limits the ADR: e.g., older patients with HIV infection have robust responses to HAART with no increased risk of metabolic disorders or other ADR. Beside safety signals, i.e., (drug, ADR) pairs, further meaningful combinations from the SRS DB involve several drugs for a single ADR. These are potential *drug interactions* (higher-order signals).

Signal detection has been approached with a variety of analysis tools [7] including statistical methods for disproportionality assessment, deviation detection, etc. However, none of these proposes a way, both automated and feasible,

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 285–294, 2009.

for generating all potential signals from the SRS DB. Moreover, even with an expert-provided potential signal, the underlying approaches would consider all drug-reaction combinations that can be derived from the signal, including many spurious ones. For instance, consider the anti-HIV drugs `Lopinavir` and `Tenofovir` in Table 1 and the ADR `HairLoss` and `Oedema`. A proportionate approach would suggest the study of signals (`Lopinavir`, `Oedema`), (`Lopinavir`, `HairLoss`), (`Tenofovir`, `Oedema`), and (`Tenofovir`, `HairLoss`). Yet the only sensible combination to study is ({`Lopinavir`, `Tenofovir`}, {`HairLoss`, `Oedema`}) as, given the dataset, the four combine to a *maximal* pattern. In summary, because of the large size of most SRS DB, the computation of all combinations is strongly combinatorial, hence their test may prove infeasible. Instead, a more careful approach would track the frequently occurring patterns in the records and use these as prototypes.

Table 1. A fragment[1] of SRS DB

Patient	Age	Gender	Prescribed drugs	Observed adverse drug reactions
Daffy	24	`Female`	`Lopinavir`, `Efavirenz`	`Nausea`, `Hives` , `Vomiting`
Farley	63	`Male`	`Lopinavir`, `Tenofovir`	`Oedema`, `Hives`, `Headache`, `Nausea`, `Heart failure`, `Hair loss`
Lane	27	`Female`	`Maraviroc`, `Efavirenz`	`Fatigue`, `Oedema`, `Hives`, `Hair loss`, `Bleeding`
Shana	15	`Female`	`Tenofovir`, `Lopinavir`	`Fatigue`, `Oedema`, `Hair loss`
Trudy	41	`Male`	`Raltegravir`	`Fatigue`, `Breath disorder`, `Nausea`, `Heart failure`, `Bleeding`, `Vomiting`

Patterns comprised of two sets, a premise and a conclusion, called *associations*, have been successfully applied to a variety of practical problems involving co-occurrencies of phenomena and seem to fit well the PV context. Yet a notorious problem of association miners is the huge number of potentially useful associations that may be extracted from even a small DB. Formal concept analysis (FCA) [6] provides the theoretical foundation for association rule bases that only withhold a tiny proportion of all valid associations while keeping the total of the information. Hence we propose an FCA-based method for signal detection which, by examining a minimal set of association rules extracted from the SRS DB helps minimize the number of (drug, ADR) pairs to be statistically analyzed.

Here, we examine the detection of safety signals and drug-drug interactions by means of FCA and a set of disproportionality measures to discard statistically non significant associations. Our approach is illustrated on a set of case reports on AIDS patients and antiretroviral drugs. A validation thereof involving the SRS DB of the French Medicines Agency is also reported.

The paper starts by a short presentation of concept lattices and association rules (Sect. 2). Follows the description of the proposed method (Sect. 3). Sect. 4

[1] Source: MedEffect, Canada vigilance online database.

presents the results of the preliminary experiments. Related work is summarised in Sect. 5 while further research directions are given in Sect. 6.

2 Background on Concept Lattices and Association Rules

2.1 Concept Lattices

Formal concept analysis (FCA)[6] is a method for designing concepts and conceptual hierarchies from collections of individuals (formal objects) described by properties (formal attributes). To apply FCA to PV data as presented in Table 1, the latter must first be encoded in standard format. The format, a binary context $\mathcal{K} = (O, A, I)$, (see Table 2) involves a set of objects O, a set of attributes A and an incidence relation $I \subseteq A \times O$ (oIa stand for "object o has the attribute a"). For instance, in Table 2, objects are patients and attributes demographic informations, drugs or reactions.

Table 2. Binary context encoding AIDS patients with their drugs and ADR

	Demographic data					Adverse reactions									Drugs				
	Young	Adult	Senior	Male	Female	Fatigue	Oedema	BreathDisorder	Hives	Headache	Nausea	HeartFailure	HairLoss	Bleeding	Vomiting	Raltegravir	Lopinavir	Tenofovir	Maraviroc	Efavirenz
Daffy		×			×				×		×				×		×			×
Farley			×	×			×		×	×	×	×	×				×	×		
Lane		×			×	×	×		×				×	×					×	×
Shana	×				×	×	×						×				×	×		
Trudy		×		×		×		×			×	×		×	×	×				

Two derivation operators, both denoted $'$ link objects and attributes [6]. Let $X \subseteq O$, $Y \subseteq A$: $X' = \{a \in A | \forall o \in X, oIa\}$, $Y' = \{o \in O | \forall a \in Y, oIa\}$. For example, following Table 2, {`Daffy`, `Trudy`}' = {`Adult`, `Nausea`, `Vomiting`}. The compound operators $''$ are *closure operators* over 2^O and 2^A, respectively. A set $Y \subseteq A$ is closed if $Y = Y''$ which means the objects sharing Y, i.e., Y', share no other attribute (i.e., from A/Y). A pair of sets corresponding to one-another through $'$ is called a (formal) *concept*: $c = (X, Y) \in \wp(O) \times \wp(A)$ is a concept of $\mathcal{K}$ iff $X' = Y$ and $Y' = X$ (here X and Y are called the *extent* and the *intent* of c, respectively). For instance, ({`Farley`, `Shana`}, {`HairLoss`, `Oedema`, `Lopinavir`, `Tenofovir`}) is a concept (c_6 in Fig. 1).

Furthermore, the set $\mathcal{C}_\mathcal{K}$ of all concepts of the context $\mathcal{K}$ is partially ordered by extent inclusion (intent containment). The structure $\mathcal{L} = \langle \mathcal{C}_\mathcal{K}, \leq_\mathcal{K} \rangle$ is a complete lattice, called the *concept lattice*. Fig. 1 shows the lattice of the context in Table 2, whereby a simplified labeling scheme is used where each object/attribute appears

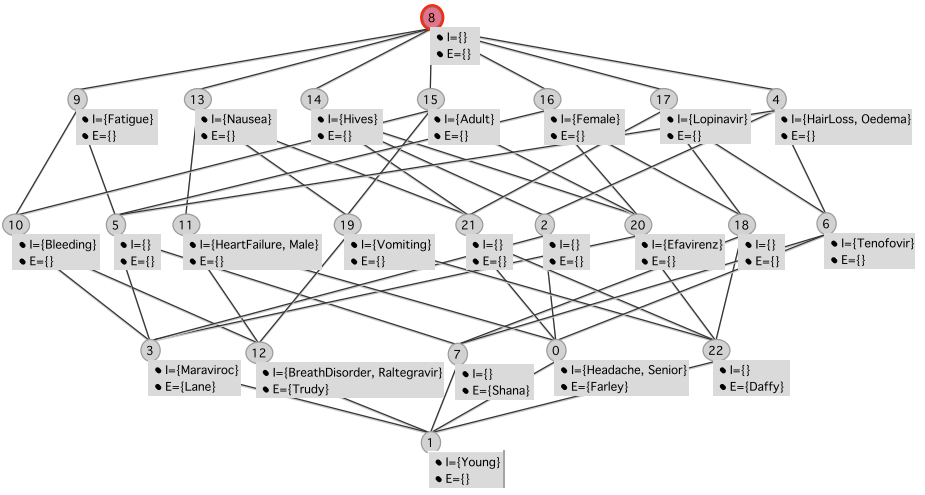


Fig. 1. Concept lattice of case reports given in Table 1

only once in the diagram. The extent of a concept is made of all objects whose labels can be reached from the concept on a downward-heading path while intent is recovered in a dual way. For example, the extent of the concept with the attribute label `Bleeding` is {`Lane`, `Trudy`} while its intent is {`Bleeding`, `Fatigue`}.

Within the lattice $\mathcal{L} = \langle \mathcal{C}_\mathcal{K}, \leq_\mathcal{K} \rangle$, concepts have a unique greatest lower bound termed *meet* ($\bigwedge$) that is defined as follows: $\bigwedge_{i=1}^{k}(X_i, Y_i) = (\bigcap_{i=1}^{k} X_i, (\bigcup_{i=1}^{k} Y_i)'')$ For instance, in Fig. 1, the meet of $c_{\#19}$ = ({`Daffy`, `Trudy`}, {`Adult`, `Vomiting`, `Nausea`}) and concept $c_{\#20}$=({`Daffy`, `Lane`}, {`Adult`, `Female`, `Hives`, `Efavirenz`}) is $c_{\#22}$=({`Daffy`}, {`Adult`, `Female`, `Efavirenz`, `Vomiting`, `Hives`, `Nausea`, `Vomiting`, `Lopinavir`}). In addition, the function $\mu : A \rightarrow \mathcal{C}_\mathcal{K}$ maps an attribute a into the *maximal* concept in the lattice having that attribute ($\mu(a) = (a', a'')$). For instance, in Fig. 1, μ(`HeartFailure`)=$c_{\#11}$.

The lattice in Fig. 1 provides the analyst with a variety of insights into the data such as the profile of the AIDS patients under study, the different anti-HIV treatments and the respective most common ADR. For instance, the concept $c_{\#20}$ = ({`Daffy`, `Lane`}, {`Female`, `Adult`, `Hives`}) represents adult female patients under anti-HIV drug regimen containing NNRTIs[2], including `Efavirenz`, and experiencing `Hives`. In summary, the lattice of case reports provides an overview of drug-reaction combinations to be explored for pharmacological associations detection. In many cases, too specific concepts are not relevant. To only keep those having extents of certain size, the support of a concept is defined as its relative extent size, $\sigma(c) = \frac{\|X\|}{\|O\|}$. The corresponding sub-order of the lattice, i.e., its upper part induced by threshold α in $]0, 1]$, is $\bar{\mathcal{L}}^\alpha = \langle \bar{\mathcal{C}}^\alpha, \leq_\mathcal{K} \rangle$ where $\bar{\mathcal{C}}^\alpha = \{c \mid c \in \mathcal{C}, \sigma(c) \geq \alpha\}$. $\bar{\mathcal{L}}^\alpha$ is called the *iceberg* lattice [10].

[2] Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs) intervene in the early stages of the HIV replication cycle.

2.2 FCA-Based Association Rule Design

FCA framework is widely-used in mining patterns from DB, including association rules that express the co-ocurrences among attribute sets (called *itemsets*). An association rule is a pair of sets '*antecedent* $\longrightarrow$ *consequent*' with no claim of causality. A rule $B \longrightarrow D$ $(B, D \subseteq A)$ has a support $\bar{\sigma}(B \longrightarrow D) = \sigma(B \cup D)$ and a confidence that is the ratio of the rule support to the support of the antecedent $(\bar{\gamma}(B \longrightarrow D) = \frac{\bar{\sigma}(B \longrightarrow D)}{\sigma(B)})$.

Table 3. Drug-reaction associations derived from the SRS data depicted in Table 1 with the corresponding support

	Support
`Tenofovir`$\longrightarrow$`HairLoss, Oedema`	0.4
`Maraviroc`$\longrightarrow$`Bleeding, Fatigue, HairLoss, Hives, Oedema`	0.2
`Efavirenz`$\longrightarrow$`Hives`	0.4
`Raltegravir`$\longrightarrow$`Bleeding, BreathDisorder, Fatigue, HeartFailure, Nausea,Vomiting`	0.2
`Lopinavir,Efavirenz`$\longrightarrow$`Hives, Nausea, Vomiting`	0.2
...	...

In FCA, mining association rules from a DB consists in: (i) extracting all frequent closed itemsets, i.e., concept intents from the DB, with support above α, (ii) generating all valid association rules, i.e., rules whose confidence exceeds a user-defined minimum threshold. The first step presents a greater challenge as the set of frequent itemsets may grow exponentially with the size of A while the second step is relatively straightforward. Moreover, several FCA-based algorithms [1] generate non-redundant bases of association rules. These bases are minimal with respect to the number of rules whereas the contained rules are informative, i.e., with minimal antecedents and maximal consequents. To extract a tractable number of association rules from PV data, we have used the Informative Generic Basis (IGB) [1] as it has been shown that this type of association rules conveys the maximum of useful knowledge, without information loss. Moreover, our IGB contains exact (versus approximative) associations rules, i.e., rules whose confidence is equal to 1 (as opposed to confidence < 1). Table 3 illustrates some of the drug-reaction associations from the IGB extracted out of data in Table 1.

3 Detecting Safety Signals Using FCA

The outline of our mining method is as follows: First, SRS data is encoded into a binary context, where formal objects represent case reports while formal attributes are either taken drugs or the observed reactions (see Table 2). Then, FCA is used to derive both the lattice and the corresponding IGB. For instance, in the case of anti-HIV drugs `Lopinavir` and `Tenofovir` and the two ADR `HairLoss` and `Oedema`, the method will consider only the pair ({`Lopinavir`, `Tenofovir`},

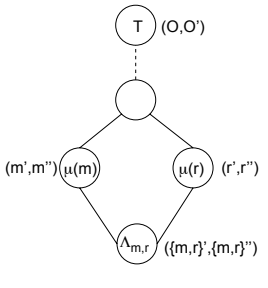
{HairLoss, Oedema}) since it represents the only combination where the four elements appear (concept c_6 in Fig. 1).

Rules of the basis are split into three groups. *Pure* association have both antecedent and consequent made exclusively of drugs and reactions respectively. *Semi-pure* associations, in contrast, admit only non-reaction items in their consequent that are further removed for analysis purposes. Finaly, *biaised* associations admit non-drug items in their antecedents as well. Their components are filtered to fit the *drugs* $\longrightarrow$ *reactions* rule scheme. Later, statistical filters are applied to detect statistically significant candidates for each of the two types of pharmacological associations, i.e., signals and drug-drug interactions.

In order to discard statistically non significant concepts, we use some of the measures of disproportionality [13] that are currently applied in various reporting centers, e.g., the British Medicines and Healthcare products Regulatory Agency (MHRA). Such measures for a suspected ADR of a drug of interest are calculated from the following variables: (a) reports including the drug of interest and the suspected reaction, (b) reports with the drug of interest and no reference to the suspected reaction, (c) reports where the suspected reaction appears without the drug of interest, (d) reports where neither the drug of interest nor the suspected reaction appear. The adopted measures are the proportional ADR reporting ratio (PRR), reporting odds ratio (ROR), and χ^2 test.

For instance, the PRR is the proportion of the suspected ADR versus all ADR reported for the drug of interest devided by the corresponding proportion for other drugs. It can be expressed as PRR $=\frac{a\times(c+d)}{c\times(a+b)}$. Fig. 2 shows how the various cells of the drug-ADR contingency table are calculated using a drug-reaction concept lattice. Hence, every meet concept $\wedge_{m,r}$ in the lattice $\mathcal{L}_P$, for a given pair of a medicine m and a reaction r is the source of a drug-reaction contingency table.

For instance, the calculation of PRR for the anti-HIV drug Lopinavir and the suspected ADR HairLoss using the concept lattice of Fig. 1 is as follows: $a = |Ext((\bigwedge_{\texttt{Lopinavir,HairLoss}})| = |Ext(c_{\#6})| = 2$, $b = |Ext(\mu(\texttt{Lopinavir}))|$-



	Adverse reaction r	$\neg r$	
Drug m	a	b	$a+b$
$\neg m$	c	d	$c+d$
	$a+c$	$b+d$	$\|O\|$

$a = |Ext(\wedge_{m,r})|$
$b = |Ext(\mu(m))| - |Ext(\wedge_{m,r})|$
$c = |Ext(\mu(r))| - |Ext(\wedge_{m,r})|$
$d = |O| - |Ext(\mu(m) \cup Ext(\mu(r))|$

Fig. 2. Left: Drug-concept and ADR-concept in the drug-reaction concept lattice $\mathcal{L}_P$. **Right:** The two-by-two contingency table for target drug m and a suspected ADR r.

$|Ext(\bigwedge_{\texttt{Lopinavir,HairLoss}})|= |Ext(c_{\#17})|-|Ext(c_{\#6})| = 1$, $c=|Ext(\mu(\texttt{HairLoss}))|-|Ext(\bigwedge_{\texttt{Lopinavir,HairLoss}})|=|Ext(c_{\#4})| -|Ext(c_{\#6})|=1$, $d=|O|-(|Ext(\mu(\texttt{Lopinavir})) \cup Ext(\mu(\texttt{HairLoss}))|) = 5-|Ext(c_{\#17}) \cup Ext(c_{\#4})|=1$, PRR $= \frac{2\times(1+1)}{1\times(2+1)}$. The obtained value of PRR is $1.33 \leq 2$. Hence, the assumption stating that `Lopinavir` causes `HairLoss` is statistically non significant.

The detection of higher-order drug-reaction associations, such as drug interactions, has been carried out so far by logistic regression modelling [7] where concomitant drugs (resp. reactions) are considered as covariates and the suspected reaction (resp. drug) as dependent variable. For instance, in our running PV data, the logistic model predicting whether `Nausea` reaction is a result of possible interaction between `Lopinavir` and `Efavirenz` would look like:

$$N = \beta_0 + \beta_1 \times L + \beta_2 \times E + \beta_3 \times L^*E$$

The variables L and E are exposure variables (or predictors) representing risk factors associated with concomitant drugs `Lopinavir` and `Efavirenz`, respectively, while L^*E is the interaction term. The intercept β_0 represents the value of the dependent variable `Nausea` (N) in a patient with no risk factors, while logistic (or logit) coefficients β_1, β_2, and β_3 basically quantify the expected variation in N associated with a unit change in the binary predictor variables `Lopinavir`, `Efavirenz`, and the interaction term, respectively.

Maximum likelihood estimation (MLE) can be used to calculate logistic coefficients. In the case of `Nausea`, `Lopinavir` and `Efavirenz`, calculating logit coefficients using the hypothetical contingency table depicted in left-hand side of Fig. 3 and R package yields $N = -1.609 - 0.993 \times L + 0.226 \times E + 2.337 \times L \times E$ with the p-values depicted in right-hand side of Fig. 3. The interpretation would be that the interaction is statistically significant as the p-value for the interaction term is 0.0381, a value that is less than the usually accepted threshold of 0.05.

L	E	L*E	N	¬N
1	1	1	9	9
0	1	0	6	23
1	0	0	2	27
0	0	0	4	20

Logit coefficient	p-value
$\beta_0 = -1.609$	0.0033
$\beta_1 = -0.993$	0.2776
$\beta_2 = 0.226$	0.7099
$\beta_3 = 2.337$	0.0381

Fig. 3. Left: $2 \times 2 \times 2$ contingency table of reports for the regression of `Nausea` (`N`) on two exposure level `Lopinavir` (`L`), `Efavirenz` (`E`) and their interaction term `L*E`. **Right:** The corresponding logit coefficients provided by the R package.

4 Tools and Experiments

SignalMiner[3], is an open source tool dedicated to mining significant drug-reaction associations. The tool is coupled with, on the one hand, Galicia

[3] http://safetyseer.cvs.sourceforge.net/signalminer/

open-source platform[4] for handling FCA data including the input contexts, concept/iceberg lattices and rule basis, and on the other hand, the open-source statistical computing and graphics environment R[5] for data pre-processing and multivariate statistics including logistic regression analysis. In addition, SIGNALMINER performs a wide range of standard calculations, e.g., PRR, ROR, χ^2 (with Yates correction), etc.

The SRS DB of the French Medicines Agency (Afssaps) was used for the validating experiments. We have tested the proposed method on several moderate-size subsets of the dataset. For instance, for a pool of 3249 case reports containing 527 drugs and 639 ADR. The obtained lattice comprises 13178 concepts while the corresponding rule basis contains 28117 rules among them only 1165 represent candidates for pharmacological associations. These candidates are further distilled by SIGNALMINER to identify pure, semi-pure or biaised associations as illustrated in Table 4. Thus, the 1165 suggested association candidates (Table 4) are further filtered, on the one hand, by focusing potential safety signals satisfying the above MHRA '*interestingness*' criteria, and on the other hand, by focusing drug interactions that have been revealed significant using regression analysis. The minimum criteria for raising hypotheses regarding safety signals are as follows: number of reports (patients) ≥ 3, PRR ≥ 2, and $\chi^2 \geq 4$ (with Yates correction).

Table 4. Left: Candidates for pharmacological associations obtained from 3249 case reports containing 527 drugs and 639 ADR. **Right:** Statistically significant candidates.

	# Pure	# Semi-pure	# Biaised	
Signals	1	88	745	834
Interactions	1	260	70	331
	2	348	815	1165

	# Pure	# Semi-pure	# Biaised	
Signals	0	4	59	63
Interactions	0	0	10	10
	0	4	69	73

Among 834 candidates representing safety signals (Table 4), we have found that 63 candidates are statisticaly significant safety signals including 36 known signals (57%), e.g., {`Abciximab`, `Thrombopenia`}, 16 new signals warranting further investigations, e.g., {`Lamivudine`, `Arthralgia`}, while the remaining potential signals are either association where the drug appears as an innocent bystander, e.g., {`Ritonavir`, `Hypophosphatemia`}, or non-interpretable association, e.g., {`Bupivacaine`, `decrease of the therapeutic effect`}. In addition, among 331 associations representing candidates for drug interactions (Table 4), 10 candidates are revealed to be statistically interesting. In a previous work [3,4],

[4] http://www.iro.umontreal.ca/~galicia
[5] http://www.r-project.org/

disproportinality measures extracted 523 and 360 statistically significant {*drug*, ADR} couples, respectively. Our approach returns a smaller set of drug-reaction associations to be further investigated.

5 Related Work

Several studies from the literature address the use of DMA to identify drug-reaction associations. In [4], the use of FCA in signal detection is briefely addressed. To assess the strength of the association between a target drug and suspected ADR, disproportionality approach introduces several parameters such as, the PRR [5], χ^2 that is often coupled with the PRR, and the ROR [13], whereas Bayesian approach consists of the Multi-item Gamma Poisson Shrinker (MGPS) algorithm [11] and the Bayesian Confidence Propagation Neural Network (BCPNN) [2].

In [9], an interpretation of mathematical structures from FCA into epidemiology is described. A comprehensive survey of state-of-the-art in statistical modelling used by a various DMA of PV data is proposed in [7]. However, to the best of our knowledge, none of them supports automatic detection of pharmacological associations involving several drugs and/or reactions.

6 Discussion

FCA has been applied in combination with statistical metrics to the detection of several types of statistically significant pharmacological associations, e.g., safety signals and drug interactions. Compared to the classical DMA-based detection, the proposed FCA method improves the quantity and quality of extracted pharmacological associations, including those involving several drugs and/or reactions. Indeed, the amount of extracted associations is reduced by targeting basis of association rules using FCA framework, yet relevant associations with respect to the referred population of case reports, thereby saving investments in time and money that would be spent in further clinial trials.

In the future, we intent to reformulate drug-reaction analysis so that detecting pharmacological association is mapped to a relational data mining problem [8]. Moreover, because drug-reaction analysis deals with a dynamic DB that comprises high volume of data, the reconstruction –from scratch– of a new concept lattice for every change in the SRS DB is so computationally expensive that it is prohibitive. We shall address the on-line analysis of pharmacovigilance data using the incremental maintenance of concept lattice [12] and the respective association basis.

Acknowledgments

This work was supported partially by project grant (Ref. ANR-07-TecSan) from the French *Agence Nationale de la Recherche, Biologie et Santé* and *Caisse Nationale de Solidarité pour l'Autonomie*, and a discovery grant of the Natural Sciences and Engineering Research Council of Canada.

References

1. Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., Lakhal, L.: Mining minimal non-redundant association rules using fci. In: Palamidessi, C., Moniz Pereira, L., Lloyd, J.W., Dahl, V., Furbach, U., Kerber, M., Lau, K.-K., Sagiv, Y., Stuckey, P.J. (eds.) CL 2000. LNCS, vol. 1861, pp. 972–986. Springer, Heidelberg (2000)
2. Bate, A., Lindquist, M., Edwards, I.R., Olsson, S., Orre, R., Lansner, A., De Freitas, R.M.: A bayesian neural network method for adverse drug reaction signal generation. European Journal of Clinical Pharmacology 54(4), 315–321 (1998)
3. Bousquet, C., Sadakhom, C., Le Beller, C., Jaulen, M.-C., Louet, A.L.: A review of potential signals generated by an automated method on 3324 pharmacovigilance case reports. Therapie 61(1), 39–47 (2006)
4. Estacio-Moreno, A., Toussaint, Y., Bousquet, C.: Mining for adverse drug events with formal concept analysis. In: Proceedings of MIE 2008. Studies in Health Technology and Informatics, vol. 136, pp. 803–808. IOS Press, Amsterdam (2008)
5. Evans, S.J., Waller, P.C., Davis, S.: Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology and Drug Safety 10(6), 483–486 (2001)
6. Ganter, B., Wille, R.: Formal Concept Analysis, Mathematical Foundations. Springer, Heidelberg (1999)
7. Hauben, M., Madigan, D., Gerrits, C.M., Walsh, L., van Puijenbroek, E.P.: The role of data mining in pharmacovigilance. Expert Opinion on Drug Safety 4(5), 929–948 (2005)
8. Huchard, M., Rouane-Hacene, M., Roume, C., Valtchev, P.: Relational concept discovery in structured datasets. Annals of Mathematics and Artificial Intelligence 49(1-4), 39–76 (2007)
9. Pogel, A., Ozonoff, D.: Contingency structures and concept analysis. In: Medina, R., Obiedkov, S. (eds.) ICFCA 2008. LNCS, vol. 4933, pp. 305–320. Springer, Heidelberg (2008)
10. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with TITANIC. Data Knowledge Engineering 42(2), 189–222 (2002)
11. Szarfman, A., Machado, S.G., ONeill, R.T.: Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the us fdas reports database. Drug Safety 25(12), 381–392 (2002)
12. Valtchev, P., Rouane-Hacene, M., Missaoui, R.: A generic scheme for the design of efficient on-line algorithms for lattices. In: Ganter, B., de Moor, A., Lex, W. (eds.) ICCS 2003. LNCS, vol. 2746, pp. 282–295. Springer, Heidelberg (2003)
13. van Puijenbroek, E.P., Diemont, W.E., van Grootheest, K.: Application of quantitative signal detection in the dutch spontaneous reporting system for adverse drug reactions. Drug Safety 26, 293–301 (2003)

Mealtime Blood Glucose Classifier Based on Fuzzy Logic for the DIABTel Telemedicine System

Gema García-Sáez[1], José M. Alonso[2], Javier Molero[1], Mercedes Rigla[3], Iñaki Martínez-Sarriegui[1], Alberto de Leiva[4], Enrique J. Gómez[1], and M. Elena Hernando[1]

[1] Bioengineering and Telemedicine Centre, Politechnical University of Madrid, CIBER-BBN Networking Research Centre, Spain
{ggarcia,jmolero,imartinez,egomez,elena}@gbt.tfo.upm.es
[2] European Centre for Soft Computing, Mieres (Asturias), Spain
jose.alonso@softcomputing.es
[3] Endocrinology Dept., Hospital de Sabadell, CIBER-BBN Networking Research Centre, Spain
mrigla@tauli.cat
[4] Endocrinology Dept., Hospital Sant Pau, Barcelona, CIBER-BBN Networking Research Centre, Spain
aleiva@santpau.es

Abstract. The accurate interpretation of Blood Glucose (BG) values is essential for diabetes care. However, BG monitoring data does not provide complete information about associated meal and moment of measurement, unless patients fulfil it manually. An automatic classification of incomplete BG data helps to a more accurate interpretation, contributing to Knowledge Management (KM) tools that support decision-making in a telemedicine system. This work presents a fuzzy rule-based classifier integrated in a KM agent of the DIABTel telemedicine architecture, to automatically classify BG measurements into meal intervals and moments of measurement. Fuzzy Logic (FL) tackles with the incompleteness of BG measurements and provides a semantic expressivity quite close to natural language used by physicians, what makes easier the system output interpretation. The best mealtime classifier provides an accuracy of 77.26% and does not increase significantly the KM analysis times. Results of classification are used to extract anomalous trends in the patient's data.

Keywords: Diabetes, Telemedicine, Fuzzy Logic, Classification.

1 Introduction

Diabetes management should avoid acute and long-term complications that can be responsible for premature death and disability [1]. Capillary Blood Glucose (BG) testing is the standard monitoring tool used in the care of people with diabetes, giving a guide to BG levels at a specific moment in time [2]. Accuracy and

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 295–304, 2009.

completeness of BG measurements is vital because physicians modify the insulin treatment on the basis of these data [3]. For physicians, a correct interpretation of BG patterns requires the evaluation of the effect of insulin doses administered by patients according to each mealtime interval. Patients should enter additional data manually after each capillary BG measure to register the moment of measurement (preprandial or before a meal, postprandial, or 2-3 hours after a meal or other moment of the day) or the meal-intake associated (breakfast, morning, lunch, afternoon, dinner, night). However, this is considered an extra effort and it is not usually completed by patients.

Patients with type 1 diabetes are also recommended to monitor daily other variables that affect their health state, such as administered insulin doses, ingested diet, exercise, or other relevant situations. This requirement increases the amount of data generated in diabetes care and the patient's and physician's workload when analysing data to take therapeutic decisions. In such scenario, it is important to have the appropriate tools for data registration and for decision-making that filter the huge amount of data both to patients and physicians and detect anomalous situations in the patient's health. Incomplete BG data is one of the factors that limit the effectiveness of decision support systems in diabetes.

The use of telemedicine services for diabetes has already proved that contributes to patient self-monitoring, improves glycaemic control [4,5,6] and provides decision support tools able to detect anomalous situations in the patient's health automatically [7]. The DIABTel telemedicine system supports the patient's self-monitoring of variables that affect the metabolism of the patient. Knowledge Management (KM) tools have been implemented in DIABTel with KM agents able to preprocess, filter and analyze automatically monitoring data, to detect anomalous patterns. A KM agent for BG data provides decision-making processes, detects risk situations in the patient's health from BG measurements and generates automatic warnings about the patient's glycaemic anomalies that are sent to patients and/or physicians.

An automatic classification of incomplete BG data helps physicians to a more accurate interpretation, being essential to detect anomalous patterns associated to meal-intakes. However, in BG classification into mealtime intervals, the limits that determine the association to a specific meal-intake interval or moment can be unclear and dependent on several normally unknown aspects of the patients' daily circumstances. In this situation, the classification should not be based on fixed time-intervals, being appropriate the use of Fuzzy Logic (FL).

FL was introduced by Zadeh (in 1965) and it is widely acknowledged for its well-known ability for linguistic concept modelling. The FL semantic expressivity, using linguistic variables [8] and linguistic rules [9], is quite close to expert natural language. In classical logic only two crisp values are admissible, what is a strong limitation in order to handle real-world problems where the available information is usually vague. In the real world things are not so simple as black and white but there is a continuous scale of greys. This is the typical situation regarding BG classification, which is not clearly determined and depends on the patient's habits. To cope with this problem, FL is a useful tool because working

with FL everything has a membership degree. Moreover, interpretability is admitted as the main advantage of fuzzy rule-based systems (FRBSs). Classical systems are usually viewed as black boxes because mathematical formulas set the mapping between inputs and outputs. On the contrary, fuzzy systems (if they are built regarding a few constraints) can be seen as white boxes in the sense that every element of the whole system can be checked and understood by a human being. Thus, interpretability is essential for those applications with high human interaction, for instance decision support systems for physicians in medicine. Notice that the use of an interpretable FRBS as a mealtime BG classifier makes easier the understanding of classification results by physicians. A FRBS for classification tasks is called fuzzy rule-based classifier (FRBC).

This work describes a FRBC that classifies automatically BG measurements into meal intervals and moments of measurement. The classifier is integrated in the KM agent of the DIABTel architecture. The rest of the paper is structured as follows. Section 2 introduces the global architecture of the DIABTel telemedicine system. Section 3 describes the implementation methodology of the Fuzzy Classification system. Section 4 shows some experimental results. Finally, a brief discussion regarding conclusions and future works is presented in Section 5.

2 The DIABTel Telemedicine System

DIABTel is a telemedicine system for patients with type 1 diabetes [10]. The architecture of the system allows different multi-access services such as: (1) *Telemonitoring*, which allows physicians to view the patient's self-monitoring data; (2) *Telecare*, which enables assessment of the patient's metabolic state, therapy modification and supervision of patient's decisions; (3) *Remote information access*, to access basic visit information, complete Electronic Health Records and stored messages; and (4) *Knowledge Management tools*, to supply doctors and patients with the knowledge needed to analyze monitoring data and/or to make diagnostic and management decisions. DIABTel includes automatic reports generation and intelligent warnings notification.

Patients can upload BG data automatically to the central database server from different glucose meters. The patient's connection to the DIABTel service is supported by a PDA-based device which runs the Personal Assistant application [11], and allows communicating with a commercial insulin pump. Insulin data administered can be uploaded automatically via infrared communication from the insulin pump to the PDA. The Personal Assistant makes possible the collection of other monitoring data such as diet or additional information (exercise, stress, illness, etc.) or to complete data such as BG measures, manually. The information registered is sent to the DIABTel center via mobile GPRS.

The KM agent in charge of BG data analysis is devoted to make easier the correct interpretation of self-monitoring data to patients and physicians, extracting as much information as possible from raw data downloaded straight from the glucose meter. It is required to separate different measurements in moments of measurement and meal associated time-intervals, in order to check the postprandial excursion (post-meal effect) of insulin boli before each meal-intake as

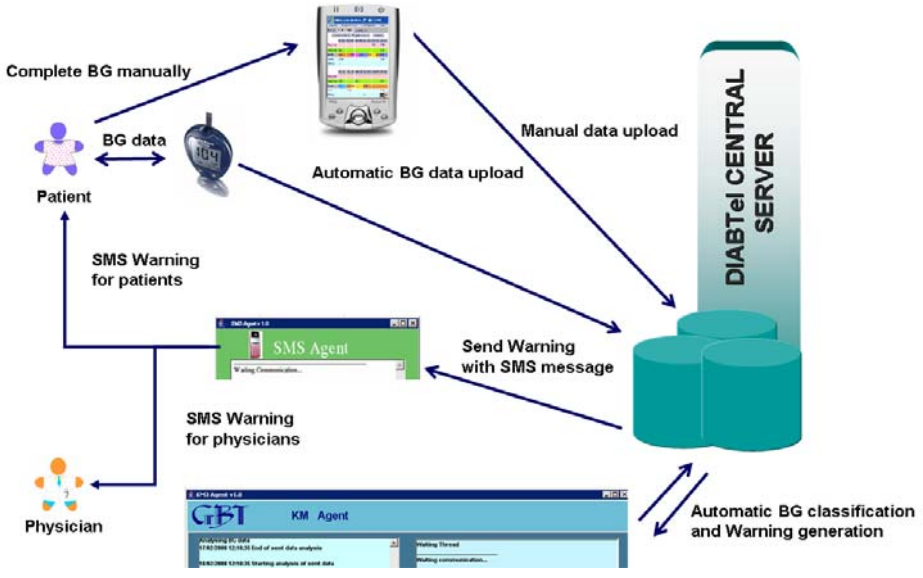


Fig. 1. BG data analysis with the KM agent in the DIABTel architecture

well as to evaluate basal insulin doses. The KM agent generates automatic reports about the patient's evolution after each data sending and it sends SMS or e-mail messages, according to the user's preferences, to warn automatically patients and/or physicians whenever any anomalous situation is detected (Fig. 1). The KM results present the information regarding BG measures divided into time intervals associated to meal-intakes and moments of measurement, which requires using the fuzzy classifier described in this paper. The next section describes the methodology followed to implement the fuzzy rule-based classifier (FRBC) integrated in the DIABTel telemedicine system.

3 Description of the Fuzzy Classifier

3.1 Fuzzy Modelling Methodology

A FRBC is usually implemented in the form of a fuzzy inference system. First, a fuzzification module converts the numerical inputs into fuzzy values. Then, an inference engine puts into effect the fuzzy inference process on the system inputs making use of the information stored in a previously generated knowledge base (linguistic variables and rules). Finally, a defuzzification module computes numerical outputs from the fuzzy outputs generated by the inference engine, establishing the output class. Our FRBC was designed and implemented using KBCT (Knowledge Base Configuration Tool) [12], a free software tool which implements the HILK (Highly Interpretable Linguistic Knowledge) fuzzy modelling methodology [13].

HILK focuses on building interpretable fuzzy classifiers, i.e., classifiers easily understandable by human beings. The semantic expressivity of FL makes easier the knowledge extraction and representation phase. In addition, HILK offers an integration framework for combining under the same FL formalism both expert knowledge (expressed in natural language) and induced knowledge (automatically extracted from data) which is likely to yield robust compact systems with a good trade-off between accuracy and interpretability. As a result, useful pieces of knowledge are formalized and represented by means of linguistic variables and rules also called Mamdani rules [9]. Defining a global semantics previous to the rule definition makes easier the rule understanding. Only if all the rules, expert and induced ones, use the same linguistic terms (defined by the same fuzzy sets) it will be possible to make a rule comparison at the linguistic level. Notice that, we have selected the Max-Min inference scheme (maximum T-conorm as aggregation and minimum T-norm as relational and conjunctive operator), and the winner rule fuzzy reasoning mechanism.

3.2 Linguistic Variables

The knowledge base of our FRBC is made up of four input variables and one output. The design of each variable was made regarding both expert and induced knowledge, and taking into consideration several options for each variable, thinking on getting a good accuracy-interpretability trade-off. The number of labels was initially defined by the expert who defined the following input variables:

- *Measurement time*: Absolute minute when the measurement was taken. This parameter considers the earliest wake-up time for each patient as the initial minute. Initially, six labels corresponding to the main meal-time intervals were associated to each intake (*Breakfast*, *Morning*, *Lunch*, *Afternoon*, *Dinner*, and *Night*).
- *Time difference with previous measurement*: The time in minutes that passed after the last registered measurement, for measurements taken the same day. Four labels (*Zero*, *Low*, *High*, and *Very high*) were initially defined with the aim of showing if the time difference between two consecutive measurements is big or small. For example, a low time difference between two measurements might mean a repeated value.
- *Blood Glucose value*: The aim of this variable is to reflect if the BG value is high or low. Four labels were defined (*Low*, *Normal*, *High*, and *Very high*).
- *Insulin bolus administered close to a BG measurement*. This variable is Boolean with two labels that show if there is an insulin bolus associated to each BG measurement (*true*) or not (*false*). It allows the correct classification of preprandial measurements, usually taken close in time to an insulin bolus administration. It is computed as the output of another fuzzy classifier that includes three Boolean inputs (*Same year*, *Same month* and *Same day*), one fuzzy input (*Time difference*) with two labels (*small*/*large*), the inferred output (*true*/*false*), and the following linguistic rules:

- **IF** *Same year is false* OR *Same month is false* OR *Same day is false* **THEN** *Insulin bolus administered is false*
- **IF** *Same year is true* AND *Same month is true* AND *Same day is true* AND *Time difference is large* **THEN** *Insulin bolus administered is false*
- **IF** *Same year is true* AND *Same month is true* AND *Same day is true* AND *Time difference is small* **THEN** *Insulin bolus administered is true*

Regarding the output variable (*Classification Result*) two options are considered. Firstly, nine labels were defined: *Breakfast Preprandial*, *Breakfast Postprandial*, *Lunch Preprandial*, *Lunch Postprandial*, *Dinner Preprandial*, *Dinner Postprandial*, *Night*, *Repeated* (for repeated measurements due to failures in the process of measurement), and *Intermediate* (for measurements along the morning or the afternoon). Secondly, the label *Intermediate* was divided into two additional labels: *Morning* and *Afternoon*.

3.3 Rule Base Design

The goal is to test different FRBCs comparing them in terms of accuracy and complexity (number of rules). First, the expert described the classification procedure, combining the previously defined linguistic variables in the form of linguistic rules. As a result, ten simple rules were formalized for the *Expert* FRBC. Two of them are presented below:

- **IF** *Measurement time is Breakfast* AND *Time difference with previous measurement is Zero* **THEN** *Classification Result is Breakfast Preprandial*
- **IF** *Measurement time is Breakfast* AND *Time difference with previous measurement is High* **THEN** *Classification Result is Breakfast Postprandial*

Then, HILK was used to automatically generate rules from data. Keeping in mind the interpretability goal, we have chosen the Fuzzy Decision Tree (FDT) [14], a fuzzy version of the popular decision trees defined by Quinlan [15]. Notice that our implementation of FDT is able to build quite general rules with the interpretable partitions previously defined. Such rules do not need to consider all available variables, what is really appreciated from the interpretability point of view. We integrated in a unique FRBC (*Expert-FDT*) the set of expert rules and rules induced from data. The integration of expert and induced rules requires a consistency checking and a simplification procedure to remove redundancies and to get more compact and understandable variables and rules. After this step, we obtained a new FRBC (*Expert-FDT-Simplified*).

In a second stage, another FRBC (*FDT-20labels*) was fully generated from data with the aim of getting higher accuracy. The granularity of the numerical input variables (*Measurement time*, *Time difference with previous measurement*, *Blood Glucose value*) was increased up to twenty equally spaced labels, while the input variable *Insulin bolus administered close to a BG measurement* was kept Boolean. After simplification, the achieved FRBC (*FDT-20labels-Simplified*) was

chosen as the basis of the final fuzzy classifier. However, other alternatives were also considered to improve accuracy results, such as varying the data set used for training or changing the number of input or output labels.

Finally, the selected FRBC has been evaluated regarding classification times of BG data sets within the KM Agent. This parameter is considered very important in order to determine the classifier's efficiency in the telemedicine system.

3.4 Validation Method

Data for training and validation of the implemented approaches for FRBCs have been obtained from two randomized cross-over studies with patients from the "Hospital de Sant Pau" [10]. BG measurements were obtained straight from the patient's glucose meter and insulin data were downloaded automatically from a Disetronic^{TM} insulin pump using the Personal Assistant. The same ten patients participated for a period of two months in each one of the two experiments: Experiment 1 and 2 (EXP-1, EXP-2).

The whole data set was divided into different parts in order to get different validation data sets: (1) EXP-1, which contains all data from the first clinical experiment with 2500 measurements; (2) EXP-2a which is limited to data for the experimental period of EXP-2 contains 1293 measurements; (3) EXP-2b which is limited to control data from EXP-2 with 949 measurements; and (4) EXP-12 which includes EXP-1 plus EXP-2a. EXP-2 was divided in two different data sets (EXP-2a and EXP-2b) because the distribution of BG values into mealtime and moments of measurement was significantly different in each period [5]. BG measurements in each data set were classified manually by the physician in charge of visiting the group of patients, who had a good knowledge on the patient's usual measurement times and their lifestyle behavior. The expert classification was used to validate results obtained with different FRBCs.

The evaluation of the designed FRBCs was based on a Z-fold cross-validation made over the whole data set. Cross-validation is a method for estimating generalization error based on re-sampling [16]. The whole data set is divided into Z parts of equal size, and each part keeps the original distribution (percentage of elements for each class). Then, the same process is repeated Z times. One part is used as test set while the other Z-1 parts are used as training set. Notice that the Z (Z=10 in our case) results are averaged.

4 Results of the Fuzzy Classification

Table 1 summarizes the classification results obtained with the five FRBC proposed initially using the cross-validation method. In a first stage, EXP-1 was used as training data set. The first column contains the abbreviations of the combined methods for generating FRBCs. Second and third columns show the accuracy regarding both training and test data sets. Accuracy is computed as the percentage of cases correctly classified. Fourth and fifth columns give an

Table 1. Fuzzy classifiers evaluation (10-fold cross-validation, EXP-1)

FRBC	Accuracy (Training)	Accuracy (Test)	Rules	Premises
Expert	63.7 %	63.7 %	10	2
Expert-FDT	80.2 %	79.8 %	33.4	2.27
Expert-FDT-Simplified	83.8 %	81.9 %	19.6	2.26
FDT-20labels	95.5 %	83.7 %	1181.9	2.89
FDT-20labels-Simplified	95.5 %	83.2 %	275.4	2.84

Table 2. Evaluation of the selected FRBC on experiment data sets

FRBC	Training data set	Accuracy (Test) (EXP-12)	(EXP-2a)	(EXP-2b)	Rules
FTD-20labels-Simplified	EXP-1	90.88 %	80.12 %	74.51 %	295

idea on the complexity of each FRBC which is measured regarding the number of rules and the number of premises (inputs used per rule). The *Expert* FRBC only considers expert knowledge and it is quite simple and interpretable but at the cost of a small accuracy. The combination of expert and induced knowledge (*Expert-FDT*) yields a more complex but robust FRBC. Notice that after applying the simplification procedure (*Expert-FDT-Simplified*), both accuracy and interpretability are improved. Finally, the *FDT-20labels* obtains better accuracy but at the cost of increasing a lot the FRBC complexity. *FDT-20labels-Simplified* keeps the high accuracy reducing significantly the number of rules. The number of premises per rule is smaller than three in all the cases.

Achieved results were presented to the experts, who selected *FDT-20labels-Simplified* as the best method according to their requirements for both accuracy and interpretability. They gave priority to get a high classification rate. Then, the selected FRBC was tested on data from EXP-12, EXP-2a and EXP-2b. Results are presented in Table 2. After iterative steps to improve accuracy, a last FRBC was generated using EXP-12 data set for rules induction. It considers 10 output labels instead of 9 because we noticed that many errors could be avoided distinguishing between morning and afternoon for the *Intermediate* measures. The number of labels for the input variable *BG value* is also increased up to 25. The total number of rules for this FRBC is 564. Since EXP-12 includes both EXP-1 and EXP-2a, the generalization of the method is evaluated over the EXP-2b data set, which was not used during the training phase. Accuracy results obtained after classifying EXP-2b were of 77.26%. The final FRBC improves accuracy at the cost of losing some interpretability. For the ten output labels

FRBC, given an input vector, the number of rules simultaneously fired is 2.192 in average, being 7 in the worst case. This supports the interpretability of the whole system despite the high number of rules. In addition, using the FRBC allows handling uncertainty associated to mealtime intervals in contrast to traditional classification based on fixed intervals getting as a result a system more robust and accurate.

Considering the good reported results of the last FRBC, it has been used to implement the BG classifier integrated in the KM agent of the DIABTel system. After receiving BG data from a patient, the extra computation time that the KM Agent requires to run FRBC ranges from 1 to 3 seconds for the usual amount of BG data sent by patients (a week).

5 Conclusions and Future Works

We have implemented a FRBC for the KM agent used to analyze BG measurements in the DIABTel telemedicine system. The final number of rules is high, although the number of rules simultaneously fired by the system is not so large, what along with the interpretable structure of the FRBC (strong fuzzy partitions, global semantics, small number of inputs per rule, etc.) favors the interpretability of the final model. Interpretability makes easier understanding the system from the physician's perspective and it is an important issue to consider when explaining to physicians the results of the classification model.

From the application point of view, the implemented classifier provides very fast classification times for the usual size of BG data sent by patients to the telemedicine centre. Consequently, the use of the fuzzy classifier is not time consuming and does not increase analysis tasks substantially, so it is a valuable tool to complement and improve the quality of analysis results obtained regarding BG measurements. The proposed FRBC requires adjusting the earliest wake-up time to each patient for configuration, not being necessary to make changes for different periods such as weekends or working days or to adjust other individual parameters daily.

The tool is very useful to complete BG reports about the patient's glycaemic control and allows determining the effect of therapeutic actions taken by patients associated to different meal-intake intervals or moments of measurement. The analysis of classified BG measurements along time detects anomalous trends in different moments of measurement and provides a detailed description of different meal-intake intervals, which is currently used for warning generation within the DIABTel system.

Future works will be addressed to check if creating different FRBCs for days with different habits such as holidays, weekends or working days would allow increasing even more the accuracy of the classifier. Other future tasks would be to determine if differentiating groups of patients with similar measurement habits is able to increase the accuracy of the tool.

References

1. The Diabetes Control and Complications Research Group: The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. New England Journal of Medicine 329, 977–986 (1993)
2. International Diabetes Federation (IDF): Clinical guidelines task force, global guideline for type 2 diabetes, 2005. Technical report, TR (2005)
3. Benjamin, E.M.: Self-monitoring of blood glucose: The basics. Clin Diabetes 20(1), 45–47 (2002)
4. Shea, S., Weinstock, R.S., Starren, J., et al., for the IDEATel Consortium: A randomized trial comparing telemedicine case management with usual care in older, ethnically diverse, medically underserved patients with diabetes mellitus. Journal of the American Medical Informatics Association 13(1), 40–51 (2006)
5. Rigla, M., Hernando, M.E., Gómez, E.J., Brugués, E., García-Sáez, G., Capel, I., Pons, B., de Leiva, A.: Real-time continuous glucose monitoring together with telemedical assistance improves glycemic control and glucose stability in pump-treated patients. Diabetes Technology & Therapeutics 10(3), 194–199 (2008)
6. Bellazzi, R., Arcelloni, M., Ferrari, P., Decata, P., Hernando, M.E., García, A., Gazzaruso, C., Gómez, E.J., Larizza, C., Fratino, P., Stefanelli, M.: Management of patients with diabetes through information technology: Tools for monitoring and control of the patients' metabolic behavior. Diabetes Technology & Therapeutics 6(5), 567–578 (2004)
7. Dinesen, B., Andersen, P.E.R.: Qualitative evaluation of a diabetes advisory system, diasnet. Journal of Telemedicine and Telecare 12(2), 71–74 (2006)
8. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning. Parts I, II, and III. Information Sciences 8, 8, 9, 199–249, 301–357, 43–80 (1975)
9. Mamdani, E.H.: Application of fuzzy logic to approximate reasoning using linguistic systems. IEEE Transactions on Computers 26(12), 1182–1191 (1977)
10. Gómez, E.J., Hernando, M.E., Vering, T., Rigla, M., Bott, O., García-Sáez, G., Pretschner, P., et al.: The inca system: A further step towards a telemedical artificial pancreas. IEEE Transactions on Information Technology in Biomedicine 12(4), 470–479 (2008)
11. García-Sáez, G., Hernando, M.E., Martínez-Sarriegui, I., Rigla, M., Torralba, V., Brugués, E., de Leiva, A., Gómez, E.J.: Architecture of a wireless personal assistant for telemedical diabetes care. International Journal of Medical Informatics 78(6), 391–403 (2009)
12. Alonso, J.M., Guillaume, S., Magdalena, L.: KBCT: A knowledge management tool for fuzzy inference systems. Free software under GPL license (2003), http://www.mat.upm.es/projects/advocate/kbct.htm
13. Alonso, J.M., Magdalena, L., Guillaume, S.: HILK: A new methodology for designing highly interpretable linguistic knowledge bases using the fuzzy logic formalism. International Journal of Intelligent Systems 23, 761–794 (2008)
14. Ichihashi, H., et al.: Neuro-fuzzy ID3: A method of inducing fuzzy decision trees with linear programming for maximizing entropy and an algebraic method for incremental learning. Fuzzy Sets and Systems 81, 157–167 (1996)
15. Quinlan, J.R.: Induction of decision trees. Machine Learning 1, 81–106 (1986)
16. Plutowski, M., Sakata, S., White, H.: Cross-validation estimates imse. In: Cowan, J.D., Tesauro, G., Alspector, J. (eds.) Advances in Neural Information Processing Systems 6, pp. 391–398. Morgan Kaufmann, San Mateo (1994)

Providing Objective Feedback on Skill Assessment in a Dental Surgical Training Simulator

Phattanapon Rhienmora[1], Peter Haddawy[1],
Siriwan Suebnukarn[2], and Matthew N. Dailey[1]

[1] School of Engineering and Technology, Asian Institute of Technology,
Pathumthani, Thailand, 12120
{phattanapon.rhienmora,haddawy,mdailey}@ait.ac.th
[2] Faculty of Dentistry, Thammasat University,
Pathumthani, Thailand, 12121
ssiriwan@tu.ac.th

Abstract. Dental students devote several years to the acquisition of sufficient psychomotor skills to prepare them for entry-level dental practice. Traditional methods of dental surgical skills training and assessment are being challenged by the complications such as the lack of real-world cases, unavailability of expert supervision and the subjective manner of surgical skills assessment. To overcome these challenges, we developed a dental training system that provides a VR environment with a haptic device for dental students to practice tooth preparation procedures. The system monitors important features of the procedure, objectively assesses the quality of the performed procedure using hidden Markov models, and provides objective feedback on the user's performance for each stage in the procedure. Important features for characterizing the quality of the procedure were identified based on interviews with experienced dentists. We evaluated the accuracy of the skill assessment with data collected from novice dental students as well as experienced dentists. We also evaluated the quality of the system's feedback by asking a dental expert for comments. The experimental results show high accuracy in classifying users into novice and expert, and the evaluation indicated a high acceptance rate for the generated feedback.

Keywords: Dental surgical training, skill assessment, objective training feedback, virtual reality.

1 Introduction

Dental students obtain their surgical skills training from various sources. Traditional methods rely on practicing procedural skills on plastic teeth or live patients under the supervision of dental experts. However, the limitations of this approach include a lack of real-world challenging cases, unavailability of expert supervision, and the subjective manner of surgical skills assessment. With recent advances in virtual reality (VR) technology, VR simulators for medical and dental surgery have been introduced [1], [2]. The advantages of these simulators are that the students are able to practice

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 305–314, 2009.

procedures as many times as they want at no incremental cost and that the training can take place anywhere. The realism of these simulators has increased with the introduction of haptic devices that provide tactile sensations to the users [3], [4], [5], [6].

Skill assessment in traditional training is usually conducted by having an expert surgeon observe the procedure or only the final outcome. However, the level of detail of human expert assessment is limited. With VR simulators, many aspects such as data about the environment and the user's precise actions can be recorded during the simulation and analyzed further to provide fine-grained objective assessment and feedback. Unfortunately, existing dental simulators do not provide this functionality. There has been some work in other fields, however. Rosen et al. [7] present a technique for objective evaluation of laparoscopic surgical skills using hidden Markov models (HMMs). The models are based on force/torque information obtained from a surgical robot. Lin et al. [8] collected various measurements from the da Vinci surgical robot while an operator performed a suturing task. The aim of their study was to automatically detect and segment surgical gestures, which is a part of their ongoing research on automatic skills evaluation. As the da Vinci surgical robot does not provide haptic feedback, their research did not consider force applied during the operation.

To add more educational value, simulators should be able to provide objective feedback to users in order to reduce the time and effort required for instructors to supervise and tutor trainees using the system. Thus, incorporation of strategies for generating objective feedback with quality comparable to that of human tutors is essential to the development of an efficient, *intelligent* training simulator.

In this paper, we describe the first virtual reality dental training system to combine realistic haptic feedback with an objective dental performance assessment and feedback generation mechanisms. While the system currently simulates the tooth preparation procedure, many of the techniques and strategies implemented should generalize well to other medical and dental procedures.

2 VR Tooth Preparation Simulator

The graphical user interface of our simulator is illustrated in Fig. 1. Movement of a virtual dental tool is controlled by a haptic device stylus. The detailed development of our simulator is explained in [9]. Currently the system simulates only labial and incisal preparations in order to avoid conflating tool skills with indirect vision skills.

The tooth preparation procedure requires that 13 stages be performed on the incisal and labial surfaces including, 1) mid-incisal depth cut, 2) distal-incisal depth cut, 3) mesial-incisal depth cut, 4) incisal reduction, 5) mid-upper-labial depth cut, 6) distal-upper-labial depth cut, 7) mesial-upper-labial depth cut, 8) upper-labial reduction, 9) mid-lower-labial depth cut, 10) distal-lower-labial depth cut, 11) mesial-lower-labial depth cut, 12) lower-labial reduction, and 13) labial marginal preparation. Examples of simulated tooth preparation outcomes are shown in Fig. 2.

We tested the ability of our simulator to produce outcomes that reflect operator skill. Ten simulated preparation outcomes completed by five students and five experienced dentists were shown to another expert, who was not aware of the nature of the experiment, who was asked to assign outcome scores based on errors found in Incisal, Labial-incisal, Labial-gingival, and Marginal. The maximum score was 16. The

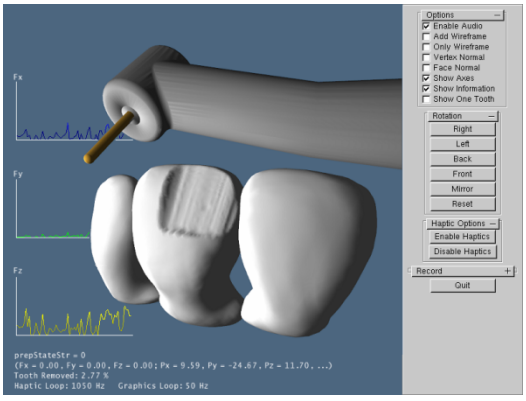


Fig. 1. Graphical user interface of our VR tooth preparation simulator

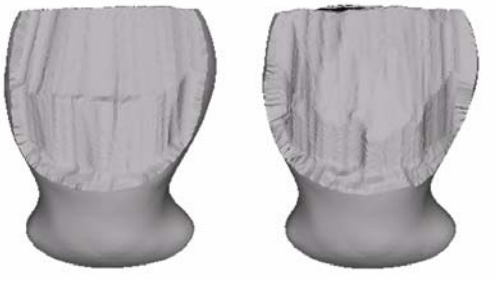


Fig. 2. Example of two outcomes of tooth preparation on the labial and incisal surfaces: an expert outcome *(left)*; a novice outcome *(right)*

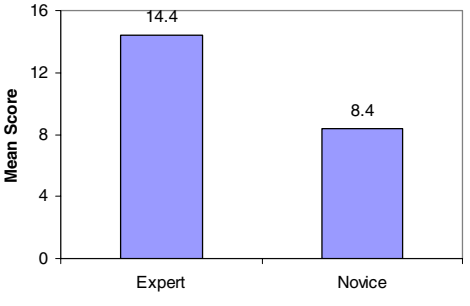


Fig. 3. Mean scores for simulated tooth preparation performed by expert *(left)* and novice *(right)*. Mean expert score = 14.4, SD = 0.89; mean novice score = 8.4, SD = 1.14.

experts' mean score (14.4) was significantly difference than the novices' mean score (8.4) ($p < 0.05$) as shown in Fig. 3. This result indicates that the simulator captures the important aspects of the physical environment.

3 Objective Assessment of Dental Surgical Skills

The current means for evaluating clinical performance and skill acquisition during training are limited to measurement of task completion time and number of errors or a subjective evaluation by an expert [10]. The aforementioned measures do not characterize the operator's movements (e.g., position, orientation, or speed). While speed is closely related to task completion time, faster is not necessarily better; the speed–accuracy trade-off is a well-known phenomenon in motor control, in which speed increases cause decreases in accuracy and vice versa [11]. More accurate movements may take more time to complete. Therefore, additional objective measures are needed to quantify surgical performance improvements and differentiate between expert and novice surgeons.

Based on interviews with experienced dentists, we hypothesized that important features for distinguishing experts from novices in dental surgery are tool movement

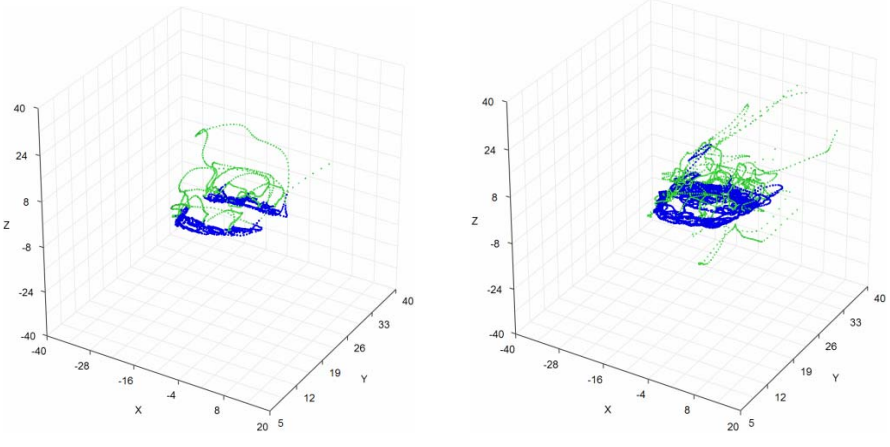


Fig. 4. Example of tool paths of an expert *(left)* and a novice *(right)*. *Darker* paths indicate cutting operation.

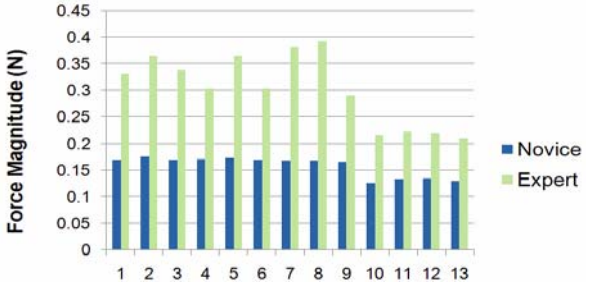


Fig. 5. An example of average force applied by an expert and a novice during 13 stages of simulated tooth preparation

(position and orientation of the tool) and applied force during a procedure. We can visualize these features by plotting tool movement of an expert and a novice in three dimensions as shown in Fig. 4 and the average magnitude of the force applied by an expert and a novice over time as shown in Fig. 5.

It can be clearly seen from Fig. 4 that expert and novice performance in tool movement is different. Expert movement, especially while cutting, is more consistent than novice movement. From Fig. 5, the force used at each stage of the procedure for experts and novices are also different. These measures can provide a foundation for quantifying surgical expertise and can be used for objective skill assessment.

3.1 Experiment

The main objective of this experiment was to test the ability of a machine learning technique, the hidden Markov model, to recognize and classify an observed procedure as novice or expert, based on a set of recorded important features.

Five novices (forth-year dental students, ages 20-22 years) and five experts in prosthodontics (ages 35-45 years) were recruited to participate in this study. All participants were right-handed.

Their task was to perform a tooth preparation on the upper left central incisor with the simulator. Experts and novices performed five trials of the task. The last trial was used for data analysis.

3.2 Evaluation

When a user performs tooth preparation on our simulator, all of the data relevant to the user's actions are monitored and recorded to a file. This data includes all important features mentioned previously as well as the active status of the drill and the indices of the vertices being cut on the tooth surface. We manually labeled the preparation stage transitions in order to facilitate later evaluation of automatic stage segmentation strategies.

After collecting the data from all participants, we developed discrete hidden Markov models (HMMs) to classify procedure sequences as novice or expert. In our model, the hidden states are the thirteen stages of tooth preparation. The observed feature set includes force calculated during the simulation as well as positions and orientations of the dental tool. Stage labels were not used in training HMMs. Since we use discrete HMMs, we first converted the feature vectors into symbols using the k-means clustering algorithm with $k = 13$. After training, we calculated the probability and log likelihood of test sequences under the novice and expert HMMs. If the log likelihood of the test sequence under the novice HMM is greater than that under the expert HMM, the system classifies the test sequence as a novice sequence; otherwise, the system classifies it as an expert sequence.

3.3 Result and Discussion

We used a different k-means for every cross validation fold and the same k-means for the novice and expert model in the same fold. For each fold, we trained the novice HMM with four novice and four expert sequences. To determine the accuracy of the

Table 1. Average log likelihood results for expert and novice performance sequences

	Log likelihool for Expert HMM	Log likelihool for Novice HMM
Expert Performance	-3.574×10^{3}	-2.229×10^{6}
Novice Performance	-6.272×10^{5}	-3.494×10^{3}

method, after training the two HMMs in each fold, we fed the test novice and expert data to each model. The average log likelihood of all sequences across all five folds for the two HMMs is shown in Table 1.

For every cross validation fold, the log likelihood of a test sequence for its corresponding HMM was higher than that for another HMM. The result demonstrated the ability of HMM to distinguish between novice and expert performance with 100 percent accuracy. However we note that the number of participants (ten) was relatively small.

4 Strategies for Objective Feedback Generation

The stage of the tooth preparation procedure and its unique force/position/orientation characteristic are the basis of our feedback generation mechanism. The average position, orientation, force, and main axis for force direction differ between procedure stages. In stage 1) (Fig. 6), for example, force and tool movement is mostly in the minus Y direction, while in stage 5) (Fig. 6) they progress mostly in the minus Z direction. These characteristics can be observed by the simulator and compared to a gold standard in order to generate useful feedback. Examples of our feedback strategy considering applied force for stage 1) and 5) are shown in Table 2.

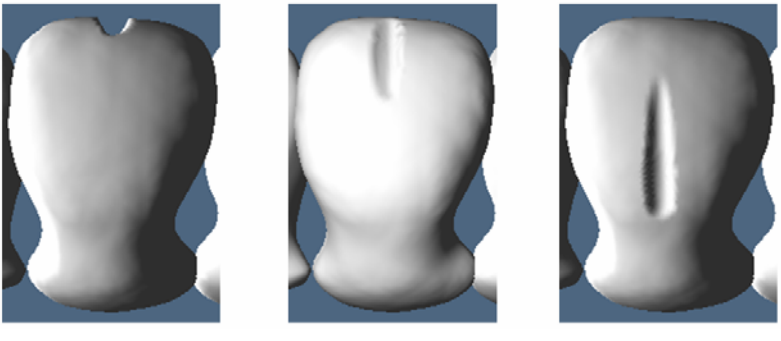


Fig. 6. Examples of stage 1) mid-incisal depth cut *(left)*, stage 5) mid-uppper-labial depth cut *(middle)* and stage 9) mid-lower-labial depth cut *(right)*

Table 2. Examples of feedback generated in stages 1) and 5) considering only applied force. Subscript *e* indicates the expert average value (out of five experts) with one standard deviation while *n* indicates the current novice value. The full table considers every feature (force, position, and orientation) and covers all 13 stages for each novice.

Stage	F_x (N)	F_y (N)	F_z (N)	Feedback
1_e	0.103±0.037	0.480±0.047	0.106±0.023	*"Force in minus Y direction should be 3 times higher"*
1_n	0.026	0.164	0.091	
5_e	0.040±0.014	0.038±0.019	0.237±0.053	*"Force in minus Z direction should be 2 times higher"*
5_n	0.028	0.019	0.129	

We generate feedback for position and orientation with the same strategy. For example, Fig. 7 shows a stage in which a novice's tool orientation was too different from that of the expert. In this case the feedback generated was *"try to lower the degree of rotation around X axis."* For states in which the operator does well, we generate a compliment such as "*well done."*

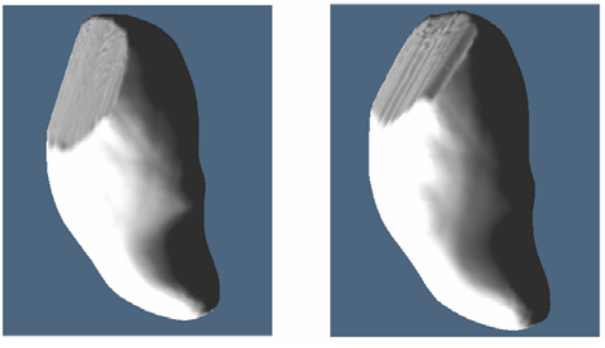


Fig. 7. Example of a difference in tool orientation between expert *(left)* and novice *(right)*

4.1 Experiment

The main objective of this experiment was to test the overall acceptability of the training feedback generated by the simulator.

The simulator loaded data files of all five novices collected during the previous experiment described in section 3.1 and replayed the procedure, one novice at a time. During playback, the system observed the characteristics of each stage, computed statistical results, compared them with those of expert performances, and then generated and displayed the tutoring feedbacks on the screen. An expert examined both the replay of the novice procedure and the feedback generated by the system. The corresponding force values in three axes were also plotted on the screen during replay to aid understanding of how the forces were applied by the operator.

During the experiment, a total of 65 tutoring feedback messages were generated. The expert was asked to rate the acceptability of each feedback message on a scale of 1-5, where 1 implied unacceptable, 2 implied not quite acceptable, 3 implied not sure, 4 implied close to acceptable and 5 implied acceptable.

4.2 Result and Discussion

Please see Fig. 6 as a reference for the desired outcomes of stages 1), 5) and 9).

From Table 3, during stage 5), where the main force should be applied in the minus Z direction, the average force applied by a user in this direction was not within one standard deviation from the expert mean (0.184 N – 0.290 N). Since the novice's average force was around half that of the expert, the generated feedback, *"Force in minus Z direction should be 2 times higher"*, was rated as *acceptable* (score 5).

Table 3. Part of the expert evaluation form for stages 1), 5) and 9). Subscript *e* indicates the expert average value (out of five experts) with one standard deviation while *n* indicates the current novice value. The full evaluation form contains all 65 cases and shows all features (force, position, and orientation) considered in the feedback generation mechanism.

Stage	F_x (N)	F_y (N)	F_z (N)	Feedback	Acceptability
1_e	0.103±0.037	0.480±0.047	0.106±0.023	*"Force in minus Y direction should be 3 times higher"*	4
1_n	0.026	0.164	0.091		
5_e	0.040±0.014	0.038±0.019	0.237±0.053	*"Force in minus Z direction should be 2 times higher"*	5
5_n	0.028	0.019	0.129		
9_e	0.064±0.024	0.035±0.019	0.285±0.033	*"Force in minus Z direction should be 2 times higher"*	3
9_n	0.108	0.115	0.159		

For stage 9), however, even though the situation in minus Z direction was almost the same as in stage 5), the feedback (*"Force in minus Z direction should be 2 times higher"*) was rated as *not sure* (score 3). The expert noticed that, during this stage, the force value in X and Y were quite high although they should have been close to zero. There might be two causes for this behavior; either the novice did not know the main direction of the force in this stage (minus Z) or he/she knew but could not control the tool to move in the right direction. The expert suggested giving a tutoring hint such as "*Do you know that minus Z should be the main direction of force in this stage?*" This kind of hint would be especially useful in online training as the system can observe a novice's reaction after the feedback is given. Note that even though we have not yet applied this strategy, the system was capable of detecting the behavior as forces in X and Y (0.108 N and 0.115 N respectively) were also higher than one standard deviation from the expert means.

For stage 1), the generated feedback, *"Force in minus Y direction should be 3 times higher"*, was *close to acceptable* (score 4). The expert commented that a novice could

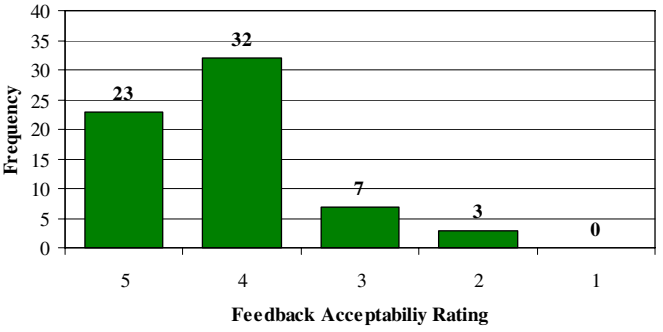


Fig. 8. Distribution of feedback acceptability ratings for 65 generated feedback messages. The average score was 4.154.

accidentally damage a tooth in this stage if he/she tried to applied force too much, therefore, suggested feedback could possibly be only *"2 times higher"* (instead of 3).

The acceptability ratings for all 65 training feedback messages generated by the system are shown in Fig. 8. The average score assigned by the expert for the generated feedback was 4.154 out of 5.

5 Conclusion

In this paper, we describe a mechanism for providing objective feedback on skill assessment using our dental training simulator. After a procedure is done, the simulator is able to classify the performance of a particular operator as novice-level or expert-level based on the force applied, tool position, and tool orientation using a hidden Markov model. Moreover, the simulator can later generate tutoring feedback with quality comparable to the feedback provided by human tutors. Additional tutoring strategies will be explored in the future work. The evaluation results are promising and prove the applicability of the simulator as a supplemental training and performance assessment tool for dental surgical skills.

Acknowledgments. This research was funded by grant NT-B-22-MS-14-50-04 from the National Electronics and Computer Technology Center, Thailand. The authors would like to thank Prabal Khanal for his contribution to the haptics software, and Kan Ouivirach for his HMM implementation. We also would like to thank Gajananan Kugamoorthy for many useful discussions and comments.

References

1. Gorman, P.J., Meier, A.H., Rawn, C., Krummel, T.M.: The future of medical education is no longer blood and guts, it is bits and bytes. Am. J. Surg. 180(5), 353–356 (2000)
2. Roberts, K.E., Bell, R.L., Duffy, A.J.: Evolution of surgical skills training. World J. Gastroenterol. 12(20), 3219–3224 (2006)

3. Kim, L., Hwang, Y., Park, S.H., Ha, S.: Dental training system using multi-modal interface. Computer-Aided Design & Applications 2(5), 591–598 (2005)
4. Yau, H.T., Hsu, C.Y.: Development of a dental training system based on point-based models. Computer-Aided Design & Applications 3(6), 779–787 (2006)
5. Yau, H.T., Tsou, L.S., Tsai, M.J.: Octree-based virtual dental training system with a haptic device. Computer-Aided Design & Applications 3, 415–424 (2006)
6. Wang, D., Zhang, Y., Wang, Y., Lu, P.: Development of dental training system with haptic display. In: 12th IEEE International Workshop on Robot and Human Interactive Communication, pp. 159–164 (2001)
7. Rosen, J., Solazzo, M., Hannaford, B., Sinanan, M.: Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden Markov model. Comput. Aided Surg. 10(1), 49–61 (2002)
8. Lin, H.C., Shafran, I., Yuh, D., Hager, G.D.: Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions. Comput. Aided Surg. 11(5), 220–230 (2006)
9. Rhienmora, P., Haddawy, P., Dailey, M.N., Khanal, P., Suebnukarn, S.: Development of a dental skills training simulator using virtual reality and haptic device. NECTEC Technical Journal 8, 140–170 (2006)
10. Hashizume, M., Shimada, M., Tomikawa, M., Ikeda, Y., Takahashi, I., Abe, R., Koga, F., Gotoh, N., Konishi, K., Maehara, S., Sugimachi, K.: Early experiences of endoscopic procedures in general surgery assisted by a computer-enhanced surgical system. Surg. Endosc. 16(8), 1187–1191 (2002)
11. Rose, D.J.: A multilevel approach to the study of motor control and learning. Allyn & Bacon, Boston (1997)

Voice Pathology Classification by Using Features from High-Speed Videos

Daniel Voigt, Michael Döllinger, Anxiong Yang,
Ulrich Eysholdt, and Jörg Lohscheller

Department of Phoniatrics and Pediatric Audiology,
University Hospital Erlangen, Bohlenplatz 21,
D-91054 Erlangen, Germany
Daniel.Voigt@uk-erlangen.de

Abstract. For the diagnosis of pathological voices it is of particular importance to examine the dynamic properties of the underlying vocal fold (VF) movements occurring at a fundamental frequency of 100–300 Hz. To this end, a patient's laryngeal oscillation patterns are captured with state-of-the-art endoscopic high-speed (HS) camera systems capable of recording 4000 frames/second. To date the clinical analysis of these HS videos is commonly performed in a subjective manner via slow-motion playback. Hence, the resulting diagnoses are inherently error-prone, exhibiting high inter-rater variability. In this paper an objective method for overcoming this drawback is presented which employs a quantitative description and classification approach based on a novel image analysis strategy called Phonovibrography. By extracting the relevant VF movement information from HS videos the spatio-temporal patterns of laryngeal activity are captured using a set of specialized features. As reference for performance, conventional voice analysis features are also computed. The derived features are analyzed with different machine learning (ML) algorithms regarding clinically meaningful classification tasks. The applicability of the approach is demonstrated using a clinical data set comprising individuals with normophonic and paralytic voices. The results indicate that the presented approach holds a lot of promise for providing reliable diagnosis support in the future.

1 Introduction

To differentiate between healthy and pathological VF vibration patterns is a vital part of the clinical diagnosis of voice functioning. This important distinction is commonly made based on the degree of symmetry and regularity of the laryngeal dynamics [1], which can only be assessed by examining the rapidly moving VFs during voice production (phonation). A variety of approaches has been developed for the observation of the underlying vibratory patterns (e.g. [2,3]), but to date endoscopic HS camera systems are the most sophisticated and promising technology for this purpose [4]. In order to clinically assess the HS recordings plenty of experience and time is needed on the part of the physician, as the human eye is much more adapted to the processing of static visual

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 315–324, 2009.

information than to moving images. Consequently, the resulting diagnoses are inherently imprecise and exhibit a rather low inter- and intra-rater reliability. To remedy this weakness, different quantitative analysis approaches have been introduced to facilitate the objective analysis of HS videos [5,6,7]. Yet, these methods lack the ability to analyze the VF oscillation pattern in its entirety.

Phonovibrography, a recently developed visualization technique, is a fast and clinically evaluated method for capturing the whole spatio-temporal pattern of activity along the entire VF length [8]. By extracting both VFs' deflections from HS videos and compactly depicting them in a so-called Phonovibrogram (PVG), for the first time a comprehensive analysis of the underlying 2-d laryngeal dynamics is enabled. Besides being a valuable diagnostic tool [9], a PVG can be used as a starting point for extracting a set of numerical features which describe the characteristic properties of the VF vibration patterns. The resulting feature space can be analyzed in order to build models for automatically distinguishing between normophonic and disordered voices.

In this paper, clinical HS recordings from healthy and pathological female subjects are processed with a novel method for capturing the PVG dynamics. The obtained numerical features are used as input to different ML algorithms with a combined evolutionary optimization strategy to model the underlying pathological concepts. Using these models the data are classified according to clinical decision tasks which are meaningful to the diagnosis of impaired voices. In order to evaluate the performance of the PVG features, another set of conventional glottal features is analyzed as well. With the resulting cross-validated classification accuracies, the different feature sets and ML approaches are compared to each other with respect to their ability to adequately capture VF movement and to support diagnostic decision making.

2 Methods

The following steps are performed to allow for the objective discrimination of the different types of VF vibrations.

2.1 Phonovibrography

By means of an endoscopic HS camera system the VFs' oscillation patterns are digitally recorded during phonation with a frame rate of 4000 fps (frames/second). For this task a conventional 24 fps-camera would be insufficient, as the fundamental frequency of normal speakers approximately ranges from 100–300 Hz. During examination the endoscope is inserted orally, yielding a top view of the larynx with a spatial resolution of 256×256 image points. The recorded HS sequence consisting of several thousands of grayscale images (see. Fig. 1 left) is segmented afterwards using a modified region-growing algorithm [8] detecting the position of the glottal opening formed between the left and the right VF. Thus, for each HS image the positions of both VFs are obtained as a function of time.

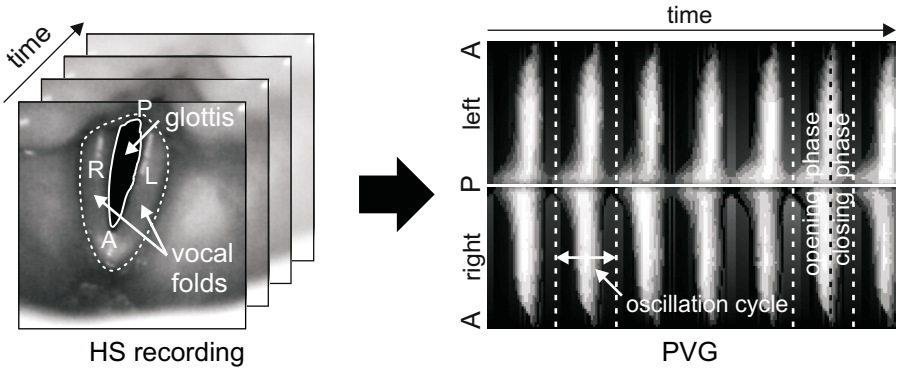


Fig. 1. PVG representation of both VFs' oscillations over time as captured in endoscopic HS recordings (left). The changing deflections yield characteristic PVG patterns (right). The letters "A" and "P" denote the anterior and posterior VF part.

The extracted spatio-temporal information can be conveniently transformed into a 2-d color representation, denoted as PVG [9]. While the deflections of all points of the left VF side are depicted in the upper PVG half, the respective points of the right side are shown in the lower half. The graded PVG color intensities reflect the VFs' changing displacements over time (see Fig. 1 right)—the brighter the color, the farther away the corresponding VF point from its initial position in the middle. Thus, a single PVG row represents the deflections at a specific VF position over time. A single PVG column displays the VF deflection alongside the entire VF length at a certain point in time. By yielding a compact image of the dynamic VF movement, the physician is enabled to quickly gain insight into the complexities of the underlying oscillation patterns. In this manner, clinical evidence for voice pathologies, e.g. the stability and symmetry of the VF dynamics, can easily be assessed. This type of information can hardly be obtained from slow-motion playback of HS videos.

2.2 Feature Extraction

The resulting PVG data matrix is subsequently analyzed in order to quantitatively describe the contained laryngeal movement information. In doing so, it is exploited that the VF vibrations exhibit periodically recurring movement patterns (see Fig. 1 right). Hence, a preliminary step of feature extraction consists in detecting the individual oscillation cycles, which normally involve a distinct opening and closing phase. Using the automatically identified cycle boundaries the continuous PVG is decomposed into smaller logical movement units, from which the corresponding features are derived. To allow for better inter-individual comparability of the features the obtained oscillation cycles are normalized to a uniform size (see Fig. 2a).

An oscillation cycle's underlying spatio-temporal changes are represented through its geometric PVG shape. This information is captured using contour

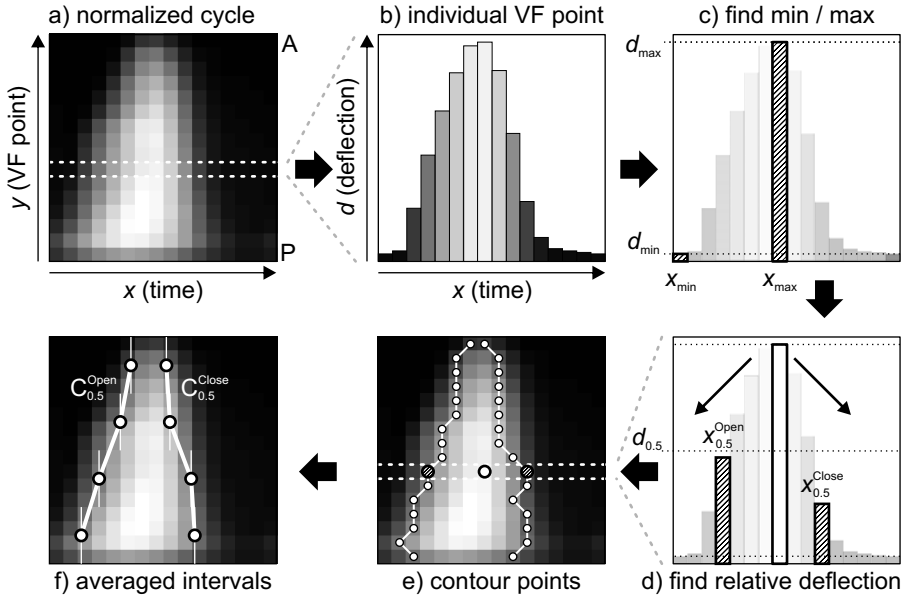


Fig. 2. Single steps of an exemplary PVG contour feature extraction. For each VF point (dashed row) of a 16×16 normalized oscillation cycle (a), the corresponding pair of opening and closing contour points is determined (c–e), and finally, it is averaged over 4 intervals, respectively, yielding a distinct opening and closing contour (f).

lines, which connect deflection states $d(x_i)$ of the VFs in the opening and closing phase of a cycle, respectively (see Fig. 2e), which are equivalent. To this end the point in time $x_{\max}$ is determined which possesses maximum deflection. Starting from $x_{\max}$ for each VF point the two points in time x_h^{Open} and x_h^{Close} are determined, when deflection first falls below a defined deflection state d_h (see Fig. 2d). This deflection threshold results from the computed difference between $d_{\min}$ and $d_{\max}$, denoting minimum and maximum displacement within the considered cycle (see Fig. 2c). Put more formally:

$$d_h = d_{\min} + h \cdot (d_{\max} - d_{\min}), \text{ with } h \in [0, 1] \tag{1}$$

$$\begin{aligned} x_h^{\text{Open}} &:= \max_i(\arg_x (d(x_i) \leq d_h)), \text{ with } x_i < x_{\max} \\ x_h^{\text{Close}} &:= \min_i(\arg_x (d(x_i) \leq d_h)), \text{ with } x_i > x_{\max}. \end{aligned} \tag{2}$$

Thus, for each VF point a pair of contour points is identified, which is characterized by a temporal (x-axis), a longitudinal (y-axis), and a lateral position (color intensity) within the cycle. To aggregate the contour line data, individual intervals are subsumed by averaging. Rather than just describing an isolated VF point the resulting averaged contour points approximate the dynamics of an entire VF section (see Fig. 2f). In this study, the PVG contour lines were extracted

at 5 different contour heights $h = [0.1;\ 0.3;\ 0.5;\ 0.7;\ 0.9]$, consisting of 8 contour intervals each, to compare their suitability for describing VF movements.

In this manner, for each PVG cycle a pair of simplified contour lines $\mathrm{C}_h^{\mathrm{Open_{L,R}}}$ and $\mathrm{C}_h^{\mathrm{Close_{L,R}}}$ is obtained, capturing the dynamics of the left and right VF separately. Since the relative behavior of both VFs is also of diagnostic relevance, the proportions $\mathrm{P}_h^{\mathrm{Open}} = \mathrm{C}_h^{\mathrm{Open_L}}/\mathrm{C}_h^{\mathrm{Open_R}}$ and $\mathrm{P}_h^{\mathrm{Close}} = \mathrm{C}_h^{\mathrm{Close_L}}/\mathrm{C}_h^{\mathrm{Close_R}}$ between corresponding left and right side contours are determined. In addition, the Euclidian distances $\mathrm{D}_h^{\mathrm{Open}} = \|\mathrm{C}_h^{\mathrm{Open_L}} - \mathrm{C}_h^{\mathrm{Open_R}}\|_2$ and $\mathrm{D}_h^{\mathrm{Close}} = \|\mathrm{C}_h^{\mathrm{Close_L}} - \mathrm{C}_h^{\mathrm{Close_R}}\|_2$ are computed as a similarity measure for opposing contour pairs. The temporal properties of VF dynamics are captured by computing the mean and the standard deviation of the features over all oscillation cycles. While the mean captures the VFs' average vibratory behavior, the standard deviation describes the variations from this average pattern in time.

Furthermore, in order to evaluate the performance of the derived PVG contours, conventional features are computed. They describe the dynamic changes of the VFs on the basis of the 1-d glottal signal, which represents the area spanned between the vibrating VFs. As yet, this description approach is standard when it comes to the analysis of voice disorders based on HS recordings. The computed features are open quotient $\mathrm{Q}^{\mathrm{Open}}$, speed quotient $\mathrm{Q}^{\mathrm{Speed}}$, glottal insufficiency $\mathrm{Q}^{\mathrm{Insuff}}$, time periodicity index $\mathrm{I}^{\mathrm{Time}}$, and amplitude periodicity index $\mathrm{I}^{\mathrm{Ampl}}$ [6]. Table 1 sums up the different features used in this study.

Table 1. Different feature sets derived from the HS recordings of the patients' VF movements

Feature sets	Contained features	Underlying signal
$\mathrm{F}_h^{\mathrm{Contour}}$	$\overline{\mathrm{C}}_h^{\mathrm{Open_{L,R}}}$, $\sigma(\mathrm{C}_h^{\mathrm{Open_{L,R}}})$, $\overline{\mathrm{C}}_h^{\mathrm{Close_{L,R}}}$, $\sigma(\mathrm{C}_h^{\mathrm{Close_{L,R}}})$, $\overline{\mathrm{P}}_h^{\mathrm{Open}}$, $\sigma(\mathrm{P}_h^{\mathrm{Open}})$, $\overline{\mathrm{P}}_h^{\mathrm{Close}}$, $\sigma(\mathrm{P}_h^{\mathrm{Close}})$, $\overline{\mathrm{D}}_h^{\mathrm{Open}}$, $\sigma(\mathrm{D}_h^{\mathrm{Open}})$, $\overline{\mathrm{D}}_h^{\mathrm{Close}}$, $\sigma(\mathrm{D}_h^{\mathrm{Close}})$	PVG (2d)
$\mathrm{F}^{\mathrm{Glottal}}$	$\overline{\mathrm{Q}}^{\mathrm{Open}}$, $\sigma(\mathrm{Q}^{\mathrm{Open}})$, $\overline{\mathrm{Q}}^{\mathrm{Speed}}$, $\sigma(\mathrm{Q}^{\mathrm{Speed}})$, $\overline{\mathrm{Q}}^{\mathrm{Insuff}}$, $\sigma(\mathrm{Q}^{\mathrm{Insuff}})$, $\overline{\mathrm{I}}^{\mathrm{Time}}$, $\sigma(\mathrm{I}^{\mathrm{Time}})$, $\overline{\mathrm{I}}^{\mathrm{Ampl}}$, $\sigma(\mathrm{I}^{\mathrm{Ampl}})$	Glottis (1d)

2.3 Machine Learning Setting

The extracted feature sets were subsequently analyzed with the following ML methods [10] with regard to clinically relevant diagnostic tasks:

- k-nearest neighbor algorithm (k-NN) with Euclidian distance measure,
- C4.5 decision tree with information gain splitting criterion,
- feed-forward artificial neural network (ANN) with backpropagation,
- support vector machine (SVM) with 1^{st}, 2^{nd}, 3^{rd}-order polynomials and radial basis function (RBF) kernels.

As most of these learning algorithms have free parameters that can significantly influence the performance of the built models, an evolutionary parameter optimization approach was employed in this study. The applied heuristic search method is the evolution strategy of Schwefel and Rechenberg [11], which finds an appropriate ML parameter combination for the problem at hand. By means of selection, recombination, and mutation steps a population of potential solutions is evaluated with respect to its fitness and then accordingly adapted. The following ML parameters are optimized: the neighborhood parameter k of k-NN; the learning rate and momentum term of ANN; and the cost parameter C for the different types of SVMs. For SVM with RBF kernel the width γ was additionally optimized. The rest of the ML parameters remain fixed at the respective standard values.

A stratified 10-fold cross-validation method (10xCV) is used for the evaluation of the learned models [12]. Splitting the data into a distinct training and test set was not advisable due to the amount of clinically available data. To reduce random outcome caused by splitting the data into individual folds, the 10xCV is repeated three times with differing random seed values for each classification task and subsequently averaging over the individual results. As classification task T_3 (see next subsection) includes a merged and undersampled class, it is performed repeatedly and averaged with different class configurations as well. Hence, a reliable estimate of classification accuracy is obtained.

2.4 Data

The efficiency of the presented PVG analysis approach was evaluated with the aid of a collective of 45 female probands. To obtain a gold standard for classification, they were thoroughly examined and diagnosed by experienced physicians. A third of the test persons had a healthy voice with no clinical signs of voice disorders. The remaining cases exhibited unilateral paralytic vibratory behavior, i.e. one VF side's motivity was impaired due to neural damage. The patients' diagnoses consisted in equal shares of left-sided and right-sided pareses. Unlike organic voice disorders (e.g. nodule, polyp, edema), which can be identified quite reliably by analyzing laryngeal still images [13], an appropriate paralytic diagnosis can only be made with respect to the dynamic properties of the VFs. For all women included in this study, the laryngeal dynamics were recorded and segmented for a sequence of 500 frames yielding the corresponding PVG and glottis signal. Using the PVG contour features and conventional glottal features (see Tab. 1) the following classification tasks were investigated:

- $\mathbf{T}_1$: Healthy vs. Paresis_L,
- $\mathbf{T}_2$: Healthy vs. Paresis_R,
- $\mathbf{T}_3$: Healthy vs. Pathological (representing a merged and undersampled class containing randomly selected cases from Paresis_L and Paresis_R),
- $\mathbf{T}_4$: Paresis_L vs. Paresis_R,
- $\mathbf{T}_5$: Healthy vs. Paresis_L vs. Paresis_R.

3 Results

In order to evaluate the obtained classification accuracies with respect to the different learning approaches and feature sets, the following results are presented:

a) classification accuracies of the employed ML algorithms averaged over all feature sets and decision tasks $T_{1,2,4}$ (see Fig. 3a),
b) the classification accuracies of the decision tasks $T_{1...5}$ averaged over contour feature sets $F^{Contour}_{0.1...0.9}$ for SVM with linear kernel (see Fig. 3b),
c) and the classification accuracies of the feature sets $F^{Contour}_{0.1...0.9}$ and $F^{Glottal}$ averaged over decision tasks $T_{1,2,4}$ for SVM with linear kernel (see Fig. 3c).

Since best average classification results were achieved by SVM with linear kernel (see Fig. 3a), Fig. 3b and 3c exclusively focus on the results of this particular ML algorithm. The omission of 2-class problem "Healthy vs. Pathological" (T_3) in the results presented in Fig. 3a and 3c is due to the fact that this decision task is a partial recombination of T_1 and T_2, and thus, its inclusion would yield an overestimation of classification accuracy. As the results of the 3-class problem "Healthy vs. $Paresis_L$ vs. $Paresis_R$" (T_5) cannot be readily compared to the outcomes of the 2-class problems, it is also left out from evaluation.

4 Discussion

From the average performance of the employed ML methods shown in Fig. 3a it can be seen that SVMs with 1st-order polynomial kernel yield the best models for solving the examined clinical decision problems (ca. 85%). By increasing the polynomial degree of the SVM kernel a gradual decline in classification accuracy is determined. In average, C4.5, ANN and SVM with 3rd-order polynomial and RBF kernel perform equally well (around 80%). The worst classification results of this study are obtained using the k-NN algorithm (ca. 68%).

The presented PVG classification method is capable of distinguishing reliably between individuals with a healthy and a disordered voice. The corresponding accuracies of $T_{1...3}$ shown in the classification results in Fig. 3b achieve approximately 90%. Since the considered decision problems "Healthy vs. Pathological", "Healthy vs. $Paresis_L$", and "Healthy vs. $Paresis_R$" are the diagnostic tasks which are most relevant clinically, the approach as a whole can be assessed as being successful. The advantages of the new PVG contours over the conventional glottal description approach can be seen from the results of the individual feature sets in Fig. 3c. The difference between the averaged classification accuracies of the best and the worst performing feature set ($F^{Contour}_{0.5}$ vs. $F^{Glottal}$) amounts to 11%. Hence, the additional amount of information gained through the PVG-based 2-d contour features significantly increases discrimination performance.

By inspecting the results of the individual contour features in Fig. 3c a distinct tendency towards mid-range contours ($F^{Contour}_{0.3...0.7}$) can be noticed. While the classification of the mid-range PVG contours yields comparable results

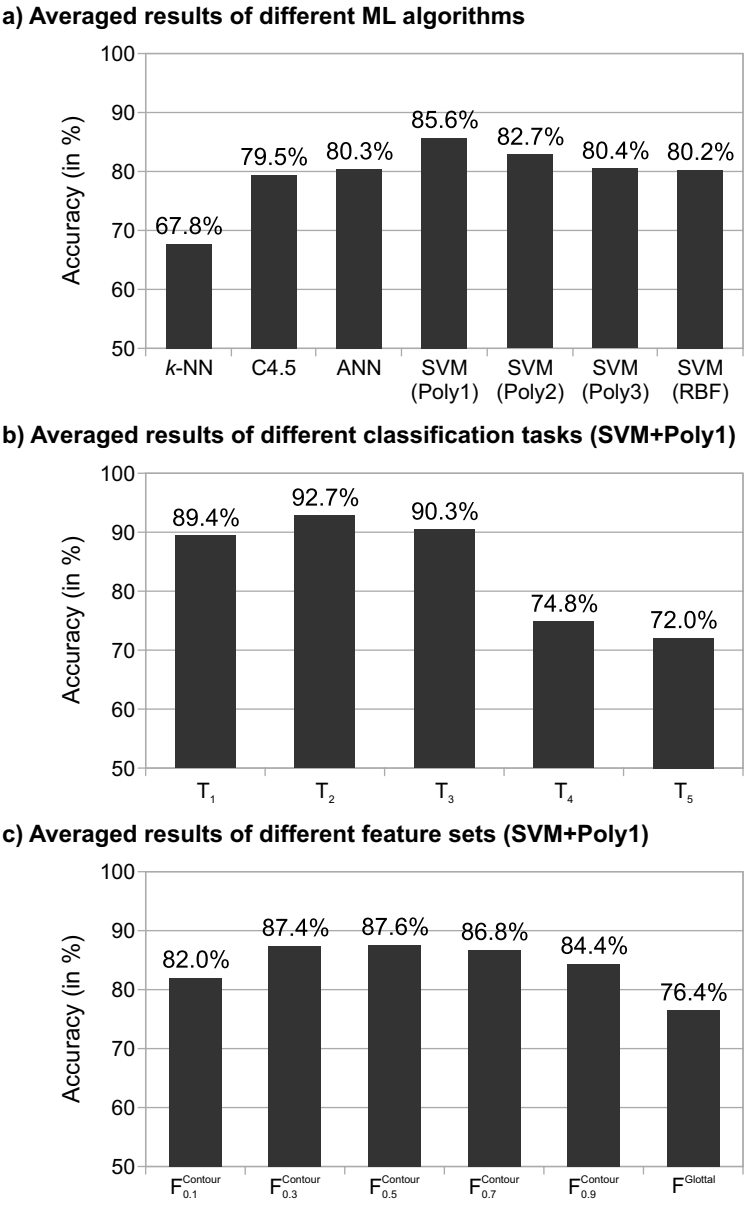


Fig. 3. Results showing the average performance of the employed ML methods (a), the considered classification tasks (b), and the analyzed feature sets (c)

(ca. 87%), the accuracy obtained for the outer contour feature sets ($F_{0.1}^{Contour}$ and $F_{0.9}^{Contour}$) decreases. This can be ascribed to the fact that the characteristic vibratory patterns in the mid-range VF deflection state are more discriminative,

and as a result, a more reliable discrimination is achieved. Thus, when the VFs approach minimum or maximum deflection, the PVG contours of the healthy and pathological vibrations converge, resulting in a reduced classification accuracy.

The assignment of the pareses to the affected VF side (T_4) yields a classification accuracy of ca. 75% (see Fig. 3b). According to that, distinct lateral classification using the derived features performed relatively poorly compared to the results of diagnostic tasks $T_{1...3}$ mentioned above. This issue is also reflected in the results of T_5 (ca. 72%). But with respect to the baseline classification accuracy of 33% for balanced 3-class problems, performance of T_5 must be considered better than the results of 2-class problem T_4. The underlying results of T_5 reveal that most misclassifications occur between the two classes Paresis_L and Paresis_R. The reason for this aggravated lateral distinction may be seen in the fact, that a paralytic impairment does not manifest itself in the deflection symmetry of the affected VF and the resulting PVG patterns as suspected. In order to understand this more thoroughly, the correlation between pathology and vibratory outcome has to be examined using more VF oscillation data. However, it can be assumed that by extending the lateral PVG feature descriptions and refining the computational parameters of the feature extraction process (e.g. amount of contour points, interval lengths) an improvement of the overall classification accuracy can be achieved.

5 Summary

In this paper a novel method for the description and classification of paralytic voice disorders was presented. For a collective of normophonic and pathological female speakers the VFs were recorded during phonation by using endoscopic HS technology. These HS videos were analyzed with a recently introduced image processing technique and the obtained VF segmentation data were transformed into a 2-d PVG representation. The spatio-temporal patterns contained in the PVGs were captured with specialized contour features at different heights of the oscillation cycles. Conventional glottal parameters served as a reference for performance. It was shown that the PVG features exceed the glottal features with regard to their ability to describe the underlying vibratory processes. Moreover, with the presented PVG description approach a reliable distinction between healthy and pathological VF movement patterns was achieved. This capability is of great diagnostic value in the objective assessment and identification of voice disorders. Nevertheless, possible starting points for improving particular aspects of the PVG description approach were also identified in this study. For the advancement of the method further studies will be conducted in the future involving an extensive collection of normophonic and pathological data. In general, the presented voice analysis approach shows a lot of potential to support the physician's clinical decision making by providing a sound objective basis.

References

1. Dejonckere, P., Bradley, P., Clemente, P., Cornut, G., Crevier-Buchman, L., Friedrich, G., Heyning, P.V.D., Remacle, M., Woisard, V.: Committee on Phoniatrics of the European Laryngological Society (ELS): A basic protocol for functional assessment of voice pathology. Eur. Arch. Otorhinolaryngol. 258, 77–82 (2001)
2. Raes, J., Lebrun, Y., Clement, P.: Videostroboscopy of the larynx. Acta Otorhinolaryngol. Belg. 40, 421–425 (1986)
3. Švec, J., Schutte, H.: Videokymography: high-speed line scanning of vocal fold vibration. J. Voice 10(2), 201–205 (1996)
4. Deliyski, D., Petrushev, P., Bonilha, H., Gerlach, T., Martin-Harris, B., Hillman, R.: Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution. Folia Phoniatr Logop 60(1), 33–44 (2008)
5. Švec, J., Sram, F., Schutte, H.: Videokymography in voice disorders: what to look for? Ann. Otol. Rhinol. Laryngol. 116(3), 172–180 (2007)
6. Qiu, Q., Schutte, H., Gu, L., Yu, Q.: An automatic method to quantify the vibration properties of human vocal folds via videokymography. Folia Phoniatr Logop 55(3), 128–136 (2003)
7. Mergell, P., Herzel, H., Titze, I.: Irregular vocal-fold vibration–high-speed observation and modeling. J. Acoust. Soc. Am. 108(6), 2996–3002 (2000)
8. Lohscheller, J., Eysholdt, U., Toy, H., Döllinger, M.: Phonovibrography: mapping high-speed movies of vocal fold vibrations into 2-d diagrams for visualizing and analyzing the underlying laryngeal dynamics. IEEE Trans. Med. Imaging 27(3), 300–309 (2008)
9. Lohscheller, J., Toy, H., Rosanowski, F., Eysholdt, U., Döllinger, M.: Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. Med. Image Anal. 11(4), 400–413 (2007)
10. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. John Wiley & Sons, Chichester (2001)
11. Beyer, H., Schwefel, H.: Evolution strategies - a comprehensive introduction. Natural Computing 1, 3–52 (2002)
12. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI, pp. 1137–1145 (1995)
13. Verikas, A., Gelzinis, A., Bacauskiene, M., Uloza, V.: Towards a computer-aided diagnosis system for vocal cord diseases. Artif. Intell. Med. 36(1), 71–84 (2006)

Analysis of EEG Epileptic Signals with Rough Sets and Support Vector Machines

Joo-Heon Shin[1], Dave Smith[2], Roman Swiniarski[3], F. Edward Dudek[5], Andrew White[2], Kevin Staley[6], and Krzysztof J. Cios[1,4]

[1] Virginia Commonwealth University,
kcios@vcu.edu
[2] University of Colorado Denver
[3] San Diego State University
[4] IITiS Polish Academy of Sciences
[5] University of Utah
[6] Massachusetts General Hospital and Harvard Medical School

Abstract. Epilepsy is a common chronic neurological disorder that impacts over 1% of the population. Animal models are used to better understand epilepsy, particularly the mechanisms and the basis for better antiepileptic therapies. For animal studies, the ability to identify accurately seizures in electroencephalographic (EEG) recordings is critical, and the use of computational tools is likely to play an important role. Electrical recording electrodes were implanted in rats before kainate-induced status epilepticus (one in each hippocampus and one on the surface of the cortex), and EEG data were collected with radio-telemetry. Several data mining methods, such as wavelets, FFTs, and neural networks, were used to develop algorithms for detecting seizures. Rough sets, which were used as an additional feature selection technique in addition to the Daubechies wavelets and the FFTs, were also used in the detection algorithm. Compared with the seizure-at-once method by using the RBF neural network classifier used earlier on the same data [12], the new method achieved higher recognition rates (i.e., 91%). Furthermore, when the entire dataset was used, as compared to only 50% used earlier, preprocessing using wavelets, Principal Component Analysis, and rough sets in concert with Support Vector Machines resulted in accuracy of 94% in identifying epileptic seizures.

Keywords: epileptic seizures detection, medical signal processing, rough sets.

1 Introduction

Epilepsy is a neurological disorder that impacts approximately two million Americans [17]. Animal models [3, 4, 13, 21] are used to better understand the mechanisms responsible for epilepsy and to develop new antiepileptic drugs. In epilepsy research, it is critical to determine the occurrence of seizures, particularly their

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 325–334, 2009.

frequency, severity and pattern. In addition to studying their behavior, electroencephalographic (EEG) recordings and associated seizures must be analyzed quantitatively.

Prior attempts at EEG analysis have focused on the identification of three specific parameters: (a) seizures [7, 10], (b) a pre-ictal state that occurs shortly before a seizure [9, 14], and (c) inter-ictal spikes occurring between seizures representing high-amplitude abnormal discharges [5, 23]. In this paper, we focus on algorithms for the detection of seizures. There is considerable literature in this field and multiple techniques have been tried, involving either direct evaluation of the signal or signal transformations (and/or filtering) [8]. Further, these techniques can also involve the use of neural networks to introduce a nonlinear detection scheme [6]. More recently, feature extraction using wavelet analysis has been combined with neural networks to improve detection sensitivity and specificity [1, 15, 18].

In this study we use Support Vector Machines (SVMs) [19] on the EEG data preprocessed with wavelets, FFT, and rough sets. Wavelet analysis localizes in time harmonic characteristics of the EEG signal while rough sets help the SVM classifier to achieve higher accuracy. The system is tested using 5- and 10-fold cross validation.

2 Methods

2.1 Dataset

In these experiments, the EEG was monitored with a radio-telemetry system [20, 22] from bilateral hippocampal electrodes and a single surface electrode connected to an implanted transmitter. The signal was transmitted to a receiver plate and then to a computer where it was stored in the file system. It was later transferred to DVD for further analysis on other computers. The sampling rate was 250 Hz. The behavioral data was recorded using 24-h video monitoring. The animal model involved kainate-induced status epilepticus, and has been used to investigate temporal lobe epilepsy [3]. In this model, Sprague-Dawley rats are given a sufficient quantity of kainic acid to produce at least 10 seizures per hour for a period of 3 h, which creates brain lesions in the hippocampus and other structures, resulting in the development of epileptic seizures after a latent period lasting from days to weeks [21]. The timing of these seizures is generally unpredictable; and therefore, continuous monitoring is needed to ensure that they are all detected. The use of EEG is essential as some seizures cannot be detected by observation alone (i.e., they have no overt convulsive behaviors, such as forelimb clonus), and so these non-convulsive seizures would be missed if only the video was reviewed.

The quantity of data produced in this experiment is huge. Each rat is monitored for approximately 3-5 months and there are three channels for each rat. At the sampling rate of 250 Hz, this results in approximately 5.8 GB per rat. For the current study there were 9 rats (4 control and 5 kainic acid rats (2 of the rats did not survive long enough to generate sufficient data)), resulting in over 50 GB

of data to analyze. Using this raw data and a series of annotations describing the time of day and duration of seizures, the filtered dataset was constructed [12]. This dataset included seizure information from the epileptic rats, as well as random samples of EEG data from the normal rats. Additionally, some relatively normal (i.e., interictal, between seizures) EEG data was included from the epileptic rats to ensure that the analysis was properly balanced. The problem was to detect whether or not a given EEG signal showed seizure activity. The EEG signal of the abnormal (epileptic) rats is comprised primarily of time periods without seizures; seizure activity makes up only a small temporal fraction of the daily EEG signal, even in rats with a high seizure frequency. We therefore included EEG data without seizures from epileptic rats to train the classifier. We used 3106 interictal and 2356 ictal (seizure) samples.

Each sample within the filtered dataset contained 230 sec of three-channel data, which yielded 172,500 data points per record. This amount of data is difficult to review by a human; therefore, computational techniques are required since the entire data set was 9 animals at 24h/d for 100 days. Additionally, significant artifact and noise exist within the signal, making it more difficult to differentiate between this and actual seizures. Note that the reliability of seizure detection is imperfect even when the EEG is manually reviewed [24].

2.2 Design Procedure

The size of the filtered dataset is large, which calls for further processing to reduce processing time. To use the EEG data efficiently we thus heuristically reduce each record by segmenting it into 90 equal partitions, and each partition is then transformed into a single number by using transforming functions (described next) on the entire partition and averaging the output. This step is to extract the reduced record as close as the original signal. We used the following functions for each partition: Fast Fourier Transform (FFT), Daubechies wavelet decomposition (db2) with detail and approximation coefficients, and mean of the absolute values. This process yielded 90 data points per sample, or 30 per channel (see Fig. 1).

Although the data are reduced, an additional feature selection step is performed to deal with artifacts and noise. For that purpose we use rough sets after performing Principal Component Analysis (PCA) projection. Although the data

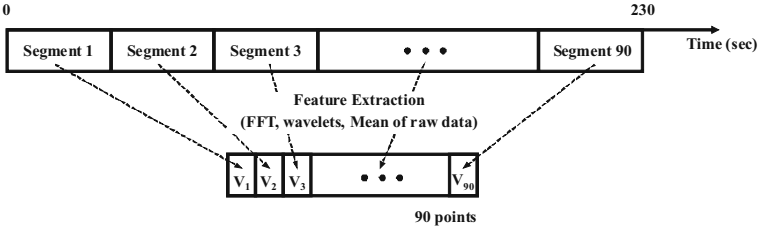


Fig. 1. Condensed record of 90 data points

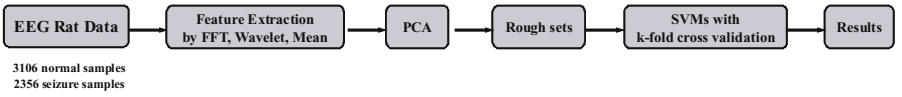


Fig. 2. High-level view of the overall approach

was collected on nine rats rough sets were shown to perform well even on such small data [2]. The vectors resulting from rough sets processing constitute input to the SVM classifier. Fig. 2 shows the sequence in which the methods are used for analyzing the data.

2.3 Feature Extraction

The key feature extraction techniques for time-series signals (like EEG) are the Fast Fourier Transform (FFT) and wavelet analysis [8, 15, 18]. The FFT converts temporal data into a frequency domain but is not able to locate position of frequency components in the original temporal signal. Wavelets, on the other hand, are able to show both the frequency components and their location in the original data [8, 12, 16]. However, choosing the most suitable mother wavelet for a given signal is a nontrivial problem. We use Daubechies wavelet db2 because it is a suitable basis for condensed record of 90 data points. We also use db2 with a number of different levels (from one to four) in multi-resolution analysis to compress and localize signals in time and frequency.

For the FFT, the power spectrum patterns after decomposition were formed to extract periodic frequency of the signal. Also, the approximation and detail coefficients of wavelet transform are treated separately to determine how they affect detection of seizures.

2.4 Feature Selection

To design EEG classifiers we further reduce high dimensional Daubechies or FFT patterns by selecting the most discernible features. This is achieved by first using *Principal Component Analysis* (PCA) feature extraction technique and then the feature selection *rough sets* technique.

Rough Sets. The rough sets theory was proposed by Pawlak [11] as a new tool for dealing with vague concepts, as an alternative to the one afforded by fuzzy sets. Rough sets were shown to significantly reduce the decision table for nondeterministic cases and pattern dimensionality by reduct selection [16]. The idea of rough sets consists of the lower approximation, which contains all objects that can be classified as certainly belonging to the concept, and the upper approximation, which contains all objects that cannot be classified categorically as not belonging to the concept [2].

Rough Sets with PCA. Let us assume that a wavelet transform or FFT has been performed on the segment of the EEG signal. The resulting wavelet/FFT

pattern is still highly dimensional. We thus first use PCA which decorrelates transformed patterns and allows for a large reduction in dimensionality. PCA, however, does not provide information about PCA pattern elements that are best for classifier design. Second, we use rough sets for selecting specific features from PCA patterns. Specifically, computation of a reduct is used for selecting some of the principal components constituting that reduct. The selected principal components describe all concepts in the data and the final pattern can be formed from the reduced PCA patterns based on the selected reduct. The pseudocode of the algorithm, after [16], is given below.

Given: A N-case data set T containing PCA patterns $\mathbf{x} = \mathbf{x}_{f,pca,red} \in \mathbb{R}^{n_{f,pca,red}}$, where f belongs to Daubechies db2 approximation and detailed coefficients, FFT power spectrum, or real-valued attributes, labeled by l classes such that $\{(\mathbf{x}^1, c^1_{target}), \cdots, (\mathbf{x}^N, c^N_{target})\}$ with $target \in \{1, \cdots, l\}$:

1) From the original labeled data set T form a pattern as $N \times n$ data pattern matrix $\mathbf{X}$.
2) Find the $m \times n$ reduced optimal Karhunen-Loéve transformation matrix $\mathbf{W}_{KLT}$ formed with the first m eigenvectors of the data covariance matrix, ordered in decreasing order of the corresponding eigenvalues.
3) Transform original patterns from $\mathbf{X}$ into reduced m-dimensional feature vectors in the principal component space by the formula $\mathbf{Y} = \mathbf{X}\mathbf{W}_{KLT}$.
4) Discretize the patterns in $\mathbf{Y}$, and store them in matrix $\mathbf{Y}_d$.
5) Form the decision table DT based on the patterns from matrix $\mathbf{Y}_d$ with the corresponding classes from the original data set T.
6) Compute a selected relative reduct from the decision table DT, and treat it as a selected set of features $\mathbf{X}_{feature,reduct}$ describing all concepts in DT.
7) Compose final (reduced) real-valued attribute decision table $DT_{final,reduced}$ containing those columns from the projected discrete matrix $\mathbf{Y}_d$, which corresponds to the selected feature set $\mathbf{X}_{feature,reduct}$, based on the selected relative reduct. Label patterns by corresponding classes from the original data set T.

Result: Reduced final data set:

$$T_{final} = \{(\mathbf{x}^1_{feature,reduct}, c^1_{target}), \cdots, (\mathbf{x}^N_{feature,reduct}, c^N_{target})\} \tag{1}$$

This set of features is used in the SVM classifier design.

2.5 Classification

Once we have selected a reduced set of features, we design an SVM classifier [16, 19]. SVMs are easily generalized by adopting kernel techniques. The SVM classifier is formed as

$$f(\mathbf{x}) = \sum_{i=1}^{N} y^i \lambda^i K(\mathbf{x}^i, \mathbf{x}) + b, \tag{2}$$

where $K(\mathbf{x}^i, \mathbf{x}) = \sum_i \phi_i(\mathbf{x}^i)\phi_i(\mathbf{x})$ is the kernel function that is symmetric and positive semi-definite and the λ^i are the *Lagrange multipliers*, respectively. This kernel substitution transforms input vectors into a high dimensional feature space (Hilbert space), and then a separating hyperplane is found with the largest margin for the vectors in the training data set. In this work we use a *Radial Basis Function kernel* defined by

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|}{2\sigma^2}\right), \tag{3}$$

with several values of the width parameter σ (as σ becomes smaller the decision boundary becomes more complex).

3 Results

We use the EEG data reported in [12] which consists of two types of labeled data: interictal and ictal (i.e., seizure). To obtain time-frequency localization characteristics we use Daubechies (db2) wavelet. Since higher compression rates can loose some important information, only Level-1, Level-2, Level-3 and Level-4 Daubechies (db2) wavelet decompositions are used. The power spectrum of the FFT and the mean of the raw data are also used as patterns. Fig. 3 shows examples of interictal and ictal (seizure) condensed records after extracting features by the FFT, the Level-2 Daubechies wavelet decomposition and the mean, using technique shown in Fig. 1.

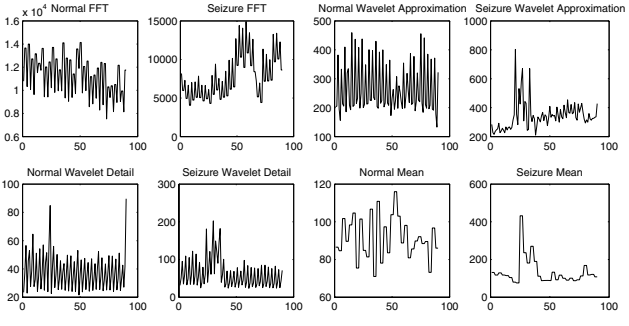


Fig. 3. Examples of seizure patterns after feature extraction

Next, we use PCA to obtain principal components from each condensed record and eliminate those components that contribute less than 0.06% to the total variation in the data. It should be noted that the PCA does not guarantee the most informative patterns for classification purposes so in order to select such patterns for classifier design we use rough sets for selecting the most discriminatory features in the PCA space. Rough sets work on discrete data so we divide each attribute into evenly spaced bins. We used from 5 to 10 bins and have chosen 7

because of its good performance. Then, the discretized data is used to find all relative reducts using pseudocode shown in Section 2.4. Since there can be several reducts we randomly select one of them as the minimal reduct, and it forms the final pattern. This selection step causes different length of vector, corresponding to a reduct.

SVMs are capable of handling class imbalance in an internal manner, so different cost parameters C_+ (for positive class) and C_- (for negative class) are used [from 1 to 10]. For the RBF kernel we also try several values for the width parameter, σ, from the range of $[0.1, 1]$. By using combinations of the cost parameters and kernel widths we end up with a total number of more than 100 classifiers. For a fair comparison of classifiers we selected results based on the setting $C_+ = C_- = 7$ and $\sigma = 0.5$. The complexity of an SVM classifier is related to the number of support vectors that are used in kernel computation. It is also known that the proportion of support vectors (sv) to all training vectors varies and is related to performance error [19]. Therefore, we show in Tables 2 and 3 the proportion of support vectors used in SVM classifiers, for each simulation, how the ratio of support vectors to all training vectors affects the classification accuracy.

Tables 1, 2 and 3 show performance of the classifiers using $\text{Accuracy}(\%) = 100 \times \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$, $\text{Sensitivity}(\%) = 100 \times \frac{\text{TP}}{\text{TP}+\text{FN}}$ and $\text{Specificity}(\%) = 100 \times \frac{\text{TN}}{\text{TN}+\text{FP}}$, where TP is true positive, TN true negatives, FP false positives, and FN false negatives.

Tables 1 and 2 show results using 5-fold cross-validation (to compare with previously reported results), while Table 3 shows results using both 5- and 10-fold cross-validation. The part of all training vectors that were used as support vectors in the SVM classifiers is calculated for each combination.

Table 1. Results reported in [12] with RBF classifier using 5-fold cross validation on the 50% of the data

Feature set	Number of neurons	Sensitivity	Specificity	Accuracy
Wavelet det.	20	77.30	93.30	86.38
Wavelet det.	200	83.20	94.10	89.40
Wavelet app.	20	79.90	83.50	81.97
Wavelet app.	200	82.00	83.20	82.67
FFT	20	79.50	89.60	85.27
FFT	200	83.60	90.30	87.45
Mean raw	20	74.10	80.50	77.77
Mean raw	200	74.50	87.90	82.11

Table 2 shows results of different classifiers, using a new way of preprocessing the data, on the data used in previous work [12]. Notice that the highest accuracy in Table 1 was obtained with the wavelet detail coefficients, while Table 2 shows highest accuracy with the FFT features. Improvement in accuracy on the 50% of data is 5% over the one reported in [12]; however, the sensitivity and specificity values are more balanced.

Table 2. SVM results using 5-fold cross validation on the 50% of the data after PCA and rough sets dimensionality reduction

Feature set	Number of sv	Sensitivity	Specificity	Accuracy
Wavelet det.	4331 (45.85)	89.27	92.93	91.35
Wavelet app.	4701 (49.77)	88.38	84.23	86.02
FFT	4373 (45.24)	87.00	94.57	91.81
Mean raw	5721 (60.57)	82.48	83.27	82.93

Table 3. Results using SVM on the entire dataset after PCA and rough sets dimensionality reduction; wav., det., and app. stand for: wavelets, detailed, and approximation coefficients, while # of sv (%) is the number of support vectors and its % proportion.

	5-fold cross-validation				10-fold cross-validation			
Feature set	# of sv (%)	Sen.	Spe.	Acc.	# of sv (%)	Sen.	Spe.	Acc.
L1 wav. det.	7805 (35.72)	93.42	92.91	93.13	17103 (34.74)	94.10	93.14	93.55
L1 wav. app.	9288 (42.50)	90.32	87.05	88.46	20574 (41.79)	90.36	87.31	88.63
L2 wav. det.	7407 (33.89)	92.86	94.55	93.83	16332 (33.17)	93.03	94.97	94.14
L2 wav. app.	9430 (43.15)	90.40	86.73	88.31	20811 (42.27)	90.28	87.31	88.59
L3 wav. det.	8004 (36.63)	92.14	94.81	93.66	17520 (35.58)	92.40	95.00	93.88
L3 wav. app.	9558 (43.74)	89.43	86.41	87.71	21118 (42.89)	89.81	86.54	87.95
L4 wav. det.	8332 (38.13)	91.42	93.36	92.53	18277 (37.12)	91.42	93.75	92.74
L4 wav. app.	10118 (46.30)	88.53	85.80	86.98	22440 (45.58)	88.87	86.02	87.25
FFT	9213 (42.16)	87.39	88.66	88.11	20480 (41.60)	87.01	88.44	87.28
Mean raw	11828 (54.13)	81.45	85.83	83.94	26163 (53.14)	81.15	86.18	84.01

To test the capability of our approach we also analyzed the entire data set (see Table 3). Wavelet detail coefficients result in much higher accuracy and required less than 40% support vectors, a significant improvement in efficiency. Note that using the FFT feature set on the entire data was not as effective as using the wavelet detail coefficient features. Table 3 shows less than 5% difference between sensitivities and specificities.

4 Conclusions

For animal studies, the ability to accurately identify seizures from EEG recordings using computational tools alone is highly desirable. This can be employed to reduce the human analysis effort for identifying seizures, decreasing the likelihood of missed seizures resulting from human fatigue, and reducing the cost of this type of translational research. Although the current work involved rat EEG recordings, the methods developed in this work can also be employed for seizure detection in humans.

Recent advances in the potential ability to abort human seizures (with administration of electrical stimulation or focal injection of pharmacologic agents) have made the accurate assessment of EEG state (i.e., the presence of seizures or inter-ictal spikes) extremely important. These tools could potentially be used to allow a sufficiently high degree of accuracy in seizure detection, such that they could be programmed into an implantable device that would automatically deliver the stimulation or drug and stop an ongoing seizure.

The novelty presented in this paper is the use of rough sets to reduce dimensionality and use of accurate and efficient SVM classifier. Rough sets are used as an additional feature selection technique after Daubechies wavelets and FFT. Our new method enabled improvement the overall accuracy significantly as compared with [12]. Furthermore, when the entire dataset was used, as compared to the 50% data used before [12], we were able to reach an accuracy of about 94% in identifying epileptic seizures.

The limitations of the method are that some characteristics of original signals can be lost when constructing 90 averaged feature vectors. It would be interesting to use feature extraction techniques on raw data without segmentation. Discretization of the data prior to using rough sets could have impacted calculation of relative reducts. Selecting the optimal reduct out of many can be also an interesting project but is beyond the scope of this work.

References

[1] Alkan, A., Koklukaya, E., Subasi, A.: Automatic seizure detection in EEG using logistic regression and artificial neural network. J. Neurosci. Methods. 148(2), 167–176 (2005)

[2] Cios, K.J., Pedrycz, W., Swiniarski, R., Kurgan, L.: Data Mining: A Knowledge Discovery Approach. Springer, Heidelberg (2007)

[3] Dudek, F.E., Clark, S., Williams, P.A., Grabenstatter, H.L.: Kainate-induced status epilepticus: A chronic model of acquired epilepsy. In: Pitkänen, A., Schwartzkroin, P.A., Moshé, S.L. (eds.) Models of Seizures and Epilepsy, ch. 34, pp. 415–432. Elsevier Academic Press, Amsterdam (2006)

[4] Dudek, F.E., Staley, K.J., Sutula, T.P.: The Search for Animal Models of Epileptogenesis and Pharmacoresistance: Are There Biologic Barriers to Simple Validation Strategies? Epilepsia 43(11), 1275–1277 (2002)

[5] Dzhala, V.I., Talos, D.M., Sdrulla, D.A., Brumback, A.C., Mathews, G.C., Benke, T.A., Delpire, E., Jensen, F.E., Staley, K.J.: NKCC1 transporter facilitates seizures in the developing brain. Nat. Med. 11, 1205–1213 (2005)

[6] Gabor, A.: Seizure detection using a self-organizing neural network: validation and comparison with other detection strategies. Electroencephalogr Clin Neurophysiol. 107(1), 27–32 (1998)
[7] Gotman, J.: Automatic seizure detection: improvements and evaluation. Electroencephalogr Clin Neurophysiol. 76(4), 317–324 (1990)
[8] Khan, Y., Gotman, J.: Wavelet based automatic seizure detection in intracerebral electroencephalogram. Clin Neurophysiol. 114(5), 898–908 (2003)
[9] Lehnertz, K., Litt, B.: The first international collaborative workshop on seizure prediction: summary and data description. Clin Neurophysiol. 116(3), 493–505 (2005)
[10] Murro, A., King, D., Smith, J., Gallagher, B., Flanigin, H., Meador, K.: Computerized seizure detection of complex partial seizures. Electroencephalogr Clin Neurophysiol. 79(4), 330–333 (1991)
[11] Pawlak, Z.: Rough sets: Theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht (1991)
[12] Schuyler, R., White, A., Staley, K., Cios, K.: Identification of Ictal and Pre-Ictal States using RBF Networks with Wavelet-Decomposed EEG. IEEE EMB 26(2), 86–93 (2007)
[13] Stables, J.P., Bertram, E., Dudek, F.E., Holmes, G., Mathern, G., Pitkanen, A., White, H.S.: Therapy discovery for pharmacoresistant epilepsy and for disease-modifying therapeutics: Summary of the NIH/NINDS/AES Models II Workshop. Epilepsia 44, 1472–1478 (2003)
[14] Staley, K.J., Dudek, F.E.: Interictal Spikes and Epileptogenesis. Epilepsy Currents 6(6), 199–202 (2006)
[15] Subasi, A.: Application of adaptive neurofuzzy inference system for epileptic seizure detection using wavelet feature extraction. Comput. Biol. Med. 37(2), 227–244 (2007)
[16] Swiniarski, R., Shin, J.: Classification of Mammograms Using 2D Haar Wavelet, Rough Sets and Support Vector Machines. In: Proceedings of the International Conference on Data Mining in Las Vegas, pp. 65–70 (2005)
[17] Theodore, W.H., Spencer, S.S., Wiebe, S., Langfitt, J.T., Ali, A., Shafer, P.O., Berg, A.T., Vickrey, B.G.: Epilepsy in North America: A Report Prepared under the Auspices of the Global Campaign against Epilepsy, the International Bureau for Epilepsy, the International League Against Epilepsy, and the World Health Organization. Epilepsia 47(10), 1700–1722 (2006)
[18] Übeyli, E.D.: Combined neural network model employing wavelet coefficients for EEG signals classification. Digital Signal Processing 19(2), 297–308 (2009)
[19] Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1999)
[20] White, A., Williams, P., Ferraro, D., Clark, S., Kadam, S., Dudek, F.E., Staley, K.: Efficient unsupervised algorithms for the detection of seizures in continuous EEG recordings from rats after brain injury. J. Neurosci. Methods 152, 255–266 (2006)
[21] Williams, P.A., White, A.M., Clark, S., Ferraro, D.J., Swiercz, W., Staley, K.J., Dudek, F.E.: Development of spontaneous recurrent seizures after kainate-induced status epilepticus. J. Neurosci. 29, 2103–2122 (2009)
[22] Williams, P., White, A., Ferraro, D., Clark, S., Staley, K., Dudek, F.E.: The use of radiotelemetry to evaluate electrographic seizures in rats with kainate-induced epilepsy. J. Neurosci. Methods 155(1), 39–48 (2006)
[23] Wilson, S.: Spike detection: a review and comparison of algorithms. Clin Neurophysiol. 113, 1873–1881 (2002)
[24] Wilson, S.B., Scheuer, M.L., Plummer, C., Young, B., Pacia, S.: Seizure detection: Correlation of human experts. Clin Neurophysiol. 114, 2156–2164 (2003)

Automatic Detecting Documents Containing Personal Health Information

Yunli Wang, Hongyu Liu, Liqiang Geng, Matthew S. Keays, and Yonghua You

Institute for Information Technology, National Research Council Canada
46 Dineen Dr. Fredericton, NB, Canada
{Yunli.Wang,Hongyu.Liu,Liqiang.Geng,Matthew.Keays,
Yonghua.You}@nrc-cnrc.gc.ca

Abstract. With the increasing usage of computers and Internet, personal health information (PHI) is distributed across multiple institutes and often scattered on multiple devices and stored in diverse formats. Non-traditional medical records such as emails and e-documents containing PHI are in a high risk of privacy leakage. We are facing the challenges of locating and managing PHI in the distributed environment. The goal of this study is to classify electronic documents into PHI and non-PHI. A supervised machine learning method was used for this text categorization task. Three classifiers: SVM, decision tree and Naive Bayesian were used and tested on three data sets. Lexical, semantic and syntactic features and their combinations were compared in terms of their effectiveness of classifying PHI documents. The results show that combining semantic and/or syntactic with lexical features is more effective than lexical features alone for PHI classification. The supervised machine learning method is effective in classifying documents into PHI and non-PHI.

1 Introduction

People are usually concerned about their information privacy, especially personal health information (PHI). We are facing the transforming from paper-based medical records to electronic medical records, but the patient's privacy is easier to be breached with electronic medical records than paper-based medical records [4]. Many countries have introduced regulations for health information privacy such as HIPAA (Health Insurance Portability and Accountability Act) in USA and PIPEDA (Privacy Information Protection in Electronic Documents Act) in Canada. Although these legal frameworks have provided guidelines for organizations to protect privacy, compliance with the law and/or organizational privacy policies within many organizations is still challenging. Some medical organizations use techniques such as access control to protect electronic medical records stored in databases. However, healthcare services are distributed across medical institutes. Personal health information is often scattered on multiple devices and stored in diverse formats. These organizations need to locate PHI first and then use some technical means to protect it.

Not only organizations but also citizens themselves are responsible to protect privacy of PHI. PHI in this study is defined as individually identifiable health

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 335–344, 2009.

information, which is a subset of health information that identifies individuals. PHI is also called protected health information. In this paper personal health information and protected health information are used interchangeably. Patients gather and keep track of PHI from many sources and in many different forms, including Web pages, appointments, prescriptions, contacts, notes and emails. Information is fragmented by location, device, and form (such as paper, email, e-documents, Web references, and notes) [9]. Therefore, not only medical records but also any other documents containing PHI should be protected. For individuals, they need to identify PHI that might be disclosed through emails, file sharing, and text chat.

In this study, we target on privacy related document filtering to meet the need of organizations and individuals to manage PHI. After PHI documents are detected, an organization can conduct de-identification or trace the information flow of these documents for the purpose of audit. For individuals, they can protect identified PHI documents by password or encryption. Our problem is to detect PHI documents from any e-documents, and we treat it as a text categorization (TC) task. TC is a task of automatic assignment of documents to a predefined set of categories. Text categorization is often used for classifying documents into general categories such as sports, news, medical, and others. Machine learning approaches have gained popularity for solving TC problems in recent years . The machine learning process automatically builds a classifier for a category by observing the characteristics of a set of documents manually classified, and then classify a new unseen document into certain category by the classifier [12]. A large amount of work has been done in TC. However, very few works focus on TC in the context of privacy protection. In this study, a supervised machine learning method is used for classifying documents into PHI and non-PHI. The performance of such a system largely relies on the ability of features that capture the characteristics of training data, and the classifier that separates the positive and negative cases. We explored different types of features that may be useful for PHI classification, analyzed the impact of these features on the performance, and compared different classifiers.

2 Related Work

In general PHI detection at the document level is a text categorization problem. Bag of words is widely used as features in TC. However, it can not reflect the relationship between words and distinct the senses of a single word under different context. In recent years, other types of features such as semantic and syntactic features have been introduced [2, 3]. Some researchers used the extracted phrase to represent documents. These phrases were extracted based on background knowledge embedded in a existing ontology such as WordNet, MeSH (Medical Subject Headings) and UMLS (Unified Medical Language System) [2]. Cai and Hofman proposed a concept-based document representation method and probabilistic latent semantic analysis was used to extract concepts from documents [3]. These studies used some kind of domain knowledge to generate

relevant features. The domain knowledge might also be useful for discriminating PHI documents. We tested this hypothesis and investigated the usage of semantic indicators as the classification features, and compared their performance with the classic lexically-based features.

Lewis used syntactic parsing to generate indexing phrases [5]. The results show that the syntactic phrases did not perform better compared with that of common bag of words. Although syntactic patterns can indicate the relationships between words in some degree, syntactic phrases representation is highly redundant and noisy [5]. In this study, we explored different semantic and syntactic features for PHI classification. We used semantic and syntactic features as a complement to the standard bag of words approach. Syntactic structures were combined with semantic features to create new syntactic features in order to capture some unique characteristics of PHI documents.

Our work focuses on a specific application - privacy protection of PHI. For this applied area, the most related work with PHI classification is de-identification of medical records. Many studies used machine learning approach for de-identification, which was treated as a term-level classification problem [11,13,14]. Sazarva et al. used a rich feature set including orthographical features and semantic features (dictionary) [11]. Uzuner et al. used a more exhaustive feature set containing lexical information, syntactic features and semantic features [13]. MeSH was used to obtain some semantic features. Our task of PHI classification is different from de-identification of medical records. We aim to classify a document into PHI or non-PHI. Our task is a TC problem at the document level. However, the goal of de-identification is to remove PHI related entities from medical records. De-identification is considered as a named entity recognition problem at the term level.

3 Method

The goal of this study is to automatically assign documents into PHI and non-PHI category. We adopted a supervised machine learning method and examined two important factors: classification algorithms and feature types in this method. Many classification algorithms were used for text categorization. We choose three classifiers: SVM, decision tree and NaiveBayes. SVM has been successfully used for different classification tasks, especially in text categorization. Decision tree and NaiveBayes are also very popular for a variety of classification problems. We used three classifiers instead of one because we are interested in whether the performances of feature types are consistent across multiple classifiers.

Features used for text categorization are considered as relevant features for PHI classification. Bag of words features are most often used in a typical document classification problem. They are used in our task as basic features, and that allow the classifier to learn certain words that distinguish between PHI and non-PHI. PHI documents are characterized by the global and also local context. The category of the whole document is the global context. Personal health information includes identifiers such as name and address and health information. We

can use identifier entities and medical terms as the semantic features to represent the global context. Local context can be obtained from the sentence and phrase level. The existing of some particular type of sentences in a document is very useful for the classification of PHI. For example, "She was diagnosed with breast cancer." We extracted syntactic features to represent syntactic patterns of some sentences in a document. In total, three types of features: lexical, semantic and syntactic features that are relevant to PHI were extracted.

1. Lexical features
We extracted each individual word when its document frequency is over 5 times in a corpus as the lexical features. Lexical features are used as the baseline for the performance measure. We did preprocessing before extracting lexical features: stop words were removed and stemming was performed on terms. TF*IDF weights were used for lexical features.

2. Semantic features
Semantic features were generated to capture some global context of a document. We conjecture that medical terms will be useful in distinguishing PHI from non-PHI. In particular, medical terms in five categories: disease, medication, anatomy, medical procedure and medical device occur very often in PHI documents. They can be extracted using a medical ontology. UMLS is widely used as a medical ontology. The categories of health/medical concepts are represented in over 200 semantic types in UMLS. We extracted medical terms in five semantic groups: disease, medication, anatomy, procedure and device in UMLS. The semantic types in these five semantic groups were described in [7]. We used MMTx(Metamap Transfer) for extracting medical terms in these five semantic groups [1].

According to privacy rule, only the information that could be used to identify individuals is considered as protected health data. HIPAA requires that 18 entities are removed from medical records. The protected health information include names, geographic locations (more precise than a state), elements of dates except years, social security numbers, telephone and fax numbers, medical record numbers and more. These entities are probably useful in distinguishing PHI from non-PHI. Although age of a person under 89 is not considered as protected health information, we intuitively consider age as an important entity for classifying documents into PHI. Some information such as doctor title and hospital names often occurs in PHI narration. We consider the occurrence frequency of these entities in the PHI documents, and choose the following entities as semantic features:

- Person names: include first names and last names.
- Location: geographic locations such as cities, provinces, street names, postal codes.
- Phone and fax numbers: include telephone, pager and fax numbers.
- Email address.
- Age: a person's age.
- Doctors: include medical doctors and other practitioners.
- Hospital: names of medical organizations.

– Dates: a specific date or general description of date like "in Aug".
– Duration: the time during which something exists or lasts such as "for three months".

The PHI related entities were extracted using a named entity recognition system developed in house. We used the term frequency normalized with the document length as the feature weight for semantic features.

3. Syntactic features
Syntactic features capture the local syntactic dependencies of adjunctive words in a sentence. Some particular sentences that occur often in PHI documents have similar syntactic structures. They can be generally represented as "A person was diagnosed with Disease" or "A person took a Medication." We hypothesize these syntactic structures might be useful for distinguishing PHI from non-PHI. We extracted such syntactic structures as syntactic features.

A few steps were taken to obtain syntactic structures. At first, the whole document was parsed into sentences using an open source sentence splitter jTokeniser [10]. Then each individual word in each sentence was also obtained by jTokeniser [10]. Brill's POS tagging was used to obtain POS tags of each word in a sentence [6]. The POS categories were noun, verb, pronoun, proper noun, etc. Chunking combined POS tags into noun phrases and verb phrases. An open source tool, CRFChunker [8], was used to obtain noun phrases and verb phrases.The syntactic structures that occurred at least once at positive cases– PHI documents were extracted. If a syntactic structure is composed of Person, a verb and a medical entity, it was used as a syntactic feature. We use "She was diagnosed with breast cancer." as an example for syntactic features.

	She	was diagnosed with	breast cancer.
POS tagging	$\langle PRP \rangle$	$\langle VBD \rangle \langle VBN \rangle \langle IN \rangle$	$\langle NN \rangle \langle NN \rangle$
Chunking	$\langle PRP \rangle$	$\langle VP \rangle$	$\langle NP \rangle$
Syntactic feature	*Person*	diagnos	*Disease*

4 Experiments

4.1 Data Sets

We collected two data sets from two online health discussion forums Dipex (http://www.dipex.org.uk/) and Lounge (http://www.doctorslounge.com/).The messages posted on these two online health forums do not have real names of patients, but they have nicknames such as "Paul123" and other identifiers such as locations, email addresses, phone numbers, posted dates and times. The writing style of these messages is similar with other informal communication methods such as emails. We downloaded over two thousands messages from these two Web sites. Since the amount of data is very large, we generated a randomized subset including 250 messages from each Web site. Then, we manually labeled these messages into PHI and non-PHI. We assume the nicknames are real names

when we label a message. If it is possible to identify a particular person based on the content of the message and it also contains a person's health information, we labeled the massage as PHI. The advantages of using such real data sets are the data is anonymous and they are publically available. To observe the performance of feature sets on a data set from different sources, we combined these two data sets and used it as the third data set for experiments.

4.2 Evaluation Metrics

The effectiveness performance is measured with F-measure on PHI category. Our task is a binary classification: PHI or non-PHI. We are more concerned about the performance of PHI category, so the F-measure of PHI category is adopted as the performance measure. TP is the number of documents correctly assigned to PHI category. FP is the number of documents incorrectly assigned to PHI category. FN refers the number of documents that belongs to PHI category but is not assigned to PHI. The precision P, recall R, and F-measure F are calculated in the following:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}$$

$$F = \frac{(\beta^2 + 1)TP}{(\beta^2 + 1)TP + FP + \beta^2 FN}$$

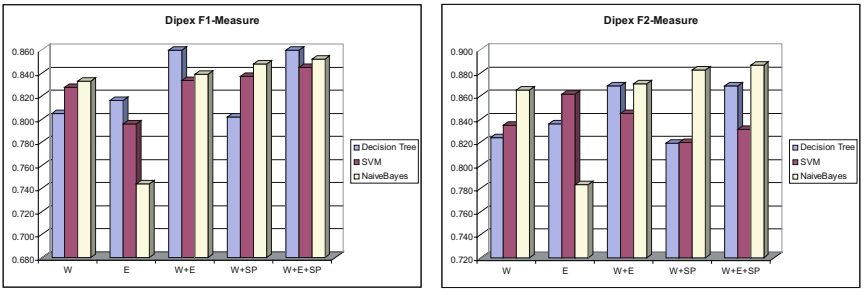
F1 ($\beta = 1$) is usually used for measuring the performance of a single category in text categorization problems. $\beta = 1$ means that FP and FN are equally important. In fact for PHI classification, we prefer a classifier with low FN rate. Classifying a PHI document into non-PHI might mean privacy leakage from this misclassified PHI document. However, classifying a non-PHI into PHI just increases the extra effort to protect non-PHI documents. Therefore, to emphasize the importance of low FN rate, we also use F2 ($\beta = 2$) to measure the performance of PHI classification.

$$F1 = \frac{2TP}{2TP + FP + FN}$$

$$F2 = \frac{5TP}{5TP + FP + 4FN}$$

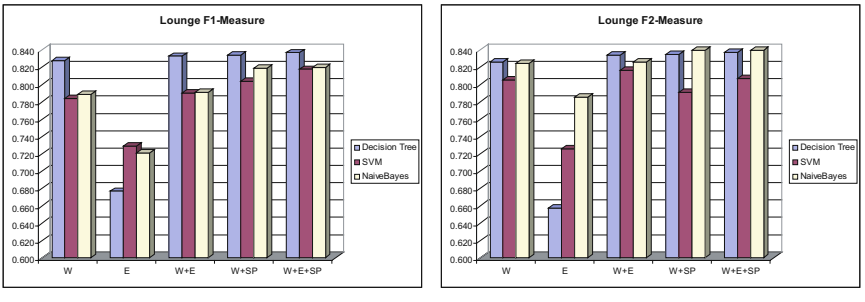
4.3 Results

Weka [15] was used as the testing platform. We used the implementation of SVM, Decision tree (J48 in Weka) and Naive Bayes algorithms and their default parameters in Weka. We performed 10 runs and 10 folds cross validation on each feature types. The lexical features were used as the baseline. Then we added semantic and syntactic features to compare the performance of each feature set and their combinations. The results of PHI classification on three data sets are shown in Figure 1-3.



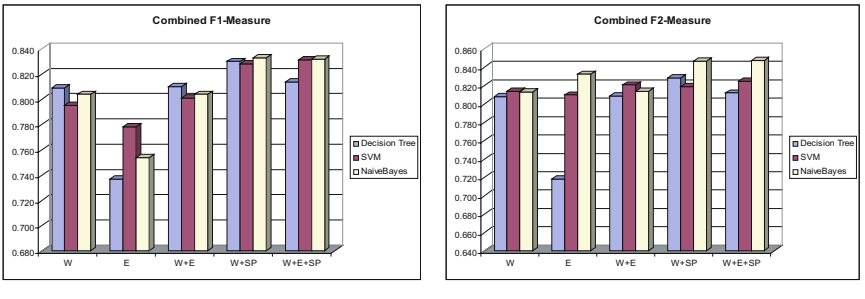
W-Lexical features, E-Semantic features, SP-Syntactic features

Fig. 1. F1 and F2 scores on Dipex for three classifiers using five feature sets



W-Lexical features, E -Semantic features, SP-Syntactic features

Fig. 2. F1 and F2 scores on Lounge for three classifiers using five feature sets



W-Lexical features, E-Semantic features, SP-Syntactic features

Fig. 3. F1 and F2 scores on combined date set for three classifiers using five feature sets

We examined two questions: which is the best feature set and which is the best classifier for this task? For the first question, we compared the performance of five feature sets in terms of F1 and F2 measures in figure 1-3. All significance tests were conducted with paired T test and P value < 0.05 is considered as

statistically significant. It shows that the combination of lexical and semantic features performs statistically significant better than lexical features in 7 out of 9 experiments with two exceptions of Naive Bayes on Lounge and Combined data sets. Also, the results obtained by combining lexical and syntactic features have shown statistically significant improvements over using only lexical features in 8 out of 9 experiments. While, the decision tree on Dipex data set is one exception. The performance of combined lexical, semantic and syntactic feature set is statistically significant better than that of combined lexical and semantic features in terms of F1 measure in 7 out of 9 experiments, with two exceptions of Decision tree on Dipex and Lounge data sets. Combining semantic and/or syntactic features with lexical features can improve the performance compared with standard bag of words approach for PHI classification. The content of Dipex is mostly patients or caregivers asked questions and other patients answered them or shared information. While, patients asked questions and medical professionals answered them on Lounge forum. The contribution of different feature sets for PHI classification might be different from these two Web sites. However, the influence of different feature sets on three data sets has the same tendency. While, based on the same feature set, the performance on Dipex is better than on Lounge and combined data set. It indicates different sources might have an impact on the performance.

For the second question, at first we looked at the performance of three classifiers in F1 measure. The performances of three classifiers are various on three data sets using five feature sets. Decision tree performs better than SVM and Naive Bayes on Lounge data set. Overall, there is no winner in majority experiments in terms of F1 measure. However, Naive Bayes outperforms Decision tree in 11, and SVM in 12 out of 15 experiments in terms of F2 measure. We looked at precision and recall more closely (data is not shown), and found that Naive Bayes has the best recall in all cases. The impact of adding semantic and syntactic features on lexical features with Naive Bayes is to improve precision slightly and improve recall significantly and that with Decision tree is to improve precision significantly and improve recall slightly. While, the influence of using semantic and syntactic features with lexical together on SVM is to improve precision but deteriorate recall. Since F2 measure is more important than F1 measure for our task, we conclude that Naive Bayes is a better classifier than SVM and decision tree for PHI classification. Interestingly, SVM does not outperform Decision tree and Naive Bayes in this study, which does not match the results previous reported for other text categorization tasks. There are a few reasons for this. Firstly, our task and data sets are different from other text categorization problems. Also, our data set is small compared with standard text categorization data sets. Secondly, we did not conduct parameter tuning on SVM. The performance of SVM could be improved by doing so. We used the RBF kernel function, but a different kernel function might improve the results.

The results show that our method is effective in classifying PHI documents. Lexical features have captured some global context of the whole document. Semantic features used in this study represent some kind of domain specific global

context. Therefore, the combined lexical and semantic features perform better than lexical feature alone. Our experiments also indicate that adding syntactic features with lexical and semantic features in PHI classification reached better performance. This is due to the combined feature set captures the global and also local context of a document. Syntactic features reached good performance in de-identification of medical records as well [13]. This relies on the ability of syntactic features in capturing the local context, which seems more important for de-identification. For PHI classification, local context and global context are equally important. For most other text categorization tasks, local context does not play a crucial role. That might explain the reason that using phrases extracted by syntactic parsing as features did not show improvements over using bag of words approach in other TC tasks.

5 Conclusions and Future Work

Privacy protection of PHI is an important issue that has raised concerns among government, institutes and individuals. As the first step to protect privacy, we need to locate PHI distributed in Intranet within organizations and stored in multiple devices in diverse formats. We studied the problem of classifying e-documents into PHI and non-PHI. A supervised machine learning method was used for automatic classification of PHI documents. We generated semantic and syntactic features and combined them with standard bag of words approach for PHI classification. The experiments were conducted on three data sets using three classifiers: SVM, decision tree and Naive Bayes. The results show that combining semantic and/or syntactic with lexical features is more effective than lexical features alone for PHI classification. Naive Bayes outperforms SVM and decision tress in terms of F2 measure, which emphasizes recall more than precision.

We tested the system on messages from online health discussion forum. Since the writing style of the data sets is similar to emails, the approach can be also applied to emails or letters. The performance of the system may not be generalized into other document types such as discharge summary. However, using semantic and syntactic features might be also applicable for these document types. We will test the system further on different document types and other data sets to verify the adoptability and generation of the method.

Since the concept of privacy is subjective, classifying a document into PHI or non-PHI seems arbitrary. In the real world, ranking documents instead of classification is more appropriate and practical. Such a ranked list would be very useful for the privacy officer to make the final decision. Another challenge we are still facing is most data sets contain a large amount of non-PHI documents but very limited PHI documents. Unbalanced data will deteriorate the performance of a supervised machine learning method. In the future, we will also explore the method for ranking documents in terms of the amount of PHI contained and dealing with unbalanced data.

References

1. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Proceedings of AMIA Symposium, pp. 17–21 (2001)
2. Bloehdron, S., Hotho, A.: Boosting for text classification with semantic features. In: Workshop on Text-based Information Retrieval (TIR 2004) at the 27th German Conference on Artificial Intelligence (2004)
3. Cai, L., Hofmann, T.: Text categorization by boosting automatically extracted concepts. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Toronto, CA, pp. 182–189 (2003)
4. Hodge, J.G., Gostin, L.O., Lacobson, P.D.: Legal issues concerning electronic health information privacy, quality, and liability. JAMA 282, 1466–1471 (1999)
5. Lewis, D.D.: An evaluation of phrasal and clustered representations on a text categorization task. In: Proceedings of SIGIR 1992, 15th ACM international conference on Research and Development in Information Retrieval, Copenhagen, Denmark, pp. 37–50 (1992)
6. Liu, H.: Monty tagger, http://web.media.mit.edu/~hugo/montytagger/
7. McCray, A.T., Burgun, A., Bodenreider, O.: Aggregating umls semantic types for reducing conceptual complexity. In: Proceedings of Medinfo 10(Pt 1), pp. 216–220 (2001)
8. Xuan-Hieu, P.: Crfchunker: Crf english phrase chunker (2006), http://crfchunker.sourceforge.net/
9. Pratt, W., Unruh, K., Civan, A., Skeels, M.M.: Personal health information management. Communication of ACM 49(1), 51–55 (2006)
10. Roberts, A.: jtokeniser (2005), http://www.andy-roberts.net/software/jTokeniser
11. Sazarva, G., Farkas, R., Busa-Fekete, R.: State-of-the-art anonymization of medical records using an iterative machine learning framework. JAMIA 14(5), 574–579 (2007)
12. Sebastiani, F.: Machine learning in automatic text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
13. Uzuner, O., Sibanda, T., Luo, Y., Szolovits, P.: A de-identification for medical discharge summaries. Artificial Intelligence in Medicine 42, 13–35 (2008)
14. Wellner, B., Huygk, M., Aberdeen, J., Morgan, A., Mardis, S., Peshkin, L., et al.: Rapidly retargetable approaches to de-identification in medical records. JAMIA 14(5), 564–573 (2007)
15. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

Segmentation of Text and Non-text in On-Line Handwritten Patient Record Based on Spatio-Temporal Analysis

Rattapoom Waranusast, Peter Haddawy, and Matthew Dailey

Computer Science and Information Management program,
School of Engineering and Technology, Asian Institute of Technology,
P.O. Box 4, Klong Luang, Pathumthani 12120, Thailand
{Rattapoom.Waranusast,haddawy,mdailey}@ait.ac.th

Abstract. Note taking is a common way for physicians to collect information from their patients in medical inquiries and diagnoses. Many times, when describing the pathology in medical records, a physician also draws diagrams and/or anatomical sketches along with the free-text narratives. The ability to understand unstructured handwritten texts and drawings in patient record could lead to implementation of automated patient record systems with more natural interfaces than current highly structured systems. The first and crucial step in automated processing of free-hand medical records is to segment the record into handwritten text and drawings, so that appropriate recognizers can be applied to different regions. This paper presents novel algorithms that separate text from non-text strokes in an on-line handwritten patient record. The algorithm is based on analyses of spatio-temporal graphs extracted from an on-line patient record and support vector machine (SVM) classification. Experiments demonstrate that the proposed approach is effective and robust.

Keywords: Automated patient record, Document segmentation, Spatio-temporal analysis, Online handwritten document.

1 Introduction

Introduction of computer-based patient record systems has become an important initiative in many countries in an attempt to improve healthcare quality and control costs [1]. While electronic patient record systems have many advantages over traditional paper based systems, acceptance among physicians has been slow. Many physicians find that electronic patient record systems are difficult to use require more time than traditional paper based systems [2, 3]. Electronic patient records are typically highly structured and make use of standardized vocabularies. While this facilitates processing of the information and communication among physicians, it has negative implications for the naturalness of the interaction. Physicians commonly use narrative and sketches to record patient information but the structured form of the records does not

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 345–354, 2009.

allow for this. Furthermore, physicians must map their concepts corresponding to their findings, diagnoses, and tests into the computer's predefined concepts, which requires time and can be constraining if the provided vocabulary is not sufficiently rich. All of these factors can result in the electronic patient record system interfering with the physician's though process rather than supporting it [4] [5]. Because of these factors, paper records still play a critical part in daily clinical work.

With the pen-based computing technology, the authors and colleagues are currently developing an electronic patient record system which can understand freehand anatomical sketches, annotations, and handwritten free-text. Understanding means that the system has the ability to recognize and interpret handwritten text, sketched drawings, and annotated symbols, and the ability to identify the context of the whole page. Such a system would combine the flexibility and user-friendliness of paper-based records with the ability to electronically process and search the information.

Normally, physicians use medical narratives in the form of unstructured text which are flexible, expressive, and familiar to them as an essential tool for diagnosis and decision making. They also often draw diagrams and/or anatomical sketches and annotate them. Some typical patient records are shown in Fig 1.

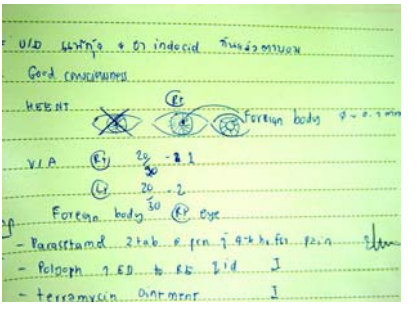


Fig. 1. Examples of paper-based patient records[1]

Because the page is potentially composed of different kinds of elements, the first and crucial step is to decompose the page into text and drawing regions. This step aims to group ink strokes of the same kind and to send them to proper recognizers in later steps.

In this paper, we present a novel approach to classifying ink strokes as text or non-text based on analyses of what we call spatio-temporal graphs. Instead of extracting features directly from ink stroke points as commonly found on other work in this field, features of the spatio-temporal graphs are calculated and extracted. The temporal neighborhoods are also taken into account based on the assumption that words and drawings are typically composed of contiguous strokes. We use a support vector machine (SVM) as a classifier to classify the strokes as text or non-text. The approach is robust to cursive and block hand writing.

---

[1] We thank to Dr. Siriwan Suebnukarn from Thammasat University for generously providing her medical expertise, data, and time.

2 Related Work

Since the late 1990's, the area of pen-based user interfaces has been very active. Applications include engineering drawings and simulations [6-8], architecture design [9], computer-aided design [10, 11], user interface and software design [12, 13], military planning [14], knowledge acquisition [15], music editing [16], and image retrieval [17, 18]. The only work in medical application of sketch-based interfaces that we know of is UNAS (UNderstanding Anatomical Sketches) [19], which is applied to COMET, a collaborative intelligent tutoring system for medical problem-based learning [20, 21].

In pen-based computing research, one of the most challenging problems is to separate the ink strokes into text, diagrams, and symbols. The objective of this process is to be able to cluster the strokes together and to send them to the right recognizers. This segmentation is necessary to design a robust interpretation of the ink even for an intelligent ink editing system [22]. Shilman et al. [23] also found that users prefer not to be constrained by facilitate ink understanding, such as pressing a button to switch modes or some special regions to identify the type of strokes [22]. Jain et al. [24] used a hierarchical approach to analyze on-line documents. They separate text and non-text stroke using only two stroke features, stroke length and stroke curvature, computed from individual strokes. This is based on the assumption that text strokes are typically short. However, this assumption cannot be applied to cursive writings and scribbles, which are ubiquitous in medical documents. Bishop et al. [25] proved that considering contexts of strokes allows better accuracy than using only stroke features. They use both features of the strokes and features of the gaps between strokes combined with temporal information using Hidden Markov model (HMM). Shilman et al. [23] use a completely different approach. They applied the bottom-up approach to separate text and graphics, that is starting with the strokes and repeatedly group them into letters, words, lines, and paragraphs. This method greatly depends on character recognition algorithms, which are not suitable for handwritten text scrawled by some physicians.

3 The System

The input data for our system are on-line digital ink documents, which are composed of sequences of strokes, separated by gaps. A stroke consists of a sequence of stroke points (x and y coordinates) recorded between a pen-down event (when the tip of the pen touch the screen) and a pen-up event (when the pen is lifted away from the screen). In addition to the spatial data of each stroke point, a time stamp indicating the time when the point was created is also recorded. These time stamps provide us the temporal ordering of the stroke points. It is obvious from other work [24-26] that the spatial information alone can give us rich useful features to separate text from non-text, we would expect to achieve better performance if we take temporal information in to account. We also believe that, people intuitively make the same kind of strokes consecutively before change to another kind of strokes. In other words, we can say that people will typically draw several graphics strokes in succession in order to draw a picture, or will scribble several text strokes in succession while writing some words. Because of this fact, we would also expect to improve the performance of the system by

extracting information from a group of neighboring strokes instead of an isolation of individual stroke. We describe our approach in detail in the following subsections.

3.1 Spatio-Temporal Graphs

To extract information from both spatial and temporal aspects of an ink document, we construct two graphs representing spatio-temporal relationships of strokes in the document. An X-T spatio-temporal graph is a graph with its horizontal axis represents temporal information of strokes (time stamp of each stroke point) and its vertical axis represents *x*-component of spatial information (*x* coordinates of each stroke points), constructed by plotting *x* value against time-stamp of each stroke point. A Y-T spatio-temporal graph also has temporal information of strokes as its horizontal axis and *y*-component of spatial information as its vertical axis. The Y-T graph is constructed in the same manner as the X-T graph, which is plotting *y* coordinates of a stroke point against its time stamp.

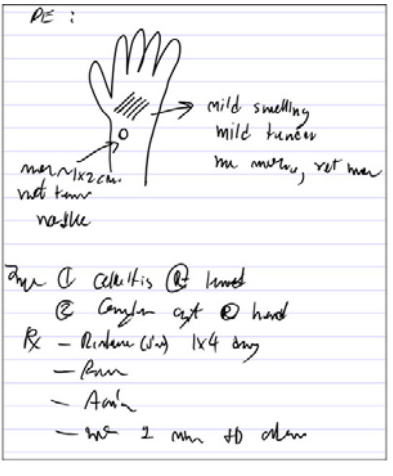


Fig. 2. A sample page of on-line patient records

Fig 2 shows a page of an on-line patient record, while Fig 3a and Fig 3b are the X-T and Y-T spatio-temporal graphs of the page in Fig 2, respectively. This patient record was written from left-to-right and top-to-bottom manner as found in the most western languages. In this example, the sketch of a patient's hand was drawn early in the page, which is corresponding to a part of the Y-T graph that look like a big U-shaped on the left side of the graph. We can notice from both X-T and Y-T graphs that there are some consistencies in the properties of text line strokes. We might say that stroke points of a text line, when plotting in X-T graph look like a slanted line, with its slope represents a horizontal speed of the writing. The fact that text letters have limited and almost the same height reflects as groups of horizontal lines in the Y-T graph. When we write a line of text, we move the pen much more up-and-downs than when we draw a picture. This fact is also shown as high frequency component

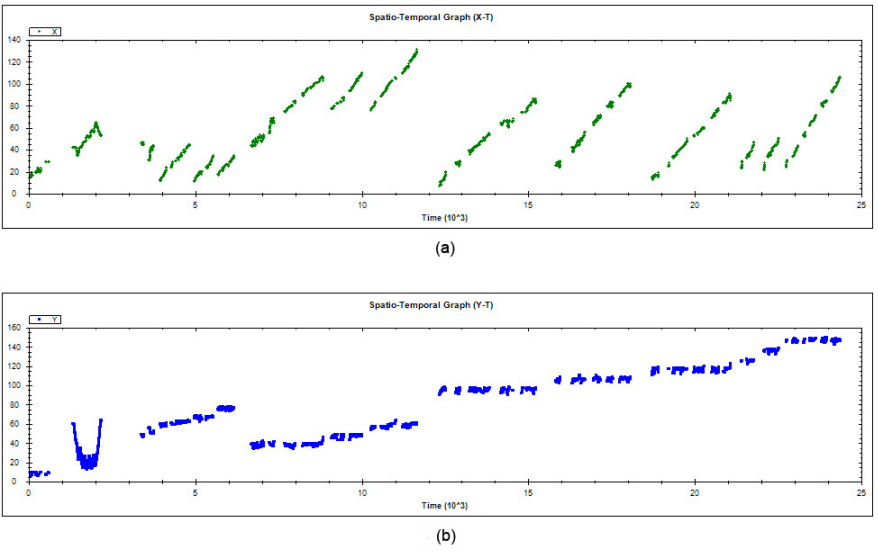


Fig. 3. Spatio-temporal graphs, X-T (a) and Y-T (b), constructed from the page in Fig 2

(looks like jagged line) in the graphs, especially the Y-T graph. From these unique properties of text lines in the spatio-temporal graphs, we can extract features from the graphs and use them to classify text or non-text strokes.

3.2 Features Extraction

At this point, we have two spatio-temporal graphs representing the whole strokes in the document, and we also know that text lines have certain unique properties, which are;

1) a text stroke and its neighbors form a line with certain angles in the X-T graph,
2) they form a horizontal line in the Y-T graph,
3) the widths of these lines (in both X-T and Y-T graphs) are consistency and limited, and
4) these lines are jagged with relatively high frequency.

We derive 9 features from these properties, which can be categorized into two groups, line-based, and frequency-based features. The followings are detail of these features.

Line-based Features: To compute the line-based features, we fitted an interesting graph point and its neighbors with a line. We use RANSAC [27] as a line-fitting method, because of its robustness to outliers. Once we fit a line to the graph points, we can compute; 1) An angle between temporal axis and fitted line in the X-T graph, 2) An angle between temporal axis and fitted line in the Y-T graph. 3) a correlation coefficiency of points in the X-T graph, 4) a correlation coefficiency of points in the

Y-T graph, 5) a standard error of estimation of the fitted line in the X-T graph, and 6) a standard error of estimation of the fitted line in the Y-T graph.

Frequency-based Features: In this category, we use two properties in computing the features. The first property is the relative extrema density [28], which is a measure for texture analysis. The relative extrema density are defined as the number of extrema per a period of time. The extrema are defined along the temporal axis in the following way. A graph point at time t is a relative minimum if its value f(t) satisfies

$$f(t) \leq f(t+1) \quad \text{and} \quad f(t) \leq f(t-1) \tag{1}$$

A graph point at time *t* is a relative maximum if

$$f(t) \geq f(t+1) \quad \text{and} \quad f(t) \geq f(t-1) \tag{2}$$

Note that with this definition, each point in the interior of any constant value run of points is considered simultaneously relative maxima. That is every points on a flat line or a plateau both are considered relative maxima. Plateaus at the local extrema, and plateaus on the way down from or up to the extremum also fall into this scenario, which is not desirable in our algorithm. To avoid this problem, we perform a preprocessing step and slightly modify the original relative extrema density algorithm. In the preprocessing step, we compress the graphs by eliminating the consecutive redundant value. To do that, the center of a plateau is considered to be a representative of that plateau, while the rest points in the plateau are ignored.

The second property is the stroke cusp density, which is the only property we compute directly from the stroke points, not the graph points. Stroke cusps are stroke points where the direction of the stroke has changed abruptly. The endpoints of each stroke are also considered as cusps. Fig 4 shows an example of cusps (depicted as superimposed red circles) on a sketch. From the figure, we can see that cusps are quite denser in a line of text (annotating words) than in a drawn picture (an eye). We define a stroke cusp density as the number of cusp in a period of time. We compute stroke cusps directly from the sketch, count them, and calculate the density.

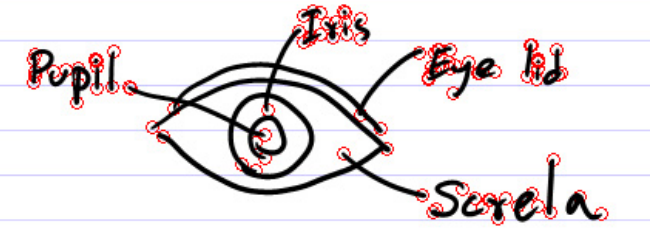


Fig. 4. An example of a sketch with cusps. Cusps are illustrated as red circles and superimposed on the sketch.

From the above definition of the relative extrema density and the cusp density, we derived three features from them as 7) a relative extrema density of the X-T graph, 8) a relative extrema density of the Y-T graph, and 9) a stroke cusp density.

To extract these 9 features, firstly, we have to cluster graph points into groups, because all of the above features can be computed only from a group of points. One possible way to compute these features is to divide a graph into some smaller segments, and extract these features from each segment. We have tried this approach, and found that this approach depends too much on divider algorithms. If the divider fails to divide a graph at proper positions, especially at the points between text and non-text graph points, the misclassifications are very high. To avoid this problem, we use moving window method to compute these features. The moving window is move along the temporal axis in both graphs. At a particular point in the graph that the window moves to, maximum weight is added to the point at the center of the window. Weights are gradually reduced and added to points that far away from the center of the window. The features at a particular point are computed from these weighted points.

At each stroke, features associated with the stroke are computed by averaging features of every stroke points in the stroke. The features of a stroke are kept to be used in the classification step.

3.3 Classification

Once all stroke features are computed, we construct a stroke classifier, which classifies strokes as text or non-text. Given these features, we believe that most standard classifiers can give us good results. We initially used k-nearest neighborhood (KNN) as our classifier. Then we tried a support vector machine (SVM) with C-SVM model and radial basis function to perform the classification. With this classifier, we performed parameter optimizations and found that the SVM is slightly more accurate than KNN.

4 Experimental Results

The experimental data was randomly collected from patient records at Thammasat University Hospital, Thailand. In total, 10,694 strokes (9,070 text strokes, and 1,624 non-text strokes) were collected from 83 patient records from approximately 8-10 physicians. Because the hospital used paper-based patient record and the physicians' time for their patients was very limited, it was unable to make the physicians sketch directly on a tablet PC. Thus we collected data by taking photos of the records at the end of examination sessions. Then, we copied the records into ink documents by displaying a patient record image on the tablet screen and sketching over it. Due to the limited number of data, we adopted 10-fold cross-validation technique to train and test the data. All strokes were stratified based on stroke class to ensure that the random sampling in cross-validation technique was done with each stroke class was properly represented in both training and test sets. To do this we employed Weka [29] as our test platform. Our algorithms achieved the accuracy of 94.61% in the 10-fold cross-validation test. Table 1 shows the confusion matrix of the results.

We also manually evaluated the results by looked through each classified page. We found that our approach provides good results, even though the text was scrawled and highly illegible. Fig 5 shows an example of our classified page, with non-text strokes marked in blue and text strokes in black.

Table 1. Confusion matrix showing the results of the evaluation. The rows correspond to the true class and the columns to the predicted classes.

		Classified as	
		Text	Non-Text
Actual	Text	8,938	132
	Non-Text	444	1,180

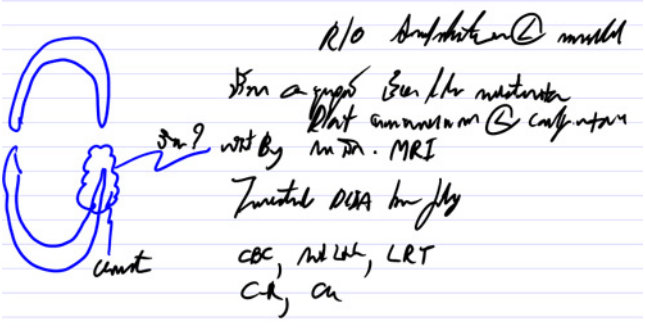


Fig. 5. An example of correctly classified scribbles in a patient record

5 Conclusion and Future Work

In this paper, we have presented a system for segmenting on-line patient records into text and non-text strokes. The system exploits features of the spatio-temporal graphs constructed from the patient record pages. The temporal neighborhoods of the spatio-temporal graph point are also taken into account, which based on the assumption that strokes of the same type tend to be correlated. We use support vector machine (SVM) as a classifier to classify the strokes. The experimental results demonstrate that the approach achieves high accuracy and is robust to cursive and scrawl writings which are ubiquitous in medical documents.

Although the accuracy of the approach is high, we believe that our current method can most probably be refined to yield better results. For example, at the moment, we ignore the properties of an individual stroke, which are usually utilized by most of other work in this area. For the moving window technique we applied in the algorithm, we move the window along the temporal axis of the graph. However, intuitively the strokes which are not in close spatial proximity should not be considered as the same group. Therefore, taking spatial axis into account in moving the window might improve the results. We also wish to gather more test data, especially on drawings, which will provide us more quantitative and qualitative insight of the nature of the medical documents.

References

1. HIMSS, 19th Annual HIMSS Leadership Survey: CIO Results Final Report. Healthcare Information and Management Systems Society (HIMSS) (2008)
2. Fraser, H.S., Biondich, P., Moodley, D., Choi, S., Mamlin, B.W., Szolovits, P.: Implementing Electronic Medical Record Systems in Developing Countries. Inform. Prim. Care. 13, 83–95 (2005)
3. Connolly, C.: Cedars-Sinai Doctors Cling to Pen and Paper. The Washington Post, March 21 (2005)
4. Walsh, S.H.: The Clinician's Perspective on Electronic Health Records and How They Can Affect Patient Care. Bmj. 328, 1184–1187 (2004)
5. Miller, R.H., Sim, I.: Physicians' Use of Electronic Medical Records: Barriers and Solutions. Health Affairs 23 (2004)
6. Stahovich, T.F., Interpreting the Engineer's Sketch: A Picture is Worth a Thousand Constraints. In: 1997 AAAI Symposium on Reasoning with Diagrammatic Representations II, Cambridge, Massachusetts, pp. 31–38 (1997)
7. Alvarado, C., Davis, R.: Intelligent Mechanical Engineering Design Environment: From Sketching to Simulation. MIT AI Laboratory Annual Abstracts. MIT, Cambridge, MA (2000)
8. Kara, L.B., Gennari, L., Stahovich, T.F.: A Sketch-based Tool for Analyzing Vibratory Mechanical Systems. Journal of Mechanical Design 130, (2008)
9. Leclercq, P.: Interpretative Tool for Architectural Sketches. In: 1st International Roundtable Conference on Visual and Spatial Reasoning in Design: Computational and Cognitive Approaches, Cambridge, MA, USA (1999)
10. Gross, M.D., Do, E.: Demonstrating the Electronic Cocktail Napkin: A Paper-like Interface for Early Design. In: Conference on Human Factors in Computing Systems, Vancouver, British Columbia, Canada, pp. 5–6 (1996)
11. Qian, D., Gross, M.: Collaborative Design with Netdraw. In: Computer Aided Architectural Design Futures Futures 1999, pp. 213–226 (1999)
12. Landay, J.A.: Interactive Sketching for the Early Stages of User Interface Design. Computer Science Dept., Vol. Ph.D. Carnegie Mellon University, Pittsburgh, Pa (1996)
13. Hammond, T., Davis, R.: Tahuti: A Sketch Recognition System for UML Class Diagrams - Extended Abstract. In: AAAI Spring Symposium on Sketch Understanding, pp. 59–68. AAAI Press, Stanford (2002)
14. Forbus, K.D., Usher, J., Chapman, V.: Sketching for Military Courses of Action Diagrams. In: 8th International Conference on Intelligent User Interfaces, pp. 61–68. ACM Press, Miami (2003)
15. Forbus, K., Usher, J.: Sketching for Knowledge Capture: A Progress Report. In: 7th International Conference on Intelligent User Interfaces, pp. 71–77. ACM Press, New York (2002)
16. Macé, S., Anquetil, E., Bossis, B.: Pen-Based Interaction for Intuitive Music Composition and Editing. In: Shen, J., Shepherd, J., Cui, B., Liu, L. (eds.) Intelligent Music Information Systems: Tools and Methodologies, pp. 261–288. Idea Group, USA (2007)
17. Fonseca, M., Barroso, B., Ribeiro, P., Jorge, J.: Sketch-Based Retrieval of ClipArt Drawings. In: Advanced Visual Interfaces (AVI 2004). ACM Press, Gallipoli (2004)
18. Leung, W.H., Chen, T.: Trademark Retrieval Using Contour-Skeleton Stroke Classification. In: IEEE Intl. Conf. on Multimedia and Expo (ICME 2002), Lausanne, Switzerland (2002)

19. Haddawy, P., Dailey, M., Kaewruen, P., Sarakhette, N.: Anatomical Sketch Understanding: Recognizing Explicit and Implicit Structure. In: Miksch, S., Hunter, J., Keravnou, E.T. (eds.) AIME 2005. LNCS, vol. 3581, pp. 343–352. Springer, Heidelberg (2005)
20. Suebnukarn, S., Haddawy, P.: A Collaborative Intelligent Tutoring System for Medical Problem-Based Learning. In: Int'l Conf. on Intelligent User Interfaces 2004, Madeira, Portugal, pp. 14–21 (2004)
21. Suebnukarn, S., Haddawy, P.: COMET: A Collaborative Tutoring System for Medical Problem-based Learning. IEEE Intelligent Systems 22, 70–77 (2007)
22. Anquetil, E., Lorette, G.: New Advances and New Challenges in On-line Handwriting Recognition & Electronic Ink Management. In: Chaudhuri, B.B. (ed.) Digital Document Processing: Major Directions and Recent Advances (Advances in Pattern Recognition), pp. 143–164. Springer, Heidelberg (2006)
23. Shilman, M., Wei, Z., Raghupathy, S., Simard, P., Jones, D.: Discerning Structure from Freeform Handwritten Notes. In: Seventh International Conference on Document Analysis and Recognition. IEEE Computer Society, Los Alamitos (2003)
24. Jain, A.K., Namboodiri, A.M., Subrahmonia, J.: Structure in On-line Documents. In: ICDAR 2001, pp. 844–848 (2001)
25. Bishop, C.M., Svensen, M., Hinton, G.E.: Distinguishing Text from Graphics in On-Line Handwritten Ink. In: 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR 2004). IEEE Computer Society, Los Alamitos (2004)
26. Namboodiri, A.M., Jain, A.K.: Robust Segmentation of Unconstrained Online Handwritten Documents. In: ICVGIP, pp. 165–170 (2004)
27. Fischler, M.A., Bolles, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Comm. of the ACM 24, 381–395 (1981)
28. Rosenfeld, A., Troy, E.B.: Visual Texture Analysis. In: Conf. Rec. Symp. Feature Extraction and Selection in Pattern Recognition, p. 115 (1970)
29. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

An Ontology-Based Method to Link Database Integration and Data Mining within a Biomedical Distributed KDD

David Perez-Rey and Victor Maojo

Artificial Intelligence Department, Facultad de Informática,
Universidad Politécnica de Madrid
28660 Boadilla del Monte, Madrid, Spain
dperezdelrey@fi.upm.es

Abstract. Over the last years, collaborative research has been continuously growing in many scientific areas such as biomedicine. However, traditional Knowledge Discovery in Databases (KDD) processes generally adopt centralized approaches that do not fully address many research needs in these distributed environments. This paper presents a method to improve traditional centralized KDD by adopting an ontology-based distributed model. Ontologies are used within this model: (i) as Virtual Schemas (VS) to solve structural heterogeneities in databases and (ii) as frameworks to guide automatic transformations when data is retrieved by users—Preprocessing Ontologies (PO). Both types of ontologies aim to facilitate data gathering and preprocessing while maintaining data source decentralization. This ontology-based approach allows to link database integration and data mining, improving final results, reusability and interoperability. The results obtained present improvements in outcome performance and new capabilities compared to traditional KDD processes.

Keywords: Database Integration, Distributed KDD, Ontologies, Preprocessing, Data Mining.

1 Introduction

Traditional KDD methodologies were conceptually based on a central data warehouse, where data are stored, manipulated and accessed for further analysis [1]. Centralized approaches do not cover the gap between data integration tasks and data mining. It is not optimal for current collaborative research in some areas such as biomedicine, where distributed environments facilitate information exchange communication networks [2]. In such distributed scenarios, the management of structural and semantic heterogeneities—which are common in many scientific and technological areas—is a major challenge. To address this issue several efforts have been carried out in the area of heterogeneous database integration, using ontologies to perform the required schema transformations [3] and preprocessing [4] in distributed KDD processes.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 355–359, 2009.

The proposed architecture address, at the same time, data integration and preprocessing in distributed KDD environments. This approach covers collaborative scenario requirements, maintaining the decentralization of the sources and solving some centralized approach problems such us periodical updates or redundancy. Ontologies are used as a framework to solve schema and instance heterogeneities, facilitating reuse and interoperability with other domains.

2 Background

Many scientists have dedicated important efforts during the last years to develop new methods for heterogeneous data source integration. Until recently, many projects approached schema integration by manually linking virtual artifacts with physical databases. Although the literature suggest that preparing the data before the data mining process is a key issue in KDD [5], there still exist few efforts in data preprocessing compared to database integration or data mining.

In classical KDD methodologies data must be preprocessed prior to the process of data mining. Such preparation is covered by different concepts and terms, such as data cleaning, data quality, preprocessing, homogenization, standardization, data integration and others. Although each of these terms have different objectives and must be approached independently, two main tasks have been considered in this paper from the functional point of view: (i) data integration to gather source schemas, and (ii) preprocessing to integrate, clean and transform data instances. The first one is where the semantic web—and ontologies—has found the best synergy within the KDD process. Commercial tools focused in the former task are usually related to data warehousing (DW) implementations. ETL—Extracting, Transforming and Loading—tools are in charge of managing the original data into the DW. Main projects on the field use ontologies: (i) to store instances of the sources and homogenize them according to the ontology hierarchy [6] and (ii) as a structure to store the information needed to perform the required transformations.

3 Method

The proposed KDD model has been designed to address both data integration and preprocessing in distributed KDD scenarios. It is based on previous developments carried out to solve, separately, schema and instance integration. For schema integration, the OntoFusion system [3], and OntoDataClean [4] for instance integration and preprocessing. To solve data heterogeneities and preprocess the data from the available sources, a two sequential step methodology is proposed. First, manual or semi-automatic inconsistency detection from one or various data sources is performed. Then, data transformation can be automated.

Schema integration is solved in the proposed model linking each original source to a first-level Virtual Schema (VS)—an ontology. Terms must be obtained from a shared domain ontology, previously reused or created, to link each concept with its corresponding element or relation in the physical database schema.

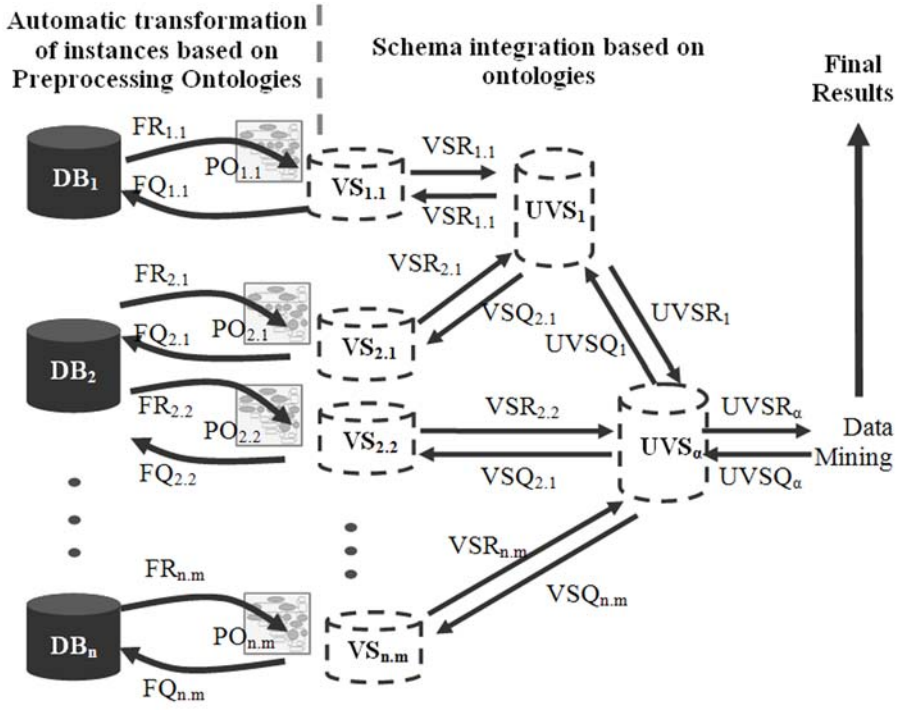


Fig. 1. Proposed Automatic Data Transformation in a Distributed KDD Process

Afterwards, the unification of VSs into Unified Virtual Schemas (UVS) is an automatic process guided by the common domain ontology. Regarding instance integration and preprocessing—e.g. solving scale, synonyms, format and other issues—, once the inconsistencies are detected the corresponding transformations are stored in a Preprocessing Ontology (PO).

Once the corresponding VSs and POs have been created, VS unification and the final task of retrieving and transforming the data before they are analyzed can be performed automatically. The distributed approach is possible since ontologies guide the process of data integration and preprocessing, maintaining the access to original data sources each time a query is executed. Figure 1 describes the different elements involved in the proposed data transformation method.

Let us suppose that a query launched at the unification of n sources (DB_1, DB_2, ..., DB_n) is required. With the proposed solution it is possible to browse the conceptual model of all the sources together by means of ontologies (UVS_α). The virtual query ($UVSQ_\alpha$) is automatically generated and the user do not need to understand the physical structure of the database, since this query is referred just to concepts, attributes and relationships of the UVS_α ontology. Then, this virtual query is divided into queries of the next level queries and sent to the corresponding VS recursively. When they reach the first-level VS ($VS_{1.1}$, ..., $VS_{n.m}$) a set of final queries are generated ($FQ_{1.1}$, ..., $FQ_{n.m}$) in the source

native language and according to the mapping between physical schemas and VSs. Data retrieved ($FR_{1.1}$, ..., $FR_{n.m}$) are transformed according to the POs ($PO_{1.1}$, ..., $PO_{n.m}$). Then results are combined and propagated through the VS hierarchy to the initial UVS, where it is retrieved by the user.

4 Results

Two datasets from the biomedical field have been applied to evaluate the suitability of distributed KDD environments using our approach. A dataset of breast patients widely used in the machine learning literature, and one obtained from the SEER database [7]. Different subsets were obtained using the proposed method from both datasets, one for each original source, one for traditional schema integration and the last one using the proposed method. These datasets were compared after applying data mining algorithms.

Outcome performance improvements were expected due to the increment of records to train the different models. Besides to quantitative outcome improvements, the main goal of this work was to test new functionalities for collaborative environments. Table 1 shows a summarized comparison of the proposed method with other database integration projects and data mining frameworks. We have considered main efforts on ontology-based database integration such as Semeda [8], KAON Reverse [9] and D2OMapper [10]; and Weka [11] to represent data mining frameworks.

Table 1. Comparison of Database Integration and Data Mining Projects Functionality (According to the Latest Information Available)

Functionality	Semeda	KAON Reverse	D2O Mapper	Weka with Workflow	Proposed Solution
Distributed Approach	√	√	√	×	√
Ontology-based	√	√	√	×	√
Schema Integration	√	√	√	×	√
Automatic Schema Mapping	×	×	√	×	×
Schema Redesign	×	×	×	×	√
Instance Integration	×	×	×	×	√
Automatic Inconsistency Detection	×	×	×	×	√
Data Preprocessing	×	×	×	√	√
Data Mining Phase	√	√	√	√	×

With new capabilities such as schema redesign, instance integration or semi-automatic detection of inconsistencies, this approach provides a bridge between traditional database integration projects and data mining frameworks.

5 Conclusions

A novel approach for distributed KDD in collaborative environments has been proposed in this paper. Promising results were obtained, with significant

improvements on final outcomes of a KDD process. The major contribution of the proposed method is to provide new capabilities in a distributed KDD. Transparent access to a set of heterogeneous sources in an ontology-based KDD framework allows to link data integration and data mining phases.

Once we have observed the viability of this ontology-based approach, further work is in progress to extend the PO structure to cover new transformations and solve other categories of inconsistencies. Automation of the mapping process is also been developed, since it can be a time-consuming task when dealing with a large number of databases. Finally, Grid frameworks can be used to deal with a distributed KDD process as a resource planning-including data sources and algorithms.

Acknowledgements

This research has been supported by funding from the Ministry of Innovation and Science (Spain) and the European Commission (ACGT and ACTION-grid projects). Thanks to M. Zwitter and M. Soklic for the Breast Cancer dataset.

References

1. Fayyad, U., Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in databases. AI Magazine 17, 37–54 (1996)
2. Gurwitz, D., Lunshof, J.E., Altman, R.B.: A call for the creation of personalized medicine database. Nature Reviews, Drug Discovery 5, 23–26 (2006)
3. Perez-Rey, D., et al.: ONTOFUSION: Ontology-Based Integration of Genomic and Clinical Databases. Comput. Biol. Med. 36, 712–730 (2006)
4. Perez-Rey, D., Anguita, A., Crespo, J.: OntoDataClean: Ontology-based Integration and Preprocessing of Distributed Data. In: Maglaveras, N., Chouvarda, I., Koutkias, V., Brause, R. (eds.) ISBMDA 2006. LNCS (LNBI), vol. 4345, pp. 262–272. Springer, Heidelberg (2006)
5. Weiss, S.M., Indurkhya, N.: Predictive Data Mining: A Practical Guide. Morgan Kaufmann, San Francisco (1998)
6. Kedad, Z., Metais, E.: Ontology-based Data Cleaning. In: Andersson, B., Bergholtz, M., Johannesson, P. (eds.) NLDB 2002. LNCS, vol. 2553, pp. 137–149. Springer, Heidelberg (2002)
7. SEER Cancer Statistics Review. Surveillance, Epidemiology and End Results (SEER) program, http://www.seer.cancer.gov/ (last accesed on April 2009)
8. Kohler, J., Philippi, S., Lange, M.: SEMEDA: ontology based semantic integration of biological databases. Bioinformatics 19(18), 2420–2427 (2003)
9. Librelotto, G.R., Souza, W., Ramalho, J.C., Henriques, P.R.: Using the Ontology Paradigm to Integrate Information Systems. In: International Conference on Knowledge Engineering and Decision Support, pp. 497–504 (2003)
10. Xu, Z., Zhang, S., Dong, Y.: Mapping between Relational Database Schema and OWL Ontology for Deep Annotation. In: International Conference on Web Intelligence, pp. 548–552 (2006)
11. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2005)

Subgroup Discovery for Weight Learning in Breast Cancer Diagnosis

Beatriz López[1], Víctor Barrera[1], Joaquim Meléndez[1], Carles Pous[1], Joan Brunet[2], and Judith Sanz[3]

[1] Institut d'Informática i Aplicacions, Universitat de Girona, Girona, Spain
{blopez,vbarrera,quimmel,carles}@eia.udg.edu
[2] Institut d'Investigació Biomèdica de Girona and Institut Català d'Oncologia, Girona, Spain
www.iconcologia.net
[3] Hospital Sant Pau, Barcelona, Spain

Abstract. In the recent years, there is an increasing interest of the use of case-based reasoning (CBR) in medicine. CBR is an approach to problem solving that is able to use specific knowledge of previous experiences. However, the efficiency of CBR strongly depends on the similarity metrics used to recover past experiences. In such metrics, the role of attribute weights is critical. In this paper we propose a methodology that use subgroup discovery methods to learn the relevance of the attributes. The methodology is applied to a Breast Cancer dataset obtaining significant improvements. ...

1 Introduction

Case-based reasoning (CBR) is an approach to problem solving that is able to use specific knowledge of previous experiences or cases. In the recent years, there has been an increasing interest in the application of CBR in Medicine and its integration with knowledge discovery techniques due to the successful history of CBR in this field. Reasoning from experiences seems to fit the evidence reasoning method with which physicians diagnose [1].

A new method to automatically score data attributes is proposed in this paper. It has been tested with complex medical data containing numerical and categorical attributes. The method is based on the analysis of rules extracted by a subgroup discovery algorithm [2] and a furhter analysis of their relevance to assign scores to attributes.

The combination of subgroup discovery and CBR is a recent field of study. In [3], for example, the authors propose the use of CBR to interpret the results of the subgroups obtained. This work is similar to ours, since they are also proposing scoring functions to determine the attribute relevance. However, their objective is to define case prototypes.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 360–364, 2009.

2 Subgroup Discovery

The objective of Subgroup Discovery (SD) is to induce rules from examples [4,2]. Rules have the form $Cond \rightarrow Class$ where $Cond$ is a conjunction of features (attribute-value pairs) and $Class$ is a class value.

The aim of subgroup descriptions is to cover a significant proportion of positive examples. Algorithms as EXPLORA [4], MIDOS [5] and CN2-SD [2] are examples of this relatively new way of addressing a learning problem.

CN2-SD is one of the state-of-the-art methods on subgroup discovery [2] based on CN2 [6], a rule learner sutable for classification, prediction applications.

CN2-SD modifies the CN2 algorithm mainly by avoiding deleting the covered positive examples from the current training set [2]. Instead, CN2-SD stores with each example a count indicating how often (with how many rules) the example has been covered so far. In order to do so, CN2-SD uses an additive and multiplicative schema.

3 Weight Learning with Subgroup Discovery

Our starting point for weight learning is the CN2-SD algorithm and a CBR system. SD identifies attribute-value pairs or features, while distance functions work with attributes. Thus, the method we propose consists of the following four main steps:

- To use subgroup discovery to extract rules that describe cases in the case base. This set of rules describes subgroups present in the examples. The set of rules provided by the subgroup discovery method (input of the second step) can be represented by the association between a set of conditions and the corresponding class plus additional information related to the quality of the rule:

$$R_j =< COND_j, Class_i, Q(R_j) > \tag{1}$$

where $COND_j$ is the set of conditions of the rule expressed as attributevalue pairs (a_k, v_l) or features, $Class_i$ the class it describes, and $Q(R_j)$ a vector containing the number of cases covered by the rule, number of conditions and other quality indices.

- Quantify the relevance of each feature in the set of rules obtained. In order to quantify the relevance of features contained in the extracted rules we defined a score for each attribute-value pair (a_k, v_l) in $COND_j$. This $score[(a_k, v_l)] \in \Re$ can be computed according to different criteria. We used the apparition frequency of each feature in the rule set as score of each one. All scores related to the same attribute a_k are combined to obtain the attribute weight, $w(a_k)$, as follows:

$$w(a_k) = F(score[(a_k, v_l)])_{\forall l} \tag{2}$$

- Map relevance of features to weights in the attribute space of the CBR system. If the a simple addition is used as the F function, we obtain the following expression:

$$w(a_k) = \sum_{l=1}^{n} score[(a_k, v_l)] \tag{3}$$

where n is the number of values that appears for the same attribute. Notice that only attribute-value pairs containing the k attribute are taken.

Apply the CBR cycle with the learned weights. Classification is performed according to similarity among the new case and the k-NN. Thus, given two cases C_1, C_2, their similarity is computed as follows:

$$Sim(C_1, C_2) = 1 - \frac{\sum_i w_i dist(a_1^i, a_2^i)}{\sum_i w_i} \tag{4}$$

where $dist(a_1^i, a_2^i)$ is the distance between the i attribute of the two cases, and w_i the learnt weight of the attribute.

Table 1 shows a rule set example that will be used to explain our score procedure.

Table 1. Rule set, feature, score and weight computing example

Rule	Feature	Score	Weight
$(Age > 59) \wedge (Family\ risk > 2.3) \rightarrow Cancer$	$Age > 59$	2/3	
$(Family\ risk > 4.1) \rightarrow Cancer$	$Familyrisk > 2.3$	1/3	$w(Age) : 2/3$
$(Age > 59) \wedge (Family\ risk < 0.9) \rightarrow No\ cancer$	$Familyrisk > 4.1$	1/3	$w(Fam...) : 3/3$
	$Familyrisk < 0.9$	1/3	

This rule set contains three rules containing two attributes (Age and $Family\ Risk$) and two classes ($Cancer$, $No\ Cancer$). The rule number one has two attribute-value pairs, $(Age > 59)$ and $(Family\ risk > 2.3)$.

Attribute-value pairs, (a_k, v_l) 2nd column in Table 1, can be scored following the rating criteria resulting in the scores shown in 3th column of the Table 1. For instance, $(Age > 59)$ feature appears in two of the three rules contained in the example rule set, so its score is 2/3. These scores represent a measure of relevance of these features and will be mapped as weights. Therefore, the weights obtained for attributes Age and $Family\ risk$ are 2/3 and 3/3 respectively (4th column in Table 1).

4 Case Study

We have used a Breast Cancer dataset with 628 cases corresponding to healthy people and 243 cases to women with breast cancer.

To test the methodology, we have used the CN2-SD implementation provided in KEEL data mining platform [7]. Particularly, we used the CN2-SD multiplicative scheme. We have mapped the scores into weights following the equation 3. eXiT*CBR framework [8] was used to implement the last step of the methodology. Two experimental settings have been considered:

- Human-filtered attributes: In this configuration, the instances are characterized by the 85 relevant attributes manually selected by a physician of the Hospital Sant Pau from the 1199 initial numerical and categorical attributes.
- Our method: The instances are characterized by the attributes learnt using the methodology based on CN2-SD algorithm from 85 attributes manually selected.

We have followed a cross-validation procedure, with 90% of the cases for trainning and 10% for testing.

As a result of applying subgroup discovery, 11 rules have been obtained with an average of 12.9 attributes per rule. Five rules correspond to examples of healthy people (Class=*No cancer*), four cancer cases and the last two rules describe unknown or borderline cases[1]. These last two rules were ignored.

After applying the score procedure, 76 different relevant features have been obtained. Most of them (69%) only appear once, whereas 21% are relevant in two rules. After the mapping procedure, these features have then been mapped into 40 attributes. Consequently, the attribute space has been reduced significantly (from 85 to 40 attributes).

Two weights with value 1 identify relevant attributes related to thyropathy and mamography. Other attributes have lower relevance, and maybe the most important benefit is the attribute reduction explained above.

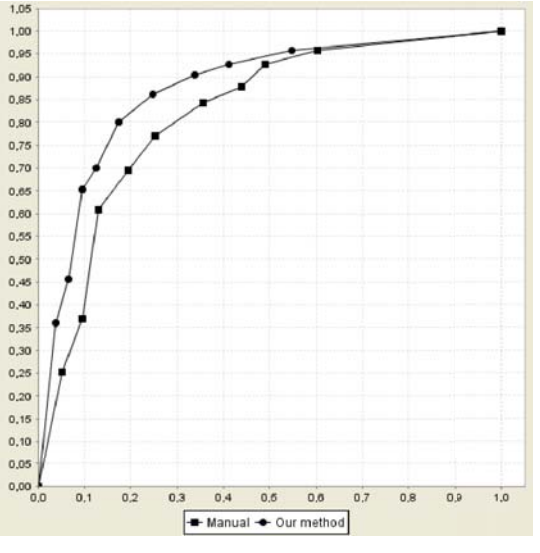


Fig. 1. ROC curves of our two experiments. Squares: results with relevant attributes provided by some physicians. Circles: results with relevant attributes according to subgroup discovery.

[1] That means SD tool has been unable to predict such cases as cancer or not.

For each experimental setup we have applied CBR and plotted in ROC (*Receiver Operator Characteristics*) [9] curves the average results of all the folds.

Figure 1 show the results obtained in each configuration. As it is possible to observe, our methodology significantly improves the results obtained. Analizing the area under the curve (AUC), we obtained 0.816 for the attribute manually selected and 0.854 for our method. So we are not loosing diagnosing capacity after reducing the space of attributes, but improving it.

5 Conclusions

In this paper we have proposed a new methodology so that subgroup discovery methods are used to learn attribute weights. These kind of methods have been designed to identify particular subgroups in a set of data, but they also allow to find regularities in the data. In this sense, we have taken advantage of this property, and we have used them to learn the attribute weights that then are further used in a medical CBR system.

We have tested our methodology with a Breast Cancer dataset. The results obtained show that our learnt weights are much better than the ones provided by human experts.

This research project has been partially funded by the Spanish MEC projects DPI 2006-09370, CTQ2008-06865-C02-02/PPQ, TIN2008-04547/TIN, and Girona Biomedical Research Institute (IdiBGi) project GRCT41.

References

1. Bichindaritz, I., Montani, S., Portinale, L.: Special issue on case-based reasoning in the health sciences. Applied Intelligence 28(3), 207–209 (2008)
2. Lavrac, N., Kavsec, B., Flach, P., Todorovski, L.: Subgroup discovery with cn2-sd. Journal of Machine Learning Research 5, 153–188 (2004)
3. Atzmueller, M., Puppe, F.: A case-based approach for characterization and analysis of subgroup patterns. Applied Intelligence 28(3), 210–221 (2008)
4. Klosgen, W.: Explora: A multipattern and multistrategy discovery assistant. In: Fayyad, U., Piatetsky- Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, p. 249. AAAI Press, Menlo Park (1996)
5. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Proceedings of Conference Principles Of Data Mining And Knowledge Discovery, pp. 78–87 (1997)
6. Clark, P., Niblett, T.: The CN2 induction algorithm. Machine Learning 3(4), 261–283 (1989)
7. Alcal-Fdez, J., Snchez, L., Garca, S., del Jesus, M., Ventura, S., Garrell, J., Otero, J., Romero, C., Bacardit, J., Rivas, V., Fernndez, J., Herrera, F.: Keel: A software tool to assess evolutionary algorithms to data mining problems. Soft Computing, http://sci2s.ugr.es/keel, doi:10.1007/s00500-008-0323-y
8. Pous, C., Pla, A., Gay, P., López, B.: exit*cbr: A framework for case-based medical diagnosis development and experimentation. In: ICDM Workshop on Data Mining in Life Sciences, pp. 96–105. Ibai Publishing (2008)
9. Fawcett, T.: Roc graphs: Notes and practical considerations for data mining researchers. HP Labs, Technical Report HPL-2003-4 (2003)

Mining Discriminant Sequential Patterns for Aging Brain

Paola Salle[1], Sandra Bringay[1,2], and Maguelonne Teisseire[3]

[1] LIRMM, Montpellier 2 University, CNRS, 161 rue Ada, 34392 Montpellier, France
[2] MIAP Dpt., Montpellier 3 University, Route de Mende, 34199 Montpellier, France
[3] Cemagref, UMR TETIS, 500 rue Jean-François Breton, F-34093 Montpellier, France
{salle,bringay}@lirmm.fr, maguelonne.teisseire@cemagref.fr

Abstract. Discovering new information about groups of genes implied in a disease is still challenging. Microarrays are a powerful tool to analyse gene expression. In this paper, we propose a new approach outlining relationships between genes based on their ordered expressions. Our contribution is twofold. First, we propose to use a new material, called sequential patterns, to be investigated by biologists. Secondly, due to the expression matrice density, extracting sequential patterns from microarray datasets is far away from being easy. The aim of our proposal is to provide the biological experts with an efficient approach based on discriminant sequential patterns. Results of various experiments on real biological data highlight the relevance of our proposal.

1 Introduction

Over recent years, thanks to the emergence of new molecular genetic technologies such as DNA microarrays, it has been possible to obtain a global view of the cell [9]. How genes are related and how cells control and regulate their expressions is crucial to understand diseases. Thus, DNA microarrays is a popular way for gene expression analysis. They provide an expression level for genes under given biological situations.

The contribution of this paper is to outline new relationships between genes from microarray data by focusing on ordered expressions. For example, it could be interesting to point out that the gene $Gene_A$ has a lower expression than the genes $Gene_B$ and the $Gene_C$ whose expressions are close. Such relationships can be represented by sequential patterns. Introduced by [1], the sequential pattern problem has been extensively addressed by the data mining community. However, microarray datasets are very dense because they contain numerous genes and each gene has one value for each DNA microarray, making the above methods ill suited for this type of data. We propose to introduce domain knowledge during the mining task. In this way, the search space is reduced and more relevant results (from a biological point of view) are obtained. We also reduce the set of extracted patterns to discriminant patterns.

This paper is organized as follows. In Section 2, we give an overview of our proposal: DSPAB (Discriminant Sequential Patterns for Aging Brain). Experiments described in Section 3 highlight the relevance of our proposal. In Section 4, our proposal is compared to related work in the biological context.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 365–369, 2009.

2 Proposal

2.1 Preliminary Definition: Discriminant Sequential Patterns

Let DB (see Table 1) be a set of records where each record r is composed of: an id (DNA microarrays), a record timestamps (gene expression) and items (genes).

An itemset it_p is a non ordered group of genes wich have similar expression, e.g. $(Gene_1 Gene_3 Gene_4)$.

A sequence $S =< it_1 it_2..it_p >$ is a non-empty and ordered list of itemsets, i.e. groups of genes ordered according to their expression, e.g. $< (Gene_4)(Gene_1 Gene_3) >$. The support of a sequence $support(S)$ is defined as the fraction of total data-sequences that contain this sequence. A sequence S is frequent if the condition $support(S)$ is greater than a parameter specified by the user called $minSupp$. In the database DB, with a minimal support equal to 100%, $< (Gene_4)(Gene_2) >$ is a sequential pattern.

Table 1. Database DB

DNAmicroarrays (id)	Expression (timestamps)	Genes (items)
1	100	$Gene_1$ $Gene_3$ $Gene_4$
	400	$Gene_2$
2	20	$Gene_4$
	120	$Gene_1$ $Gene_3$
	430	$Gene_2$

Discriminant Sequential Patterns are inspired from "Emerging Patterns" (EPs) [4] which enable to capture significant changes between two classes D_1 and D_2. In our context, we aim at discovering sequential patterns that discriminate two classes: healthy and sick animals, in particular patterns present in the first class D_1 and absent in the second class D_2. Therefore, we define two thresholds : If $support(S)_{D1} \geq minSupport_{D1}$ and $support(S)_{D2} \leq maxSupport_{D2}$ then S is considered as discriminant.

2.2 Overview

Microarray datasets are so dense that we have to deal with a huge search space which is intractable by traditional mining methods. To overcome these problems, we first reduce the search space by using domain knowledge and Pearson correlation. Then, we extract discriminant sequential patterns.

Search space reduction by using knowledge source. Some genes are known by the biologists to be involved in a particular disease. Our aim is to outline relationships between these *preferred genes* and other genes not already known for their implication in the disease. The extraction of sequential patterns is supervised by these preferred genes. Thus, the benefits are twofold: the search subspace is defined thanks to the knowledge source and the results have biological meaning.

Pearson Correlation. Pearson correlation is used in order to reduce the number of candidate sequences. We generate only sequential patterns composed of correlated genes.

$$\alpha_{XY} = \frac{Cov(X,Y)}{\sigma x \sigma y} \quad (1)$$

$Cov(X,Y)$ is the covariance of the two variables, σx (resp. σy) is the standard deviation of X (resp. Y) and $-1 \leq \alpha_{XY} \leq 1$. $\alpha_{XY} = 1$ (resp. $\alpha_{XY} = -1$) means that the variables have an increasing (resp. decreasing) linear relation.

Main Algorithm. Usually, sequential patterns are extracted with a "generate and prune" paradigm. It is composed of two phases: first, candidate $k - sequences$ (sequences of length k) are generated from frequent $(k-1) - sequences$ using a breadth-first search strategy. Then, for each candidate, the support is computed and the ones which do not respect the predefined $supportMin$ are pruned.

In our algorithm, in order to ensure the scalability of the approach, 2-sequences are generated with the preferred genes only. Then, the frequent sequences are extended by adding other genes selected with the Pearson Correlation: "**before**" (resp. "**after**") the first (resp. the last) itemset of the sequence, "**Between**" each itemset, "**In**" each itemset. Then, for each candidate, the support is computed in the two classes and the sequences which do not respect the constraints with $minSupport_{D1}$ and $minSupport_{D2}$ are pruned. By this way, we garantee to generate all sequential patterns.

3 Experiments

In the framework of the PEPS-ST2I Gene Mining project, conducted in collaboration with the laboratory MMDN[1], we mine real data produced by the analysis of DNA microarrays (Affymetrix DNA U133 plus 2.0) to study the Alzheimer disease (AD). This dataset is composed of 14 microarrays corresponding to 2 AD animals and 12 healthy animals. The dataset aims at discovering classification tools for dicrimating between AD animals and healthy animals.

We thus took the genes for which the difference of expression is significative between these two classes by using the SAM (Significant Analysis MicroArray) method, usually used by biologists, which uses the FDR and q-value method. 579 genes were extracted by this process.

To reduce the search space, we use 7 preferred genes which are obtained by mapping the Alzheimer disease pathway in KEGG[2] and the database.
An example of sequential pattern expressed in AD animals and absent in healthy animals is $S =< (MRVI1)(PGAP1)(PLA2R1)(GARNL1)(A2M)(GSK3B) >$ with 0.9 Pearson correlation. Interestingly, those proteins are involved in signaling,

[1] "Molecular mechanisms in neurodegenerative dementias" from the University of Montpellier 2 www.mmdn.univ-montp2.fr

[2] "Pathway" is a generic term corresponding to a graphical representation of interaction between genes. This representation is often used by the biologists to model the relation between genes implied in a specific disease. http://www.genome.jp/kegg/

metabolism and interfere with Alzheimer's disease cellular events. In order to analyse the results, a visual tool has been developed and used by biologists which are outlined some relevant genes to be investigated [12].

4 Related Work

Early work to analyse microarrays were based on statistical methods [2]. They compare the genes expression and they reveal genes (candidate genes) with different behaviours according to several experimental conditions. The SAM analysis, developed by [14], is one such method. Several Machine Learning methods such as [6] or [7] have also been proposed in order to select discriminative genes. Clustering methods have been proposed in order to identify groups of co-expressed genes. One popular method is proposed by [5]. Some association rules methods have also been proposed. [8] extract closed patterns. [3] discover association rules and then classify them into groups such as $R = GeneA, GeneB \rightarrow Cancer$. [10] extract association rules which contain temporal abstraction in order to find temporal relations between specific patterns that the gene may assume over time. [11] propose a new kind of temporal association rules and apply them in the context of DNA microarrays. To our knowledge, there is only one approach dealing with sequential patterns, [13] who extract sequential patterns from microarray datasets ordered by biological situations.

None of these approaches extracts patterns associated to relationships between gene expressions as achieved by our proposal. Note that our approach is not meant to replace existing methods. It is rather viewed as a complementary approach that biologists will use on top of their current tools.

5 Conclusion

In this paper, we have proposed a new approach to discover discriminant sequential patterns in the frame of microarray datasets. It exploits relationships between genes taking into account their ordered expressions and it provides a new material to be investigated by biologists. Due to the specificities of microarray datasets, we have proposed an adapted algorithm DSPAB. Indeed, introducing preferred genes from a knowledge source reduces the search space and ensures more biological meaning. In order to limit the size of results, we use the Pearson correlation as an interest measure. The extraction of discriminant sequential patterns offers a new point of view from gene data to biologists. Further work includes evaluating the accuracy of the obtained discriminant sequential patterns as predictive tool [2]. We will also investigate how a fuzzy approach could improved the biological relevance of the sequential patterns.

References

1. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Eleventh International Conference on Data Engineering, pp. 3–14. IEEE Computer Society Press, Los Alamitos (1995)
2. Bellazzi, R., Zupan, B.: Towards knowledge-based gene expression data mining. Journal of Biomedical Informatics 40(6), 787–802 (2007)

3. Cong, G., Tung, A.K.H., Xu, X., Pan, F., Yang, J.: Farmer: Finding interesting rule groups in microarray datasets. In: SIGMOD Conference 2004, pp. 143–154. ACM, New York (2004)
4. Dong, G., Li, J.: Efficient mining of discriminant patterns: discovering trends and differences. In: KDD 1999: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 43–52. ACM, San Diego (1999)
5. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Science 85(25), 14863–14868 (1998)
6. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 7(6), 673–679 (2001)
7. Li, L., Weinberg, C.R., Darden, T.A., Pedersen, L.G.: Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 17(12), 1131–1142 (2001)
8. Pan, F., Cong, G., Tung, A.K.H., Yang, J., Zaki, M.J.: Carpenter: finding closed patterns in long biological datasets. In: Getoor, L., Senator, T.E., Domingos, P., Faloutsos, C. (eds.) KDD, pp. 637–642. ACM, New York (2003)
9. Piatetsky-Shapiro, G., Tamayo, P.: Microarray data mining: facing the challenges. SIGKDD Explor. Newsl. 5(2) (2003)
10. Sacchi, L., Larizza, C., Magni, P., Bellazzi, R.: Precedence Temporal Networks to represent temporal relationships in gene expression data. Journal of Biomedical Informatics 40(6), 761–774 (2007)
11. Sacchi, L., Larizza, C., Magni, P., Bellazzi, R.: Data mining with Temporal Abstractions: learning rules from time series. Data Min. Knowl. Discov. 15(2), 217–247 (2007)
12. Salle, P., Bringay, S., Teisseire, M., Chakkour, F., Roche, M., Devau, G., Lautier, C., Verdier, J.M.: GeneMining: Identification, Visualization, and Interpretation of Brain Ageing Signatures. In: MIE 2009 Medical Informatics Europe, Sarajevo (2009)
13. Tanasa, D., López, A., Trousse, B.: Extracting Sequential Patterns for Gene Regulatory Expressions Profiles. In: López, J.A., Benfenati, E., Dubitzky, W. (eds.) KELSI 2004. LNCS, vol. 3303, pp. 46–57. Springer, Heidelberg (2004)
14. Tusher, V., Tibshirani, R., Chu, C.: Significance Analysis of microarrays Applied to Ionizing Radiation Response. Proc. Nat'l Academy of Sciences 98(9), 5116–5121 (2001)

The Role of Biomedical Dataset in Classification

Ajay Kumar Tanwani and Muddassar Farooq

Next Generation Intelligent Networks Research Center (nexGIN RC)
National University of Computer & Emerging Sciences (FAST-NUCES)
Islamabad, 44000, Pakistan
{ajay.tanwani,muddassar.farooq}@nexginrc.org

Abstract. In this paper, we investigate the role of a biomedical dataset on the classification accuracy of an algorithm. We quantify the complexity of a biomedical dataset using five complexity measures: correlation-based feature selection subset merit, noise, imbalance ratio, missing values and information gain. The effect of these complexity measures on classification accuracy is evaluated using five diverse machine learning algorithms: J48 (decision tree), SMO (support vector machines), Naive Bayes (probabilistic), IBk (instance based learner) and JRIP (rule-based induction). The results of our experiments show that noise and correlation-based feature selection subset merit – not a particular choice of algorithm – play a major role in determining the classification accuracy. In the end, we provide researchers with a meta-model and an empirical equation to estimate the classification potential of a dataset on the basis of its complexity. This well help researchers to efficiently pre-process the dataset for automatic knowledge extraction.

Keywords: Classification, Complexity Measures, Biomedical Datasets.

1 Introduction

A diverse set of machine learning and data mining algorithms have been proposed to extract useful information from a dataset. But, the learning behavior of all these algorithms is dependent on the complexity of underlying data [1]. Biomedical datasets, in this context, pose a unique challenge to machine learning techniques for classification because of their high dimensionality, multiple classes, noisy data and missing values [2]. Therefore, we advocate the need to separately study the impact of the complexity of biomedical dataset in classification.

In this paper, we systematically investigate the role of biomedical dataset in classification using a number of complexity measures and a diverse set of machine learning algorithms. The empirical study is performed on 31 biomedical datasets publicly available from the UCI Machine Learning repository [3]. The goal is to resolve the uncertainties associated with the complexity of biomedical dataset and the resulting accuracy of classification. The outcome of our study is a novel framework to estimate the classification potential of a biomedical dataset. This will prove useful in understanding the nature of a biomedical dataset for efficient pre-processing and automatic knowledge extraction.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 370–374, 2009.

2 Complexity Measures of Biomedical Datasets

1) Correlation-Based Feature Selection Subset Merit - CfsSubset Merit. Correlation-based feature selection (Cfs) is used to select a subset of attributes that are highly correlated with the class but have low inter-correlation. The correlation between the two attributes A and B with entropies $H(A)$ and $H(B)$ is measured using the *symmetrical uncertainty* [4]:

$$U(A,B) = 2\frac{H(A) + H(B) - H(A,B)}{H(A) + H(B)} \tag{1}$$

where H(A,B) is the joint entropy calculated from the joint probabilities of all combination of attribute values. The merit of a subset formed with correlation-based feature selection M_{cfs} is calculated using [4]:

$$M_{cfs} = \frac{\sum_{i=1}^{N_a} U(A_j, C)}{\sqrt{\sum_{i=1}^{N_a}\sum_{j=1}^{N_a} U(A_i, A_j)}} \tag{2}$$

where N_a is the number of attributes in the set and C is the class attribute. The CfsSubset merit provides a measure of the quality of attributes in a dataset.

2) Noise. Brodley and Friedl characterized noise as the proportion of incorrectly classified instances by a set of trained classifiers [5]. We quantify noise as the sum of all off-diagonal entities where each entity is the minimum of all the corresponding elements in a set of confusion matrices. The advantage of our approach is that we separately identify misclassified instances of every class and only categorize those as noisy which are misclassified by all the classifiers. The percentage of class noise N_o in a dataset with I_n instances can be computed as below:

$$N_o = \left(\frac{1}{I_n}\sum_{i=1}^{N_c}\sum_{j=1}^{N_c} min(C_1(i,j), C_2(i,j).....C_n(i,j))\right)100 \qquad (i \neq j) \tag{3}$$

where C_n is the n^{th} confusion matrix in a set of n classifiers, N_c is the number of classes, $min(C_1(i,j), C_2(i,j).....C_n(i,j))$ is an entity for corresponding i and j that represents minimum number of class instances misclassified by all the classifiers. We have used the same set of classifiers as used for our comparative study to determine percentage of noise levels in the datasets. It is evident from Table 1 that biomedical datasets are generally associated with high noise levels.

3) Imbalance Ratio. We propose the following definition of imbalance ratio I_r to cater for proportion of all class distributions in a dataset:

$$I_r = \frac{N_c - 1}{N_c}\sum_{i=1}^{N_c}\frac{I_i}{I_n - I_i} \tag{4}$$

where I_r is in the range ($1 \leq I_r < \infty$) and I_r = 1 is a completely balanced dataset having equal instances of all classes. I_i is the number of instances of i^{th} class and I_n is the total number of instances.

Table 1. The Table shows: (1) Complexity of datasets quantified in terms of CfsSubset Merit (CfsSub Merit), Noise, Imbalance Ratio (Imb Ratio), Average Information Gain (Info Gain) and Missing Values; (2) Classification accuracies of compared algorithms; bold entry in every row represents the best accuracy

Dataset	Complexity of Dataset					Classifiers					Mean
	CfsSub Merit	Noise	Imb Ratio	Info Gain	Missing Values	J48	SMO	NB	IB*k*	JRIP	
Ann-Thyroid	0.64	0.11	8.37	0.037	0	**99.69**	93.79	95.42	94.12	99.53	96.51 ± 2.89
Breast Cancer	0.70	2.72	1.21	0.451	0.23	94.56	96.71	95.99	**96.99**	95.42	95.94 ± 0.98
Breast Cancer Diagnostic	0.67	2.11	1.14	0.303	0	92.97	**97.89**	92.62	97.19	93.67	94.87 ± 2.48
Breast Cancer Prognostic	0.12	13.64	1.76	0.004	0.06	73.74	**75.76**	67.17	73.23	**75.76**	73.13 ± 3.52
Cardiac Arrhythmia	0.47	11.28	1.57	0.047	0.32	65.49	**70.57**	61.73	58.85	69.25	65.18 ± 4.94
Cleveland-Heart	0.26	17.82	1.37	0.115	0.15	55.45	**59.08**	55.45	**59.08**	53.14	56.43 ± 2.59
Contraceptive Method	0.08	31.98	1.05	0.041	0	52.14	48.20	50.78	48.47	**52.41**	50.40 ± 1.98
Dermatology	0.77	0.82	1.05	0.442	0.06	93.99	95.35	**97.27**	95.63	86.88	93.82 ± 4.05
Echocardiogram	0.65	6.06	1.24	0.084	4.67	**90.84**	86.26	87.02	86.26	**90.84**	88.24 ± 2.39
E-Coli	0.67	6.55	1.25	0.678	0	84.23	**87.20**	85.12	86.01	81.25	84.76 ± 2.25
Haberman's Survival	0.08	16.67	1.57	0.023	0	71.89	73.53	**74.84**	69.28	72.22	72.35 ± 2.07
Hepatitis	0.32	10.97	2.0	0.058	5.67	81.94	**87.10**	83.22	85.16	76.77	82.84 ± 3.91
Horse Colic	0.32	11.96	1.14	0.061	19.39	85.33	83.42	80.43	81.79	**86.96**	83.59 ± 2.62
Hungarian Heart	0.27	13.61	1.74	0.079	20.46	**68.71**	68.02	65.31	66.33	64.63	66.60 ± 1.74
Hyper Thyroid	0.30	0.34	28.81	0.012	2.17	**98.94**	97.77	95.39	97.83	98.49	97.68 ± 1.37
Hypo-Thyroid	0.42	0.54	9.99	0.024	6.74	**99.24**	97.44	97.91	97.28	**99.24**	98.22 ± 0.96
Liver Disorders	0.06	9.86	1.05	0.011	0	**68.70**	58.26	55.36	59.13	64.64	61.22 ± 5.36
Lung Cancer	0.50	21.88	1.02	0.152	0.28	50.00	40.62	**62.50**	40.62	43.75	47.50 ± 9.22
Lymph Nodes	0.41	10.81	1.46	0.138	0	77.03	**86.49**	83.11	83.78	76.35	81.35 ± 4.45
Mammographic Masses	0.33	14.15	1.01	0.193	3.37	82.73	78.88	83.14	80.12	**83.25**	81.62 ± 1.99
New Thyroid	0.74	2.79	1.78	0.602	0	92.09	89.77	**96.74**	93.95	93.02	93.12 ± 2.56
Pima Indians Diabetes	0.16	20.18	1.20	0.064	0	73.83	**77.34**	76.30	73.18	75.13	75.16 ± 1.72
Post Operative Patient	0.05	30.00	1.90	0.016	0.44	**70.00**	**70.00**	67.78	68.89	**70.00**	69.33 ± 0.99
Promoters Genes	0.36	4.72	1.00	0.078	0	81.13	**93.40**	90.57	79.24	83.02	85.47 ± 6.17
Protein Data	0.02	45.48	1.19	0.065	0	**54.52**	**54.52**	**54.52**	**54.52**	**54.52**	54.52 ± 0.00
Sick	0.42	0.71	7.72	0.013	2.24	**98.75**	93.86	92.68	95.96	98.21	95.89 ± 2.65
Splice-Junction Genes	0.48	4.60	1.15	0.022	0	93.05	91.70	**93.90**	79.80	93.20	90.33 ± 5.94
Statlog Heart	0.32	15.19	1.02	0.092	0	76.67	82.59	**84.81**	81.11	77.04	80.44 ± 3.54
Switzerland Heart	0.09	32.52	1.14	0.023	17.07	29.27	39.02	35.77	30.89	**39.84**	34.96 ± 4.74
Thyroid0387	0.42	1.35	2.99	0.091	5.5	**95.76**	77.77	78.42	81.81	93.93	85.54 ± 8.65
VA-Heart	0.07	27.00	1.04	0.023	26.85	34.00	**35.00**	34.00	32.00	30.00	33.00 ±2.00
Mean	**0.36**	**12.53**	**2.97**	**0.13**	**3.73**	$76.99^{(2)}$ ± 19.02	$\mathbf{77.01}^{(1)}$ ± **18.62**	$76.62^{(3)}$ ± 18.28	$75.11^{(5)}$ ± 19.34	$76.53^{(4)}$ ± 18.82	

4) Missing Values. The datasets obtained from clinical databases contain several missing fields as their inherent characteristic. Therefore, we quantify the percentage of missing values in a dataset to study their effect on classification accuracy.

5) Information Gain. Information gain is an information-theoretic measure that evaluates the quality of an attribute in a dataset based on its entropy [4]. We use the average information gain of a dataset to give a measure of the quality of its attributes for classification.

3 Experiments, Results and Discussions

We now present the results of our experiments that we have done to analyze the complexity of 31 biomedical datasets with five different algorithms: J48, SMO, Naive Bayes, IBk and JRIP [2]. We have used the standard implementations of these schemes in Wakaito Environment for Knowledge Acquisition (WEKA) [4] to remove any customized bias in our study. A careful insight into the results in Table 1 helps to draw an

important conclusion: ***the variance in accuracy of classifiers on a particular dataset is significantly smaller compared with the variance in accuracy of the same classifier on different datasets***. This implies that the nature of dataset has a very strong impact on the classification accuracy of a dataset compared to the choice of classifier. In Figure 1, we present the effect of our complexity measures with mean classification accuracy of all algorithms. It is obvious that the percentage of noise in a dataset effectively determines the classification accuracy; the CfsSubset merit is directly proportional to the classification accuracy; the high average information gain of a dataset yields better classification accuracy; the high percentage of missing values significantly degrade the classification accuracy while the imbalance ratio in a dataset has a minor impact on the resulting accuracy.

In order to get better insights in Figure 1, a meta-dataset is created consisting of our five complexity measures as its attributes. The output classification potential of a dataset is categorized into three classes depending upon the classification accuracy: good (greater than 85%), satisfactory (65-85%) and bad (less than 65%). The classification models are extracted using JRIP and J48 with resulting classification accuracies of 80.64% and 77.42%. The meta-model shows that noise and CfsSubset Merit are the two most important attributes in estimating the classification potential of a dataset.

Classification Rules of JRIP

```
(noise >= 17.82) => class=bad (8.0/2.0)
(noise <= 6.06) => class=good (12.0/0.0)
 => class=satisfactory (11.0/1.0)
```

Decision Tree of J48

```
noise <= 4.72: good (12.0)
noise > 4.72
|   CfsSubset <= 0.263: bad (9.0/1.0)
|   CfsSubset > 0.263: satisfactory (10.0/1.0)}
```

To generalize the findings of our study, we map an equation on the obtained results of noise (N_o) and CfsSubset Merit (M_{cfs}) to determine the classification potential (C_P)

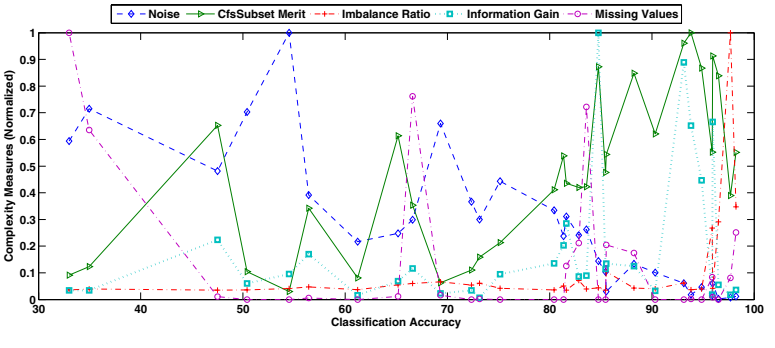


Fig. 1. Effect of Complexity Measures on Classification Accuracy

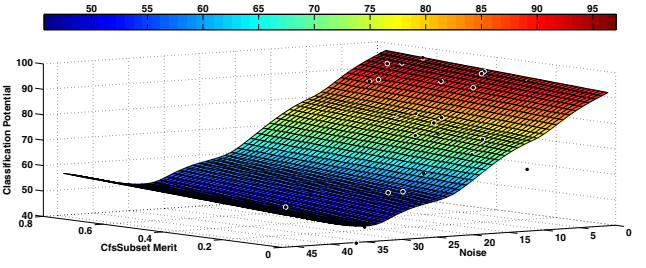


Fig. 2. Classification Potential as Function of Noise and CfsSubset Merit

of a dataset (see Figure 2). The equation is obtained using fitness criteria of lowest sum of squared absolute error:

$$C_P = 98.66 - 1.22 * N_o - \frac{1.43}{M_{cfs}} - 0.06 * {N_o}^2 - \frac{0.09}{{M_{cfs}}^2} + 0.19 * \frac{N_o}{M_{cfs}} \qquad (5)$$

4 Conclusion

In this paper, we have quantified the complexity of a biomedical dataset in terms of correlation-based feature subset merit, noise, imbalance ratio, missing values and information gain. The effect of complexity on classification accuracy is evaluated using five well-known diverse algorithms. The results show that the complexity measures – noise and CfsSubset merit – predominantly determines the classification accuracy of a biomedical dataset rather than the choice of a particular classifier. The major contribution of this paper is a novel methodology for estimating the classification potential of a dataset using its complexity measures.

References

1. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 289–300 (2002)
2. Tanwani, A.K., Afridi, J., Shafiq, M.Z., Farooq, M.: Guidelines to select machine learning scheme for classifcation of biomedical datasets. In: Pizzuti, C., Ritchie, M.D., Giacobini, M. (eds.) EVOBIO 2009. LNCS, vol. 5483, pp. 128–139. Springer, Heidelberg (2009)
3. UCI repository of machine learning databases, University of California-Irvine, Department of Information and Computer Science, http://www.ics.uci.edu/~mlearn/MLRepository.html
4. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
5. Brodley, C.E., Friedl, M.A.: Identifying mislabeled training data. Journal of Artificial Intelligence Research 11, 131–167 (1999)

Online Prediction of Ovarian Cancer

Fedor Zhdanov, Vladimir Vovk, Brian Burford, Dmitry Devetyarov, Ilia Nouretdinov, and Alex Gammerman

Computer Learning Research Centre, Department of Computer Science
Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK
fedor@cs.rhul.ac.uk

Abstract. In this paper we apply computer learning methods to the diagnosis of ovarian cancer using the level of the standard biomarker CA125 in conjunction with information provided by mass spectrometry. Our algorithm gives probability predictions for the disease. To check the power of our algorithm we use it to test the hypothesis that CA125 and the peaks do not contain useful information for the prediction of the disease at a particular time before the diagnosis. It produces p-values that are less than those produced by an algorithm that has been previously applied to this data set. Our conclusion is that the proposed algorithm is especially reliable for prediction the ovarian cancer on some stages.

Keywords: Online prediction, aggregating algorithm, ovarian cancer, mass spectrometry, proteomics.

1 Introduction

Using the antigen CA125 significantly improves the quality of diagnosis of ovarian cancer, though sometimes CA125 signal appears too late to make use of it.

We consider prediction in *triplets*: each *case sample* (sample of the patient with the disease) is accompanied by two samples from healthy individuals, *matched controls*, which are chosen to be as close as possible to the case sample with respect to several attributes. We have 179 triplets in which case samples are taken from 104 individuals. Each triplet is assigned a *time-to-diagnosis* defined as the time of taking the sample to the moment of diagnosis of the case sample in this triplet. We predict the case sample by values of CA125 and intensity of mass-spectrometry peaks. This framework was first described in [1].

For our research we use an ovarian cancer data set (see [2]) processed by the authors of [3]. We combine decision rules proposed in [3] by using an online prediction algorithm[1] described in [5] and thus get our own decision rule. To estimate classification accuracy, we convert probability predictions into strict predictions by the *maximum rule*: we assign weight 1 to the labels with maximum predicted probability within a triplet, weight 0 to the labels of other samples, and

[1] A survey of online prediction can be found in [4].

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 375–379, 2009.

then normalize the assigned weights. The detailed description of our approach can be found in [6].

The paper is organized as follows. In Section 2 we describe methods we use to give predictions. We explain our experiments and results in Section 3. Section 4 concludes the paper.

2 Online Prediction Framework and Aggregating Algorithm

The mathematical framework used in this paper is called prediction with expert advice. This is an online framework, where at each time step experts give their predictions about an event, and the learner's goal is to combine their predictions in such a way that his loss accumulated over several time steps is as close as possible to the loss accumulated by the best expert over these time steps.

Let Ω be a finite and non-empty set of outcomes, $\Gamma := \mathcal{P}(\Omega)$ be the set of predictions, defined to be the set of all probability measures on Ω. The Brier loss function $\lambda : \Omega \times \Gamma \to [0, \infty)$ (see [7]) is defined by

$$\lambda(\omega, \gamma) = \sum_{o \in \Omega} (\gamma\{o\} - \delta_\omega\{o\})^2 . \tag{1}$$

Here $\delta_\omega \in \mathcal{P}(\Omega)$ is the probability measure concentrated at ω: $\delta_\omega\{\omega\} = 1$ and $\delta_\omega\{o\} = 0$ for $o \neq \omega$. We interpret ω as the index of the diseased patient in a triplet. Let us denote the Brier cumulative loss of the learner at a time step N by L_N, and the cumulative loss of the k-th expert at this step by L_N^k. The theoretical bound (see [5] for the proof) for the cumulative loss of the Aggregating Algorithm (AA) at a prediction step N is

$$L_N(\mathrm{AA}) \leq L_N^k + \ln K \tag{2}$$

for any expert k, where the number of experts equals K. The way AA makes predictions is described as Algorithm 1.

Algorithm 1. Aggregating algorithm for the Brier game

$w_0^k := 1$, $k = 1, \ldots, K$.
for $N = 1, 2, \ldots$ **do**
  Read the experts' predictions γ_N^k, $k = 1, \ldots, K$.
  Set $G_N(\omega) := -\frac{1}{\eta} \ln \sum_{k=1}^K w_{N-1}^k e^{-\eta \lambda(\omega, \gamma_N^k)}$, $\omega \in \Omega$.
  Solve $\sum_{\omega \in \Omega} (s - G_N(\omega))^+ = 2$ in $s \in \mathbb{R}$.
  Set $\gamma_N\{\omega\} := (s - G_N(\omega))^+ / 2$, $\omega \in \Omega$.
  Output prediction $\gamma_N \in \mathcal{P}(\Omega)$.
  Read observation ω_N.
  $w_N^k := w_{N-1}^k e^{-\eta \lambda(\omega_N, \gamma_N^k)}$.
end for

3 Experiments

This section describes two experiments.

3.1 Probability Prediction of Ovarian Cancer

For each sample we calculate values

$$u(v, w, p) = v \ln C + w \ln I_p, \tag{3}$$

where C is the level of CA125, I_p is the intensity of the p-th peak, $p = 1, \ldots, 67$, $v \in \{0, 1\}, w \in \{-2, -1, -1/2, 0, 1/2, 1, 2\}$. For each triplet we predict as a case the sample with a maximum value of $u(v, w, p)$ and use the obtained predictions as experts' predictions. The total number of experts is 537: $402 = 6 \times 67$ for $v = 1, w \neq 0$, $134 = 2 \times 67$ for $v = 0$, and 1 for $v = 1, w = 0$. For each triplet our algorithms gives the probability of being diseased for each person in the triplet. The evolution of the cumulative Brier loss of all the experts minus the cumulative loss of our algorithm over all the 179 triplets is presented in Figure 1. The x-axis presents triplets in the chronological order. We can see from Figure 1

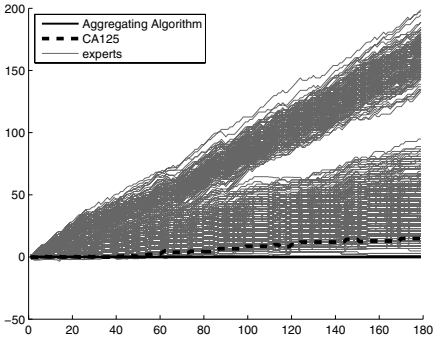


Fig. 1. Difference between the cumulative losses of different predictors and AA over all the triplets

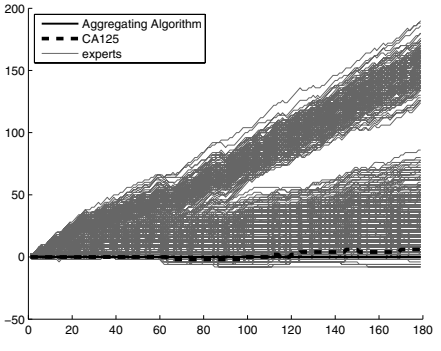


Fig. 2. Difference between the cumulative losses of different predictors and categorical AA over all the triplets

that the Aggregating Algorithm predicts better than most experts in our class after about 54 triplets, in particular better than CA125. At the end the AA is better than all the experts.

Further we use the maximum rule to convert probability predictions into strict predictions, and refer to this algorithm as to the *categorical AA*. If we calculate the Brier loss, we get Figure 2. We can see that the categorical AA still beats CA125 at the end in the case where it gives strict predictions.

3.2 Prediction on Different Stages of the Disease

We consider 6-month time periods with starting point $t = 0, 1, \ldots, 16$ months before diagnosis. For each time period we select only those triplets from the corresponding time period, the latest for each case patient if there are more than one. In this experiment we use different initial distribution on experts: the combinations with peak 1 have initial weight $1 = d^0$, the combinations with peak 2 have initial weight d^{-1}, etc. We empirically choose the coefficient for this distribution $d = 1.2$, and the parameter η for the AA $\eta = 0.65$. Figure 3 shows the fraction of erroneous predictions of the diseased patient made by different algorithms at each time period. We calculate p-values for testing the

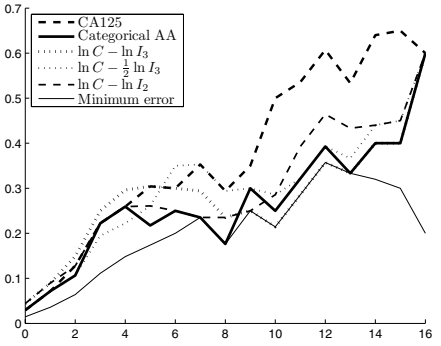


Fig. 3. Fraction of erroneous predictions over different time periods of different predictors

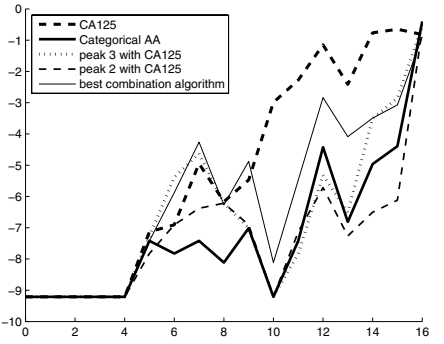


Fig. 4. The logarithm of p-values for different algorithms

null hypothesis to check that our results are not accidental. The p-value can be defined as the value taken by a function p satisfying

$$\forall \delta \text{ Probability}(p \leq \delta) \leq \delta$$

for all $\delta \in (0, 1)$ under the null hypothesis. For each time period we make 10^4 trials as follows. We vary $d \in \{1.1, 1.2, \ldots, 2.0\}$ and $\eta \in \{0.1, 0.15, \ldots, 1.0\}$. For each trial we randomly permute labels in each triplet, find the error of the best algorithm among algorithms with different parameters d, η, and calculate the proportion of the trials when the error is less than or equal to the error of the best algorithm on true labels. The detailed procedure is described in [6]. The logarithms of p-values for different algorithms are presented in Figure 4. We include the algorithm proposed in [3] (it appears in Figure 4 as the "best combination algorithm") and p-values for peaks 3 and 2 calculated there. As we can see, our algorithm has small p-values, comparable with or even smaller than p-values for other algorithms. And our algorithm has fewer adjustments than other algorithms presented in Figure 4. If we choose 5% as our significance level,

then CA125 classification is significant up to 9 months in advance of diagnosis (the p-values are less than 5%). At the same time, the results for peaks combinations and for AA are significant for up to 15 months. Thus we can say that instead of choosing one particular combination, we should use the Aggregating Algorithm to mix all the combinations.

4 Conclusion

Our results show that the CA125 criterion, which is a current standard for the detection of ovarian cancer, can be outperformed especially at early stages. We have proposed a way to give probability predictions for the ovarian cancer in the triplet setting. We found that the Aggregating Algorithm we use to mix combinations predicts better than almost any combination at different stages before diagnosis. Our test statistic produces p-values that are never worse than the p-values produced by the statistic proposed in [3].

Acknowledgments

We would like to thank Mike Waterfield, Ali Tiss, Celia Smith, Rainer Cramer, Alex Gentry-Maharaj, Rachel Hallett, Stephane Camuzeaux, Jeremy Ford, John Timms, Usha Menon, and Ian Jacobs for sharing this data set with us and useful discussions of experiments and results. This work has been supported by EPSRC grant EP/F002998/1, EU FP7 grant, MRC grant G0301107, ASPIDA grant from the Cyprus Research Promotion Foundation, Veterinary Laboratories Agency of DEFRA grant, and EPSRC grant EP/E000053/1.

References

1. Gammerman, A., et al.: Serum Proteomic Abnormality Predating Screen Detection of Ovarian Cancer. The Computer Journal (2008) bxn021
2. Menon, U., et al.: Prospective study using the risk of ovarian cancer algorithm to screen for ovarian cancer. J. Clin. Oncol. 23(31), 7919–7926 (2005)
3. Devetyarov, D., et al.: Analysis of serial UKCTOCS-OC data: discriminating abilities of proteomics peaks. Technical report (2009), http://clrc.rhul.ac.uk/projects/proteomic3.htm
4. Cesa-Bianchi, N., Lugosi, G.: Prediction, learning, and games. Cambridge University Press, Cambridge (2006)
5. Vovk, V., Zhdanov, F.: Prediction with expert advice for the Brier game. In: ICML 2008: Proceedings of the 25th International Conference on Machine Learning, pp. 1104–1111. ACM, New York (2008)
6. Zhdanov, F., et al.: Online prediction of ovarian cancer. Technical report, arXiv:0904.1579v1 [cs.AI], arXiv.org e-Print archive (2009)
7. Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly Weather Review 78, 1–3 (1950)

Prediction of Mechanical Lung Parameters Using Gaussian Process Models

Steven Ganzert[1,⋆], Stefan Kramer[2], Knut Möller[3], Daniel Steinmann[1], and Josef Guttmann[1]

[1] University Hospital Freiburg, Department of Experimental Anesthesiology, D-79106 Freiburg, Germany
steven.ganzert@uniklinik-freiburg.de
[2] Technische Universität München, Institut für Informatik / I12, D-85748 Garching b. München, Germany
[3] Furtwangen University, Department of Biomedical Engineering, D-78054 Villingen-Schwenningen, Germany

Abstract. Mechanical ventilation can cause severe lung damage by inadequate adjustment of the ventilator. We introduce a Machine Learning approach to predict the pressure-dependent, non-linear lung compliance, a crucial parameter to estimate lung protective ventilation settings. Features were extracted by fitting a generally accepted lumped parameter model to time series data obtained from ARDS (adult respiratory distress syndrome) patients. Numerical prediction was performed by use of Gaussian processes, a probabilistic, non-parametric modeling approach for non-linear functions.

1 Medical Background and Clinical Purpose

Under the condition of mechanical ventilation a high volume distensibility – or *compliance* C – of the lung is assumed to reduce the mechanical stress to the lung tissue and hence irreversible damage to the respiratory system. A common technique to determine the *maximal compliance* C_{max} inflates the lung with almost zero flow (so-called 'static' conditions) over a large pressure-volume (PV) range (inspiratory capacity). The inspiratory limb of the corresponding PV curve typically shows a sigmoid shape. As C is determined by the change of respiratory volume V divided by the change of applied respiratory pressure P, i.e. $C = \Delta V/\Delta P$, C_{max} is found at the curve interval with the steepest slope. This is supposed to be the optimal PV range for lung protective ventilation [1] (see Fig. 1a). Within Super-syringe maneuvers [1] rapid flow interruptions are iteratively performed after consecutive, equally sized volume inflations. These flow interruptions reveal characteristic stress relaxation curves, exponentially approximating the plateau pressure level P_{plat} (see Fig. 1b, insert). The spring-and-dashpot model [2] is assumed to represent the viscoelastic behavior of the lung tissue. Fitting this model to flow interruption data provides the four parameters C, *respiratory resistance* R, *viscoelastic compliance* C_{ve} and *resistance* R_{ve}

⋆ Corresponding author.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 380–384, 2009.

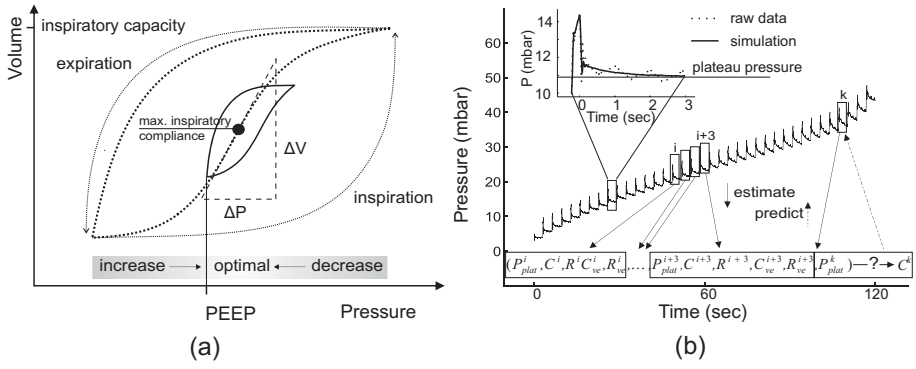


Fig. 1. (a) Schematized PV loops measured under static (dotted large loop: inspiratory flow ≈ 0 ml/sec) and dynamic (straight small loop: inspiratory flow ≫ 0 ml/sec) conditions. For the dynamic loop, the PEEP (positive end-expiratory pressure) is optimized. The pressure gap between the static and the dynamic curve is effected by the flow induced pressure fraction under dynamic conditions. (b) Sample pressure time series for Super-syringe maneuver: Based on the estimated values of 4 consecutive occlusion steps i to $i+3$ the compliance C was predicted for the plateau pressure at step k. Insert: Sample volume step and airway occlusion of raw data and model simulation.

which are non-linearly related to P_{plat}. In its entirety these parameters numerically reflect the mechanical status of the respiratory system. The purpose of this study is to model statistically the pressure-dependent non-linear behavior of the compliance. The individualized prediction of this mechanical lung parameter could assist the physician when individually adjusting the applied pressure level in order to reduce the risk of ventilator induced lung injury.

2 Gaussian Processes and Modeling Task

Non-linear regression problems are generally modeled by parametrizing a function $f(x)$ with parameters w to $f(x;w)$. Gaussian Processes (GPs) [3] introduce a probabilistic approach to this field: the parametrized function can be rewritten as a linear combination of non-linear basis functions $\phi_h(x)$, i.e. $f(x;w) = \sum_{h=1}^{H} \omega_h \phi_h(x)$. Under the assumption that the distribution of w is Gaussian with zero mean, the linear combination of the parametrized basis functions produces a result which is distributed Gaussian as well. Assuming that the target values differ by additive Gaussian noise from the function values, the prior probability of the target values is also Gaussian. The linear combination of parametrized basis functions can be replaced by the covariance matrix of the function values and inference from a new feature observation is done by evaluation of this matrix. A prior assumption on our modeling task is that the hypothesis space consists of the derivatives of sigmoid-like shaped functions ($C = \Delta V/\Delta P$) (see Fig. 1a). As all measured patient data sets imply general as well as individual characteristics

of the ARDS lung, the shape of these functions ought to be distributed according to a prior probability. Therefore, we hypothesized that our modeling task would benefit from the probabilistic modeling of (possibly) non-linear functions as provided by GP modeling. In the present study inferences were made from the status of the respiratory system at a distinct plateau pressure range to the compliance value at a different pressure level:

(i) Prediction of the compliance-pressure curve covering the range of the inspiratory capacity.
(ii) Prediction of C_{max} and its corresponding plateau pressure $P_{plat}(C_{max})$.
(iii) Prediction, if the pressure level should be increased, decreased or retained in order to achieve C_{max} (which we refer to as *trend* in the following).

3 Materials and Methods

Raw data: 18 mechanically ventilated patients suffering from ARDS were included [4,5]. Automatized Super-syringe maneuvers were performed. During a single maneuver, the ventilatory system repetitively applied volume steps of 100 ml with constant inspiratory airflow rates. At the end of each volume step, airflow was interrupted for 3 seconds (see Fig. 1b).

Feature extraction and preprocessing: For each flow interruption step k, the attributes C^k, R^k, C^k_{ve} and R^k_{ve} were estimated by fitting a spring-and-dashpot model [2] to the data and P^k_{plat} was measured (see Fig. 1b). The initial system status was represented by the fitted parameter and measured plateau pressure values of 4 consecutive steps i to $i+3$. For all possible states that can be determined by such quadripartite steps the compliance was predicted as target value C^k for all plateau pressure levels P^k_{plat}. Therefore the feature samples consisted of the 22-tupel ($P^{i\ldots i+3}_{plat}$, $C^{i\ldots i+3}$, $R^{i\ldots i+3}$, $C^{i\ldots i+3}_{ve}$, $R^{i\ldots i+3}_{ve}$, P^k_{plat}, C^k) (see Fig. 1b). The M5P algorithm [6] generates a combination of decision trees with linear regression functions as model trees. M5P was evaluated as reference method for prediction task (i). Two experimental settings were evaluated:

1. Separately for each single patient data set (i.e. Super-syringe measuring) the three prediction tasks were performed. This experimental setting should confirm the suitability of GP models for the given data.
2. To investigate the practical applicability of the approach, training and test set were repetitively built for each patient data set. For each run, the training set consisted of all patient-data sets except one, which was used as test set.

Performance measures: For prediction task (i), the performance was measured by the correlation coefficient (CC) of the model prediction. Other performance measures like error estimations were supposed to be inadequate as the raw data showed high variability. For task (ii) the percentage difference between the maximum compliance (respectively its corresponding P_{plat}) determined from the raw data and the predicted maximum compliance (respectively P_{plat}) was

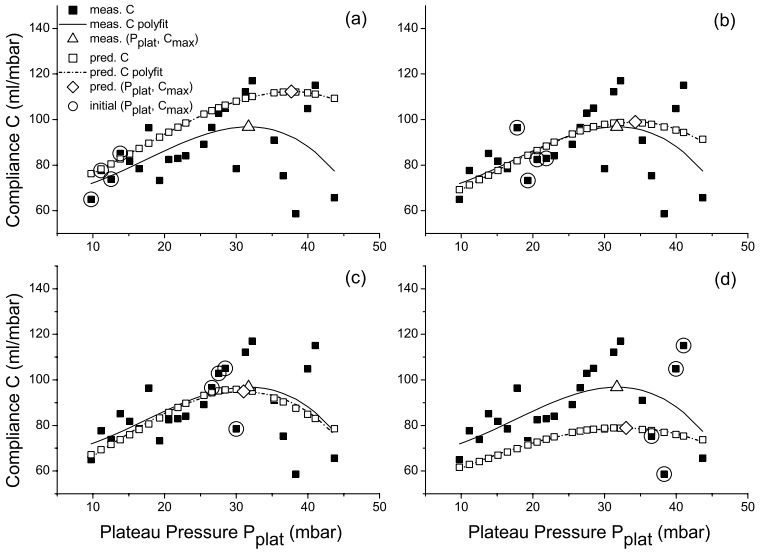


Fig. 2. Sample results for one patient in experimental setting (2). Predictions of C (*pred.* C) are based on the respiratory status at low (a), intermediate (b), (c) and high (d) levels of P_{plat} (*initial* (P_{plat}, C_{max})). The variability of the measured (i.e., fitted) C values (*meas.* C) is clearly exhibited. Measured and predicted C_{max} and the corresponding P_{plat} (*meas.* (P_{plat}, C_{max}), *pred.* (P_{plat}, C_{max})) are determined by polynomial fits (*meas.* C $polyfit$, *pred.* C $polyfit$).

calculated. Task (iii) was evaluated by the percentage of correct predictions of the *trend*. Results are given as $mean \pm sd$. For full details of materials and methods, see [7].

4 Results and Discussion

Experimental setting (1): (i) Prediction of the compliance curve by GP modeling reached an averaged CC of 0.78 ± 0.16, the reference Method (M5P) an averaged CC of 0.92 ± 0.23. (ii) While the predicted maximum compliance C_{max} averagely differed with $9.7 \pm 6.5\%$ from C_{max} estimated from the raw data, $P_{plat}(C_{max})$ differed with an average of $5.4 \pm 10.6\%$. (iii) The prediction, if the pressure level should be increased, decreased or retained was correctly answered in $93.2 \pm 11.1\%$.

Experimental setting (2): (i) The CC of the learned model had an average of 0.34 ± 0.24 for GP modeling and 0.18 ± 0.22 for the M5P algorithm. (ii) Predicted C_{max} differed with an average of $34.3 \pm 34.3\%$ and predicted $P_{plat}(C_{max})$ with $40.7 \pm 70.1\%$ from the maximum compliance and corresponding pressure values derived from the raw data. (iii) Prediction of the trend for the pressure correction was in 2/3 of the cases ($66.3 \pm 30.3\%$) correct. For full details of the results, see [7].

Performance within experimental setting (1) confirmed that GP modeling is basically suitable for the present task and problem representation. As hypothesized, the compliance-pressure curves were adequately modeled, having slopes of the first derivative of a sigmoid-like function (see Fig. 2). Differentiated characteristics for the individual patient datasets were expressed in differing curve slopes. Comparing the results of GP modeling and M5P for prediction task (i) within settings (1) and (2) leads to the assumption, that the M5P tends more to overfitting than GP modeling. This was perhaps down to the fact that for the present problem the modeling of functions might provide a higher degree of abstraction and reduce the impact of noise. Nevertheless, individual compliance curves for new observations according to setting (2) showed rather poor results. While the prediction of C_{max} and $P_{plat}(C_{max})$ (task ii) as well as the prediction of the correct *trend* for the pressure correction (task iii) showed failure rates below 10% in setting (1), which might be sufficiently precise for an indication in medical practice, the results again were impaired within setting (2). Predictions with divergences of more than 30% for C_{max} and $P_{plat}(C_{max})$ and failure rates in a similar range for trend prediction provide at most a rough estimates. This implies that learning an individualized model might require an individualized feature selection.

To the best of our knowledge, this is the first time that mechanical lung parameters have been predicted by a statistical modeling approach. The results indicate that the combination of classical model fitting and statistical modeling is generally capable of solving this task. Nevertheless an individualized feature selection as pre-processing step should be brought into focus in future efforts.

References

1. Matamis, D., Lemaire, F., Harf, A., Brun-Buisson, C., Ansquer, J.C., Atlan, G.: Total respiratory pressure-volume curves in the adult respiratory distress syndrome. Chest 86, 58–66 (1984)
2. Jonson, B., Beydon, L., Brauer, K., Mansson, C., Valind, S., Grytzell, H.: Mechanics of respiratory system in healthy anesthetized humans with emphasis on viscoelastic properties. J. Appl. Physiol. 75, 132–140 (1993)
3. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, MA (2006)
4. Stahl, C.A., Möller, K., Schumann, S., Kuhlen, R., Sydow, M., Putensen, C., Guttmann, J.: Dynamic versus static respiratory mechanics in acute lung injury and acute respiratory distress syndrome. Crit. Care Med. 34, 2090–2098 (2006)
5. Ganzert, S., Guttmann, J., Kersting, K., Kuhlen, R., Putensen, C., Sydow, M., Kramer, S.: Analysis of respiratory pressure-volume curves in intensive care medicine using inductive machine learning. Artif. Intell. Med. 26, 69–86 (2002)
6. Quinlan, J.R.: Learning with continuous classes. In: Proceedings of the Australian Joint Conference on Artificial Intelligence, pp. 343–348. World Scientific, Singapore (1992)
7. Ganzert, S., Kramer, S., Möller, K., Steinmann, D., Guttmann, J.: Prediction of mechanical lung parameters using Gaussian process models. Technical report, TUM-I0911, Fakultät für Informatik, TU München (2009)

Learning Approach to Analyze Tumour Heterogeneity in DCE-MRI Data During Anti-cancer Treatment

Alessandro Daducci[2], Umberto Castellani[1], Marco Cristani[1], Paolo Farace[2], Pasquina Marzola[2], Andrea Sbarbati[2], and Vittorio Murino[1]

[1] VIPS lab, University of Verona, Italy
[2] Department of Morphological-Biomedical Sciences, Section of Anatomy and Histology, University of Verona, Italy

Abstract. The paper proposes a learning approach to support medical researchers in the context of in-vivo cancer imaging, and specifically in the analysis of Dynamic Contrast-Enhanced MRI (DCE-MRI) data. Tumour heterogeneity is characterized by identifying regions with different vascular perfusion. The overall aim is to measure volume differences of such regions for two experimental groups: the treated group, to which an anticancer therapy is administered, and a control group. The proposed approach is based on a three-steps procedure: (i) robust features extraction from raw time-intensity curves, (ii) sample-regions identification manually traced by medical researchers on a small portion of input data, and (iii) overall segmentation by training a Support Vector Machine (SVM) to classify the MRI voxels according to the previously identified cancer areas. In this way a non-invasive method for the analysis of the treatment efficacy is obtained as shown by the promising results reported in our experiments.

1 Introduction

In the context of cancer imaging, machine learning techniques are becoming important to automatically isolate areas of interest characterized by heterogeneous tumoural tissues. In particular, the identification of tumour heterogeneity is crucial for diagnosis and therapy assessment. In this paper, tumour morphology and functional perfusion are obtained by Dynamic Contrast Enhanced MRI (DCE-MRI) techniques. We propose a learning-by-example approach[1] to classify tumoral regions characterized by heterogeneous vascular perfusion. In fact, DCE-MRI techniques represent noninvasive ways to assess tumour vasculature, that are accepted surrogate markers of tumour angiogenesis [2]. Data are analyzed with the aim of investigating the volume changes in the identified regions for both untreated and treated tumours. The proposed analysis is based on three main phases: (i) features extraction from raw time-intensity curves, (ii) representative tumour areas identification, and (iii) overall voxel-by-voxel classification. In the first phase, few robust features that compactly represent the response of

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 385–389, 2009.

the tissue to the DCE-MRI analysis are computed. The second step provides a manual identification of tumour samples that are representative of the typical tumour aspects. Such samples are carefully and manually chosen by medical researchers on a small portion of input data by observing the different behavior of the time-intensity signals within different kind of tumoural regions (i.e., necrotic or still alive zones). Finally, in the third step, a Support Vector Machine (SVM) is trained to classify voxels according to the regions (i.e., typologies of tumour tissue) defined by the previous phase. In this way, the SVM is able to automatically detect the most discriminative characteristics of the manually identified regions by extending such capability to classify unseen subjects.

Several works are based on the use of machine learning techniques for DCE-MRI tumour analysis [3,4,5,6]. In [3], a visual data-mining approach is proposed to support the medical researchers in tumoral areas characterization by clustering data according to the transendothelial permeability (kPS) and fractional plasma volume (fPV). Although kPS and fPV are accepted estimate of tissue vasculature, their instability under small perturbation of the chosen pharmacokinetics model was proved [4,5]. Therefore, different works are addressed the idea of analyzing directly the raw signals by exploiting possible other compact parameters of the curve shapes. As an example, in [4] the raw signals of the DCE-MRI voxels are analyzed in the context of musculoskeletal tissue classification, where the classification is carried out by introducing a thresholding approach. In [5] the authors propose the use of the Mean Shift algorithm [1] for the clustering of breast DCE-MRI lesions. In particular, voxels are clustered according to the area under the curve feature. Since the results are over-segmented, an iterative procedure is introduced to automatically select the clusters which better represent the tumour. Similarly, in our previous work [6], tumoral regions are characterized by combining Mean Shift clustering[1] with Support Vector Machine (SVM) classification. In the present work we extend the basic framework proposed in [6], to asses the treatment efficacy of anticancer therapies.

2 Materials and Methods

Tumours were induced by subcutaneous injection of human carcinoma cells in nude mice ($n = 11$). Ten days after cells injection animals were randomly assigned to the treated ($n = 6$) and control group ($n = 5$). Animals belonging to the treated and control group received an experimental drug and vehicle, respectively, for a period of 7 days. All animals were observed by MRI before (time T_0) and after the treatment (time T_1). A further group of mice ($n = 5$) bearing the same kind of tumour was used in the training step of the classification procedure. Animals were examined using DCE-MRI as described [2]. To account for tumour heterogeneity, seven classes have been fixed by combining a-priori knowledge of medical experts with the observation of signal shape behaviors (see `A-G` in Figure 1):

– **Classes `A`, `D`**, are characterized by a contrast agent wash-out (i.e., clear defined peak followed by a decrease). These regions correspond to very active

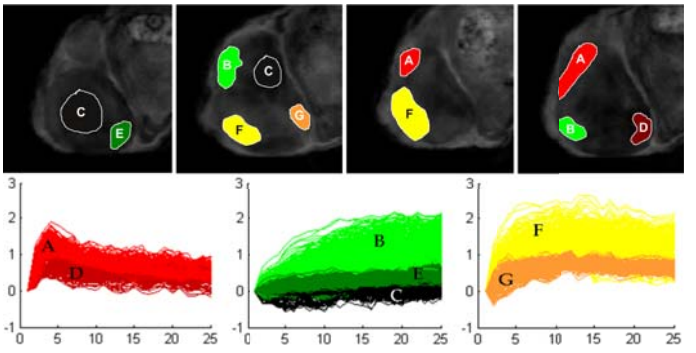


Fig. 1. Sample regions used for SVM training. Seven classes (A-G) were chosen. Source data used to select the sample regions (TOP), and the relative DCE-MRI curves of the whole training set are reported (BOTTOM).

areas where the tumor is still alive. Regions A and D differs by the value of maximum intensity.
- **Classes** B, E, reveal a contrast agent accumulation (i.e., increasing trend). Typically, the activity in such areas are very few by evidencing the transition of the tumor toward a necrotic state.
- **Class** C, contains voxels with negligible enhancement, presumably due to necrotic tissue.
- F **and** G, have been introduced to account for intermediate patterns (i.e., initial increasing trend followed by a plateau phase). These regions are likely to correspond to tumor transitions from hight active to low active state.

By following the proposed pipeline, few and stable signal features are identified to model the different DCE-MRI curve class. In particular, the following curve characteristics are chosen: time-to-peak (TTP), peak value (PEAK), area under curve (AUC), initial area under curve (AUCTTP), and washout rate (WR).

Therefore, in order to apply a learning-by-example approach, several samples of each identified class need to be fed to the classifier. As mentioned above, such phase is carried out manually by medical experts. Figure 1(TOP) shows some representative regions which are used to build the training set. In Figure 1(BOTTOM) the signals curve of the whole selected samples are reported. Signals are colored according to their respective class by evidencing the expected curve shape. A binary Support Vector Machine (SVM) classifier[1] is used to distinguish among the several tumoral tissue classes. In particular, the SVM is able to automatically detect the most discriminative characteristics of the manually identified regions therefore allowing also the classification of new subjects. Finally, in each tumour, the percentage volumes covered by each of the seven classes have been then calculated to evaluate the time-dependent changes in control and treated tumours. The values obtained have been averaged over the experimental groups and statistically compared by paired t-test.

3 Results

The rate of tumour growth is strongly affected by the treatment; in fact average tumour volume, as determined by MR images, increased from $575 \pm 104\ mm^3$ to $1821 \pm 191\ mm^3$ in the control group and from $553 \pm 187\ mm^3$ to $788 \pm 227\ mm^3$ in the treated group. Figure 2 shows percentage volumes of the different classes, averaged over the whole experimental group. In the control group there is a significant ($p < 0.05$) increase of the percentage volume covered by the classes `C` and `E` (i.e., the less enhancing portions of the tumours). Concomitantly, a significant ($p < 0.01$) decrease of the class `F` is observed. An increase in the

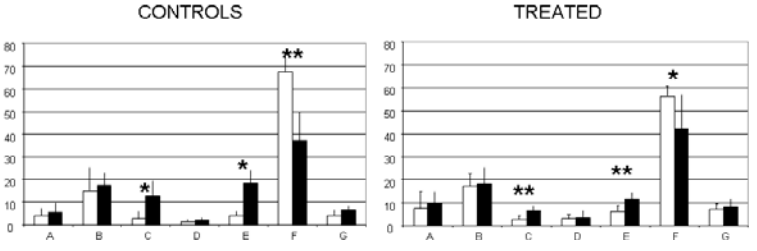


Fig. 2. Percentage volume attributed to each of the seven classes `A-G` averaged over the different experimental groups (control and treated). White and black bars represent values at time T_0 and T_1 respectively. Asterisks indicate t-test significance at a 0.05 (*) and 0.01 (**).

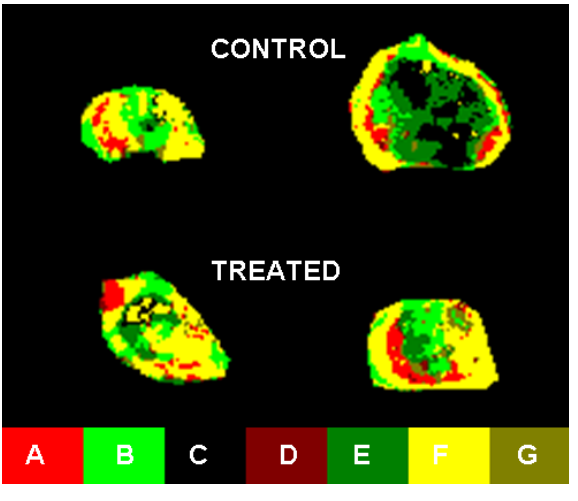


Fig. 3. Segmentation of tumour images in representative control and treated tumours at time T_0 (left) and at time T_1 (right). The colorbar shows the colors used to identify the segmented classes.

scarcely enhanced tissue (necrotic tissue) is typically observed during fast tumour growth. The increase of this tissue is less pronounced in treated tumours as expected from their reduced rate of growth. The percentage volume attributed to `A` and `D` classes (wash-out regions that correspond to well vascularized tissue) is not significantly affected by treatment (or normal tumour growth) in agreement with the fact that the biological target of the herein investigated therapeutic treatment is represented by tumour cells and not by vasculature.

Figure 3 shows segmentation of tumour images obtained with SVM in two representative animals (vehicle and drug treated) before and after the treatment. The substantial increase in the necrotic portion of the control tumour, typical of fast growing tumours, can be visually appreciated.

4 Conclusions

In this paper we emphasize the use of machine learning techniques as a mean to produce automatic and meaningful segmentation results in the quantitative evaluation of DCE-MRI data. The proposed approach permits the computation of percentage tumour volumes of above defined regions and to follow their modifications during the treatment with an experimental drug. Our results suggest that this approach can be useful in the analysis of heterogeneous tumour tissues and of their response to therapies.

References

1. Duda, R., Hart, P., Stork, D.: Pattern Classification, 2nd edn. John Wiley and Sons, Chichester (2001)
2. Marzola, P., Degrassi, A., Calderan, L., Farace, P., Crescimanno, C., Nicolato, E., Giusti, A., Pesenti, E., Terron, A., Sbarbati, A., Abrams, T., Murray, L., Osculati, F.: In vivo assessment of antiangiogenic activity of su6668 in an experimental colon carcinoma model. Clin. Cancer Res. 2(10), 739–750 (2004)
3. Castellani, U., Cristani, M., Combi, C., Murino, V., Sbarbati, A., Marzola, P.: Visual MRI: Merging information visualization and non-parametric clustering techniques for mri dataset analysis. Artificial Intelligence in Medicine 44(3), 171–282 (2008)
4. Lavinia, C., de Jongea, M., Van de Sandeb, M., Takb, P., Nederveena, A.J., Maas, M.: Pixel-by-pixel analysis of DCE MRI curve patterns and an illustration of its application to the imaging of the musculoskeletal system. Magnetic Resonance Imaging 25, 604–612 (2007)
5. Stoutjesdijk, M.J., Veltman, J., Huisman, M., Karssemeijer, N., Barents, J., et al.: Automatic analysis of contrast enhancent in breast MRI lesions using Mean Shift clustering for roi selection. Journal of Magnetic Resonance Imaging 26, 606–614 (2007)
6. Castellani, U., Cristani, M., Daducci, A., Farace, P., Marzola, P., Murino, V., Sbarbati, A.: DCE-MRI data analysis for cancer area classification. Methods of information in medicine 48 (2009)

Predicting the Need to Perform Life-Saving Interventions in Trauma Patients by Using New Vital Signs and Artificial Neural Networks

Andriy I. Batchinsky, Jose Salinas, John A. Jones, Corina Necsoiu, and Leopoldo C. Cancio

U.S. Army Institute of Surgical Research, 3400 Rawley E. Chambers Avenue, Building 3611, Fort Sam Houston, Texas, USA
andriy.batchinsky@amedd.army.mil

The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Department of the Army or the U.S. Department of Defense.

Abstract. Previous work in risk stratification of critically injured patients involved artificial neural networks (ANNs) of various configurations tuned to process traditional vital signs and demographical, clinical, and laboratory data obtained via direct contact with the patient. We now report "new vital signs" (NVSs) that are superior in distinguishing the injured and can be derived without hands-on patient contact. Data from 262 trauma patients are presented, in whom NVSs derived from electrocardiogram (EKG) analysis (heart-rate complexity and variability) were input into a commercially available ANN. The endpoint was performance of life-saving interventions (LSIs) such as intubation, cardiopulmonary resuscitation, chest-tube placement, needle chest decompression, and blood transfusion. We conclude that based on EKG-derived NVS alone, it is possible to accurately identify trauma patients who undergo LSIs. Our approach may permit development of a next-generation decision support system.

Keywords: electrocardiography, heart rate complexity, heart rate variability, ANN, prediction of life-saving interventions.

1 Introduction

Clinical decision making by physicians is a complicated, subjective, and nonlinear process. More objective tools stemming from advances in artificial intelligence science are available for medical prognosis [1, 2] and have been used with encouraging results [3, 4]. Most studies to date have used traditional vital signs (such as heart rate and blood pressure), demographical data, disease symptoms, laboratory findings, and injury severity scores as descriptors of patient state. Generation of this information is laborious and time-consuming, and requires direct patient contact--which may not be possible during rural rescue operations, mass-casualty events, or combat. Even when

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 390–394, 2009.

available, traditional vital signs can be misleading as they lack diagnostic accuracy, especially in early, compensated shock.

We demonstrated that semiautomatic analysis of the electrocardiogram (EKG) using linear and nonlinear statistical approaches produces a panel of descriptive variables that are sensitive markers of physiologic state during blood loss [5] and trauma [6]; and are associated with mortality [6] and the need to perform life-saving interventions (LSIs) [7] in trauma patients. Most importantly, generation of these descriptive variables does not require contact with the patient, can be performed remotely via telemetry, and requires small sections of EKG acquired over minutes. The objective of this study was to evaluate whether EKG-derived data submitted to off-the-shelf ANN software could be used for assessing the need to perform life-saving interventions in a mixed cohort of prehospital and emergency department trauma patients.

2 Methods

To select the cohort, 464 patients were screened from the Trauma Vitals database developed at our Institute. The cohort consisted of patients with both blunt and penetrating injuries admitted to emergency departments in Houston and San Antonio, Texas, as well as in Baghdad, Iraq. Patients were excluded from the study if 1) EKG sections did not contain at least 800 R-to-R intervals (RRIs); 2) ectopic beats were present within the analyzed data segments; or 3) the EKG contained electromechanical noise. Demographical and vital sign data was analyzed from charts.

Continuous 20- to 30-minute sections of EKG waveforms were recorded from each patient only once and stored on a computer. The earliest available 800-beat data sets were imported into WinCPRS software (Absolute Aliens Oy, Turku, Finland) and analyzed as previously described [5]. Heart rate complexity as well as time- and frequency-domain analyses were used to generate the list of new vital signs (NVS) as previously described [5] (see Tables 1 and 2). A commercially available feed-forward back-propagation ANN (NeuralWare, Carnegie, PA) was used with training on 70% and validation on 30% of the data with tenfold cross validation. SAS version 9.1 (SAS Institute, Cary, NC) was used for statistical analysis. A receiver-operating-characteristic (ROC) curve was constructed to assess the diagnostic performance of the ANN. Estimated odds ratio and 95% confidence intervals (CI) were determined by the maximum-likelihood method.

3 Results

Data from 192 prehospital trauma patients and 70 emergency room patients (n=262) patients were included. Of these, 65 patients received a total of 88 life-saving interventions (LSIs). LSIs in the 65 patients were intubation (n=61), cardiopulmonary resuscitation (n=5), cricothyroidotomy (n=2), emergency blood transfusions (n=4), and decompression of pneumothorax (n=16). LSI and non-LSI patients were clinically indistinguishable with respect to age, sex, mechanism of injury, heart rate, or blood pressure. LSI patients were, however, more severely injured based on injury severity scores, had higher heart rates (lower RRI, Table 1), and had higher mortality (data not shown).

Results are provided in Table 1 for linear (time- and frequency-domain) variables and show a clear separation between the two groups.

Table 1. Linear time- and frequency-domain analysis variables associated with the need to perform life-saving interventions by ANN

[RRI, mean R-to-R interval of the EKG, ms; RMSSD, the square root of the mean squared differences of successive normal-to-normal (NN) RRIs; TP, total R-to-R interval spectral power (0.003-0.4 Hz, ms^2); HF, RRI spectral power at the high frequency (0.15-0.4, ms^2); LF/HF, the ratio of LF (RRI spectral power at the low frequency (0.04-0.15 Hz, ms^2) to HF; HFnu, spectral power at the high frequency normalized to TP; CDM LF, amplitude of the LF oscillations by complex demodulation; CDM HF, amplitude of the HF oscillations. CDM LF/HF, ratio of the CDM LF and CDM HF. Data are means ± SEM.]

Variable	Non LSI (n=197)	LSI (n=65)	p value	Reflects parasympathetic nervous system	Reflects sympathetic nervous system
RRI	650.50 ± 9.70	565.63 ± 16.19	<.0001	yes	yes
RMSSD	13.89 ± 0.88	6.17 ± 0.77	<.0001	yes	
TP	1107.75 ± 131.81	305.98 ± 58.73	<.0001	yes	yes
HF	95.57 ± 13.73	21.49 ± 7.03	<.0001	yes	
LF/HF	150.04 ± 104.68	104.68 ± 46.39	<.0001	yes	yes
HFnu	0.20 ± 0.01	0.25 ± 0.02	0.013	yes	
CDM LF	16.22 ± 0.78	5.75 ± 0.86	<.0001	yes	yes
CDM HF	8.28 ± 0.57	3.35 ± 0.53	<.0001	yes	
CDM LF/HF	2.40 ± 0.09	1.79 ± 0.13	<.0001	yes	yes

Specifically, the LSI group had lower values for all time-domain and frequency-domain descriptive metrics used, but the HFnu which was higher. The nonlinear analysis methods also showed decreased values in the majority of metrics with the exception of SOD and FW, which were higher (Table 2).

Table 2. Heart-rate complexity variables associated with the need to perform life-saving interventions by ANN

[ApEn, approximate entropy; FDDA, fractal dimension by dispersion analysis; DFA, short-term correlations within the RRI by Detrended Fluctuations Analysis; SOD, similarity of distributions; StatAV, signal Stationarity; FW (%), forbidden words; DisnEn, normalized signal distribution entropy. All variables are unitless. Data are means ± SEM.]

Variable	Non LSI (n=197)	LSI (n=65)	p value
ApEn	1.10 ± 0.02	0.93 ± 0.04	<.0001
FDDA	1.13 ± 0.01	1.07 ± 0.01	<.0001
DFA	1.35 ± 0.03	1.07 ± 0.05	<.0001
SOD	0.15 ± 0.00	0.20 ± 0.01	<.0001
StatAV	0.82 ± 0.01	0.95 ± 0.01	<.0001
FW	52.59 ± 0.93	60.84 ± 1.17	<.0001
DisnEn	0.64 ± 0.01	0.55 ± 0.01	<.0001

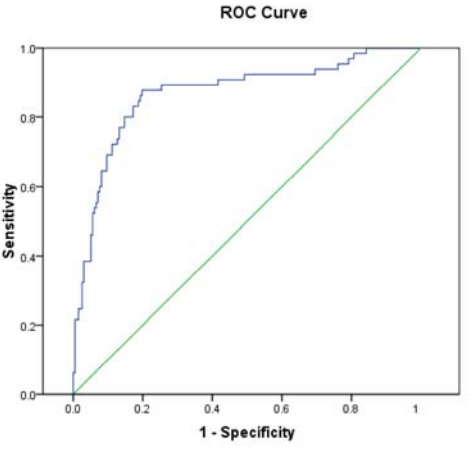


Fig. 1. ROC curve for model derived by ANN for prediction of LSIs using EKG variables alone. Area under the curve (AUC) = 0.868; 10-fold cross validation; standard error 0.028. Asymptotic significance was 0.001; and the lower and higher asymptotic 95% confidence intervals were 0.812 and 0.924, respectively.

Sixteen of the calculated variables were associated with the need to perform LSIs (Tables 1 and 2). Areas under the ROC curves prepared after threefold, fivefold, and tenfold cross-validation did not differ significantly and were 0.864, 0.861, and 0.868, respectively (Fig. 1). In each respective case, the list of descriptives associated with the LSIs differed by a few variables.

4 Discussion

A cornerstone of medical decision making at present is direct contact with the patient, during which visual assessment, physical examination, traditional vital signs (heart rate, blood pressure), and determination of injury severity are obtained. Traditional vital signs, however, are relatively inaccurate descriptors of trauma patient status and frequently do not distinguish patients in need of immediate LSIs until their condition suddenly deteriorates. Particularly during combat, direct contact with the injured may expose the medic to unacceptable risk until the scene is safe, which has engendered the concept of remote monitoring via telemetry.

We introduce a new way of assessing critically injured patients. The main finding of this study is that the ANN identified patients who received an LSI based on EKG-derived data alone with a significant and clinically relevant degree of accuracy. The implications of this retrospective work are threefold. First, these results support previous reports in the literature that injury leads to decrease in heart-rate complexity and heart-rate variability as assessed by multiple linear and nonlinear statistical tools [5, 6, 8].

Second, the need to perform LSIs could be predicted in entirely automatic fashion, pending further improvements in computerized waveform analysis, signal processing, and transmission.

Third, the clinical relevance of our findings as acquired on a mixed cohort of critically injured patients will add additional credibility to the ability of artificial intelligence systems to interpret medical descriptive data that is otherwise too complicated for fast processing by humans. Potential applications of this approach may include development of personal diagnostic and monitoring systems to be used in automobile accident alert systems; for remote assessment and triage of patients in austere environments and mass-casualty settings; and during combat. This approach could also serve as an evidence-based decision assistance tool helping medical providers distinguish patients in imminent danger of dying during times when changes in their traditional vital signs are noninformative.

References

1. Abu-Hanna, A., Lucas, P.J.: Prognostic Models in Medicine. AI and statistical approaches. Methods of Information in Medicine 40, 1–5 (2001)
2. Kononenko, I.: Machine Learning for Medical Diagnosis: History, State of the Art and Perspective. Artificial Intelligence in Medicine 23, 89–109 (2001)
3. Baxt, W.G., Shofer, F.S., Sites, F.D., Hollander, J.E.: A Neural Network Aid for the Early Diagnosis of Cardiac Ischemia in Patients Presenting to the Emergency Department with Chest Pain. Ann. Emerg. Med. 40, 575–583 (2002)
4. DiRusso, S.M., Sullivan, T., Holly, C., Cuff, S.N., Savino, J.: An Artificial Neural Network as a Model for Prediction of Survival in Trauma Patients: Validation for a Regional trauma area. J. Trauma 49, 212–220; discussion 220–213 (2000)
5. Batchinsky, A.I., Cooke, W.H., Kuusela, T., Cancio, L.C.: Loss of Complexity Characterizes the Heart-Rate Response to Experimental Hemorrhagic Shock in Swine. Crit. Care Med. 35, 519–525 (2007)
6. Batchinsky, A.I., Cancio, L.C., Salinas, J., Kuusela, T., Cooke, W.H., Wang, J.J., Boehme, M., Convertino, V.A., Holcomb, J.B.: Prehospital Loss of R-to-R Interval Complexity Is Associated with Mortality in Trauma Patients. J. Trauma 63, 512–518 (2007)
7. Cancio, L.C., Batchinsky, A.I., Salinas, J., Kuusela, T., Convertino, V.A., Wade, C.E., Holcomb, J.B.: Heart-rate complexity for prediction of prehospital lifesaving interventions in trauma patients. J. Trauma 65, 813–819 (2008)
8. Winchell, R.J., Hoyt, D.B.: Spectral Analysis of Heart Rate Variability in the ICU: A Measure of Autonomic Function. J. Surg. Res. 63, 11–16 (1996)

Causal Probabilistic Modelling for Two-View Mammographic Analysis

Marina Velikova[1], Maurice Samulski[1], Peter J.F. Lucas[2], and Nico Karssemeijer[1]

[1] Department of Radiology, Radboud University Nijmegen Medical Centre
6525 GA, Nijmegen, The Netherlands
{m.velikova,m.samulski,n.karssemeijer}@rad.umcn.nl
[2] Institute for Computing and Information Sciences, Radboud University Nijmegen
6525 ED Nijmegen, The Netherlands
peterl@cs.ru.nl

Abstract. Mammographic analysis is a difficult task due to the complexity of image interpretation. This results in diagnostic uncertainty, thus provoking the need for assistance by computer decision-making tools. Probabilistic modelling based on Bayesian networks is among the suitable tools, as it allows for the formalization of the uncertainty about parameters, models, and predictions in a statistical manner, yet such that available background knowledge about characteristics of the domain can be taken into account. In this paper, we investigate a specific class of Bayesian networks—causal independence models—for exploring the dependencies between two breast image views. The proposed method is based on a multi-stage scheme incorporating domain knowledge and information obtained from two computer-aided detection systems. The experiments with actual mammographic data demonstrate the potential of the proposed two-view probabilistic system for supporting radiologists in detecting breast cancer, both at a location and a patient level.

1 Introduction

The interpretation of screening mammograms is routine work for radiologists involved in national breast-cancer screening programs. Although routine, it is still a task fraught with difficulty with much space for improvement. The difficulty of reading mammograms is due to a number of factors such as variations in female breast tissue, cancer appearance on a mammogram, and image quality. To facilitate their screening work, radiologists are typically provided with two projections, or *views*, of each breast: mediolateral oblique (MLO), taken under 45° angle and showing part of the pectoral muscles, and craniocaudal (CC), taken head to tail. If cancer is present, then it is expected to be observed in both views. Observing that radiologists still miss too many cancer cases, offering them some form of assistance, for example, through computer-aided detection (CAD) systems is important. However, without exploiting principles that radiologists have

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 395–404, 2009.

used successfully for decades, it is unlikely that CAD systems will ever outperform highly trained human interpreters. It is only recently that researchers have started to study ways to incorporate such principle into CAD systems.

Bayesian networks are especially promising in bridging this gap between the capabilities of humans and computer-aided interpretation, as they have the virtue of supporting the explicit representation of expert knowledge, handle uncertainty and missing information, and allow combining multiple types of knowledge. One of the principles used by radiologists in analysing mammograms is combining information obtained from different views of the same breast, which provides the basis for the development of a model for two-view mammographic analysis presented in this paper. The aim of the current study was two fold:

- to improve the breast cancer detection rate at a location and a patient level in comparison to a single-view CAD system;
- to get more insight into the mechanisms underlying mammographic analysis, which then can act as a basis for improvement of current CAD systems.

To achieve these goals, we developed a multi-stage system, using (i) a specific class of Bayesian networks, called causal independence models, and (ii) knowledge derived from an analysis of the way radiologists interpret mammograms. The method here builds on our previous research presented in [1,2].

We adopt the following terminology from the breast cancer domain throughout this paper. By *lesion* we refer to a physical cancerous object detected in a patient; see Fig. 2. We call a contoured area on a mammogram a *region*, marked, for example, manually by a human or detected automatically by a CAD system. A region detected by a CAD system is described by a set of continuous (real-valued) *single-view features*, e.g., size, location, contrast. By *link* we denote established correspondence, between two regions in MLO and CC views, respectively. Every link is described by a set of *multi-view features*, such as contrast difference and location difference. The most recent mammographic *exam* for a woman is called *current*, whereas the previous exam(s) are *prior(s)*.

Previous research has already demonstrated the potential of exploring multi-view dependencies to improve the automatic detection of breast cancer on mammograms. The approaches in [3,4] focused on improving the lesion-based results, mostly for prompting purposes, based on the distinction between true and false positive links of regions in MLO and CC views. In other studies multi-view information was used to increase both lesion-based and exam-based performance, i.e., fraction of true positive exams where an exam is true positive if cancer is found in the MLO *or* CC views, in comparison to a single-view CAD system ([5,6]). In contrast to the other research, which mostly explores neural networks or linear discriminant analysis, the probabilistic methodology proposed in the current study has the advantage of providing not only strong predictive power but also explicit modelling of expert knowledge and insight in the results obtained–properties desired especially by medical domain experts.

2 Causal Probabilistic Modelling

2.1 Bayesian Networks

A *Bayesian network* is defined as a pair $\mathrm{BN} = (G, P)$, where G is an acyclic directed graph (ADG) $G = (V, E)$ and P is a joint probability distribution of the random variables X. There exists a 1–1 correspondence between the nodes in V and the random variables in X; the (directed) edges, or arcs, $E \subseteq (V \times V)$ correspond to direct causal relationships between the variables. We say that G is an *I–map* of P if any independence represented in G, denoted by $A \perp\!\!\!\perp_G B \mid C$, with $A, B, C \subseteq V$ mutually disjoint sets of nodes, is satisfied by P, i.e.,

$$A \perp\!\!\!\perp_G B \mid C \quad \Longrightarrow \quad X_A \perp\!\!\!\perp_P X_B \mid X_C \ ,$$

where A, B and C are sets of nodes of the ADG G and X_A, X_B and X_C are the corresponding sets of random variables, indexed by A, B, and C. The acyclic directed graphical part of a Bayesian network G is by definition an I–map of the associated joint probability distribution P. A Bayesian network BN offers a compact representation of the joint probability distribution P in terms of local *conditional probability distributions* (*CPDs*), by taking into account the conditional independence information represented by the ADG.

2.2 Causal Independence Models

Causal independence arises when multiple causes (parent nodes) lead to a common effect (child node) through interaction of independent uncertain processes. Causal independence models provide a way to specify interactions among random variables in a compact fashion [7]. A definition of the notion of causal independence is given following the one from [8].

The general structure of a causal-independence model is shown in Fig. 1; it expresses the idea that causes $C_1, \ldots, C_n$ influence a given common effect E through intermediate variables $I_1, \ldots, I_n$; the intermediate variable I_j is considered to be a contribution of the cause variable C_j to the common effect E. The *interaction function* f represents in which way the intermediate effects I_j, and indirectly also the causes C_j, interact. This function f is defined in such way that when a relationship between the I_j's and $E = \mathit{true}$ is satisfied, then it holds that $f(I_1, \ldots, I_n) = \mathit{true}$; otherwise, it holds that $f(I_1, \ldots, I_n) = \mathit{false}$.

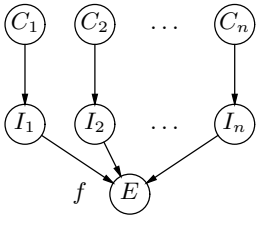


Fig. 1. Causal-independence model

Note that each variable I_j is only dependent on its associated cause C_j and the effect variable E. Furthermore, the graph structure expresses that the effect variable E is conditionally independent of each cause C_j given the associated intermediate variable I_j. It is assumed that absent causes do not contribute to the effect, i.e., $P(I_j = \mathit{true} \mid C_j = \mathit{false}) = 0$. As an example from the breast cancer domain, we can consider the regions detected by a single-view CAD system as cause variables and presence of breast cancer as the effect variable.

2.3 Exact and Threshold Functions

A natural class of interaction functions f are the *symmetric* Boolean functions where the order of the arguments does not matter. There are 2^{n+1} of such functions, with n the number of arguments; typical examples are the logical OR, AND and XOR. In the breast cancer domain, for example, the order of regions does not play a role in determining whether the breast is or is not cancerous.

A useful feature of symmetric Boolean functions is their decomposability in terms of *exact* Boolean functions. The exact function e_k checks whether there are *exactly* k trues among its arguments, i.e., $e_k(I_1, \ldots, I_n) = \mathit{true}$, if $\sum_{j=1}^{n} I_j = k$. In decision making under uncertainty there is a natural tendency to aggregate available uncertain information until a threshold is passed. The *threshold function* τ_k is a symmetric Boolean function that allows us to model this principle; it checks whether there are *at least* k trues among its arguments, i.e., $\tau_k(I_1, \ldots, I_n) = \mathit{true}$, if $\sum_{j=1}^{n} I_j \geq k$. Note that the logical OR function is a threshold function τ_k with $k = 1$ and the AND function is a threshold function τ_k with $k = n$. The conditional probability of the effect variable E given the causes $C_1, \ldots, C_n$ in a noisy threshold model with interaction function τ_k is given by:

$$P_{\tau_k}(e \mid C_1, \ldots, C_n) = \sum_{k \leq l \leq n} \sum_{e_k(I_1, \ldots, I_n)} \prod_{j=1}^{n} P(I_j \mid C_j) \ . \tag{1}$$

From Equation (1), it follows that $P_{\tau_k}(e \mid C_1, \ldots, C_n) \geq P_{\tau_{k+1}}(e \mid C_1, \ldots, C_n)$, for each $k \geq 0$, i.e., with a lower value of threshold k the probability of the effect e is non-decreasing. A more detailed description of exact and threshold functions can be found in [8]. Causal independence models with threshold functions are the basic elements for the model presented in the next section.

3 Causal Modelling for Mammographic Analysis

3.1 Two-View Analysis

The objective of a two-view mammographic analysis is to determine whether or not a breast exhibits cancerous characteristics by establishing correspondences between regions in the two breast views. Fig. 2 depicts the general multi-view detection scheme used in this study.

A lesion (cancerous object) is represented by a circle in a view. Clearly, if a lesion is detected in both views, the breast is cancerous and the patient has breast

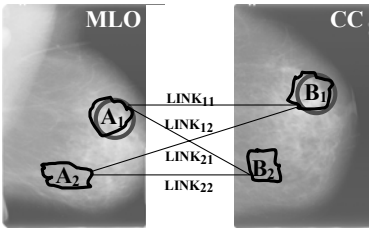


Fig. 2. Schematic representation of mammographic two-view analysis with automatically detected regions. The circle represents a lesion (cancerous object).

cancer. An automatic single-view CAD system attempts to establish whether there are regions that are suspicious for cancer in each view separately. In the figure regions A_1 and B_1 have correctly been detected as lesions, i.e., these are true positive (TP) regions, whereas regions A_2 and B_2 are false positive (FP) regions. Since we deal with projections of the same physical object—the lesion—correspondence between lesions is represented by a *link* (LINK_{ij}) between regions in each view, A_i and B_j. To every link, a value $\text{LINK}_{ij} = \ell_{ij}$ is assigned, where $\ell_{ij} \in \{\text{TPTP}, \text{TPFP}, \text{FPTP}, \text{FPFP}\}$. For every region, breast, exam and patient, a class with values of *true* (presence of cancer) or *false* is assumed to be provided by pathology or a human expert.

3.2 Probabilistic Model

The architecture of our model for two-view mammographic analysis is inspired by the way radiologists analyse images. They do this by distinguishing several levels in the interpretation process. At the lowest (image) level, radiologists look for suspicious regions with cancer characteristics. If suspicious regions are observed on both views of the same breast, then the (individual) suspiciousness of these regions increases implying that a lesion is likely to be present. As a result, the whole breast as well as the exam (patient) is considered suspicious for cancer.

The first steps—identifying suspicious regions and establishing links between them on both views of the same breast—have already been tackled in previous research conducted by our group. Here, we build upon the resulting systems to model the following stages in the mammographic analysis as described above. Fig. 3 presents an overview of the probabilistic model.

We start by modelling the two-view dependencies between the regions in MLO and CC. For each of the four link values ℓ_{ij} we consider the links LINK_{ij} with their respective set of multi-view features **MVFeat** and the correspondence scores $\text{CorrSc}(\text{LINK}_{ij})$, $\text{CorrSc}(\text{LINK}_{ji})$ obtained from the system described in [2]. We have used logistic regression to reliably compute the conditional probability distribution $P(\text{LINK}_{ij} = \ell_{ij} \mid \text{CorrSc}(\text{LINK}_{ij}), \text{CorrSc}(\text{LINK}_{ji}), \mathbf{MVFeat})$. Thus, for every link LINK_{ij} we obtain four probabilities corresponding to each link value and every link probability is symmetric.

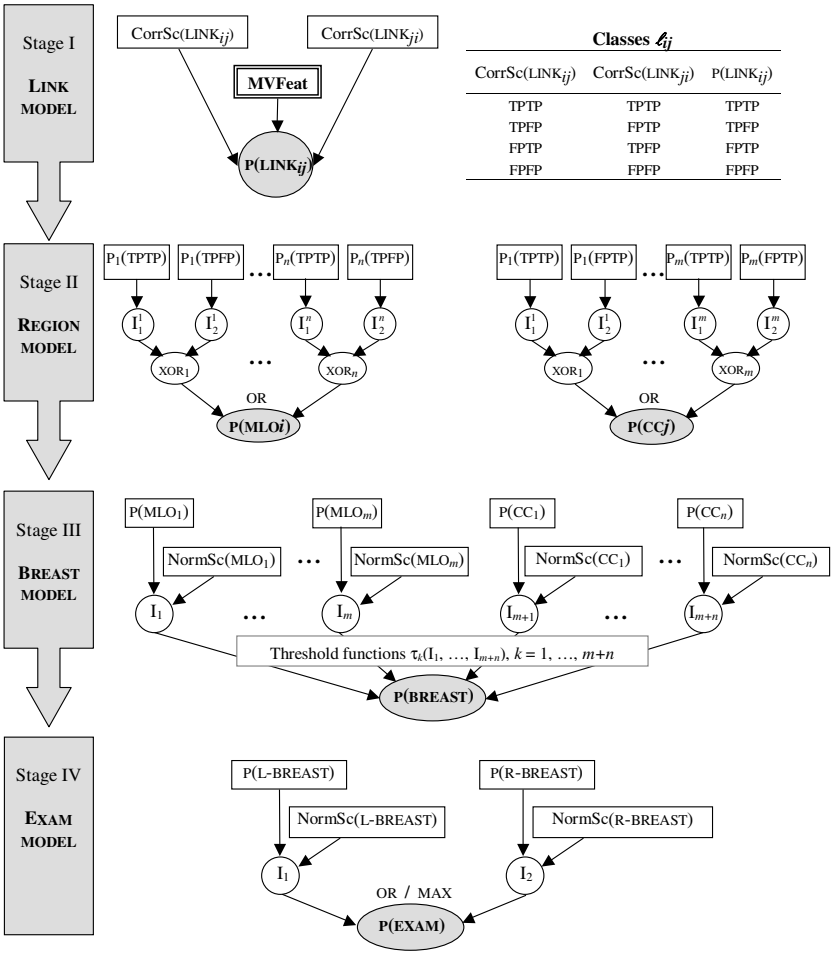


Fig. 3. Multi-stage causal probabilistic model for two-view mammographic analysis

At the second stage, we compute the probability of a region being cancerous given the link information about the regions in the complementary view. This is done by combining link probabilities obtained from the first stage using a causal independence model, where the link probabilities are the cause variables and the region probability is the effect variable. In computing the probability of a region of MLO being cancerous, we combine only the link probabilities for the classes TPTP and TPFP as they correspond to a TP region of MLO. With respect to a region of CC, the link classes considered are TPTP and FPTP. These link probabilities interact through the XOR function, as only one of them can be true. Next the logical OR is used to represent the knowledge that the probability of a region being cancerous is true if at least one of the link probabilities is true.

At the third stage, we focus at the breast level where the region probabilities from the respective MLO and CC views are combined using a causal independence model with a threshold function. In the combining scheme, we also use the suspiciousness measure for the region (NormSc(MLO_i), NormSc(CC_j)) computed by the single-view CAD system ([2]), which is already a good indicator for discriminating between normal and cancerous regions. By varying the threshold k from 1 to $m + n$ (the maximum number of regions in the breast), we try to get insight into the causal interactions between the regions and the breast. One can expect that models with small threshold values would be able to distinguish well between cancerous and normal breasts whereas models with larger values of k might not be able make the distinction. This expectation follows from the fact that breast cancer in its early stages is mostly unifocal, i.e., located in a single region, and not observed on multiple locations in the breast (view).

At the last stage, we combine the probabilities for the left and right breast and their respective single-view measures for suspiciousness (NormSc(BREAST)) to compute the probability for an exam being cancerous. Two combination functions are used and compared: the logical OR and the MAX function.

Finally, having patients with more than one exam (current and prior(s) available), we compute the probability for the patient having cancer by taking the probability of her current (most recent) exam, which is presumably most informative, i.e. $P(Patient = cancerous) = P(CurrentExam = cancerous)$.

3.3 Data Description

The data set contains 392 (332 current + 60 prior) exams from which 218 (185 current + 33 prior) were cancerous. The exams of one patient were considered as independent. All exams contained both MLO and CC views. All cancerous breasts had one visible lesion in at least one view, which was verified by pathology reports to be cancerous. Lesion contours were marked by a mammogram reader.

For each image (mammogram) we have a number of regions detected as suspicious by the single-view CAD system presented in [2]. This number varies between 1 and 5 per image (2 and 10 per breast). For each region, based on the ground-truth data, we have a class value of $true$ (TP) if the detected region hits a cancerous finding and $false$ (FP) otherwise. Every region from MLO was linked with every region in CC. Every link was described by a set of multi-view features. For every link we assigned one of the four link class values depending on the region class values. We assign binary classes of $true$ (cancerous) and $false$ (normal) for a breast, exam and patient based on the ground-truth information.

3.4 Training, Evaluation and Results

The proposed model has been built, trained and tested using the Matlab-based Bayesian Network Toolbox ([9]). The evaluation of the model is done using ten-fold cross validation with the same data split as the one used in [2]. For every split of the data, the test set is used only for testing and never for training at different stages of the model. Thus, we used an unbiased evaluation procedure.

Table 1. AUCs obtained from `MV-CAD-Causal` and `MV-CAD-kNN` at a *link* level

Link class	Current exams		All exams	
	MV-CAD-Causal	MV-CAD-kNN	MV-CAD-Causal	MV-CAD-kNN
TPTP	0.935	0.918	0.936	0.914
TPFP	0.838	0.660	0.844	0.650
FPTP	0.888	0.809	0.881	0.785
FPFP	0.887	0.829	0.874	0.824

Table 2. AUCs obtained from `MV-CAD-Causal` and `SV-CAD` at a *breast* level

k	Current exams		All exams	
	MV-CAD-Causal	SV-CAD	MV-CAD-Causal	SV-CAD
1	0.919		0.902	
2	0.917	0.904	0.899	0.895
3	0.872		0.859	

Table 3. AUCs obtained from `MV-CAD-Causal` and `SV-CAD` at an *exam* level

k	Current exams			All exams		
	MV-CAD-Causal		SV-CAD	MV-CAD-Causal		SV-CAD
	MAX	OR		MAX	OR	
1	0.903	0.899		0.889	0.879	
2	0.897	0.892	0.877	0.884	0.882	0.865
3	0.897	0.891		0.883	0.877	

At a link level the performance of our multi-view model is compared with the multi-view model based on k-Nearest Neighbour (`MV-CAD-kNN`) presented in [2]. At a region, breast and exam level, the benchmark for comparison is the single-view CAD system (`SV-CAD`). For the latter, the likelihood for a view, breast and exam being cancerous is computed by taking the likelihood of the most suspicious region. The comparison analysis is done using the (Free-response) Receiver Operating Characteristic ((F)ROC) curve and the Area Under the Curve (AUC). We next report the test results for all and current (patient) exams at all levels.

Link level. Table 1 presents the AUCs at a link level obtained from `LinkModel` and `MV-CAD-kNN`. The results show that our link model outperforms `MV-CAD-kNN` for all link values, with a considerable difference for the link types: TPFP, FPTP and FPFP. Except the classification improvement, our method has the advantage of making the links symmetric, i.e., for every two regions there is only one probability per link class.

Region level. To evaluate the performance of both multi- and single-view CAD systems at a region level we use the FROC curve which plots the lesion-detected and localized fraction (y-coordinate) vs. the average number of false positives

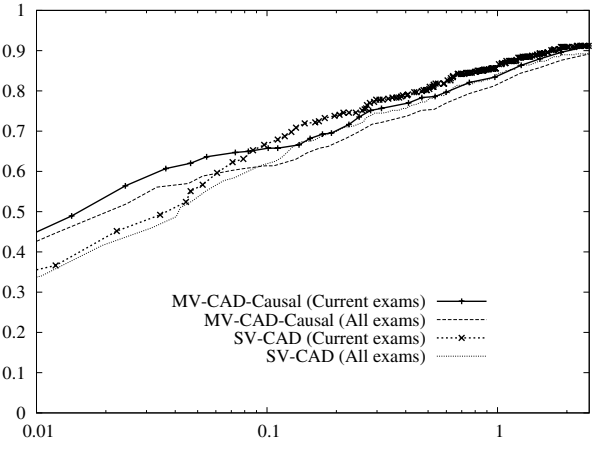


Fig. 4. FROC curves for the performance of `MV-CAD-Causal` and `SV-CAD` based on all and current exams at a *region* level

per image (x-coordinate). Fig. 4 depicts the results for the multi- and single-view CAD system for all and current exams. It is clear that taking two-view information into account helps increasing the cancer detection rate while keeping the number of false positives per image very low. This trend is even clearer for the regions in the current exams. This is a desirable outcome for screening programs where radiologists operate at high specificity (low false positive) rates because of the very low cancer incidence rates.

Breast level. At this level, we compare the performance of `BreastModel` using different threshold functions τ_k, where $k = 1, \ldots, 10$ with respect to `SV-CAD`. Table 2 presents the results for the three best multi-view models for all and current exams. The results confirm our expectation–the correct detection of at least one ($k = 1$) or two cancerous regions ($k = 2$) is sufficient to classify the breast as cancerous. For these thresholds the multi-view CAD system outperforms the single-view CAD system. For threshold functions with $k \geq 4$ the performance of `BreastModel` drops significantly reaching AUCs of 0.506 when $k = 10$ for all and current exams. This result is in line with the domain knowledge and the local nature of the early developing breast cancer.

Exam level. At the screening practice the most important question eventually is whether or not a patient is suspicious for cancer and needs to be referred for further examination. To answer this question, for the current comparison study, we focus on the results obtained from the last stage of our model (`ExamModel`) based on the best three `BreastModel` with $k = 1, 2, 3$; see Table 3. For all and current exams, we observe overall improvement in the breast cancer detection rate achieved by our two-view system. It is interesting to note that the causal independence modelling in `ExamModel` helps improve the performance of the exam model even if the performance of the breast model is less satisfactory as for `BreastModel` with a threshold of $k = 3$. Furthermore, we notice that

using MAX as a combination function for the breast probabilities leads to a better distinction between cancerous and normal exams than using the logical OR. A possible explanation might be that the latter tends to overestimate the probability of normal exams by considering both breasts, whereas the MAX function seems to be more appropriate given that in screening mammography mostly one of the breasts is cancerous.

4 Conclusion

We presented a unified Bayesian network framework for two-view mammographic analysis motivated by the radiologists' practice and the organization of the domain. The foundation of the framework is based on the notion of causal independence where the interaction between the variables is modelled by threshold functions. The definition of link used in this paper better captures the correspondences between the regions in both views, in comparison to our previous approach ([1]), where we used binary links. Through experimental results we showed that for lower thresholds the proposed two-view probabilistic model not only is in line with the domain knowledge but also outperforms a single-view CAD system by increasing the breast cancer detection rate for low false positive rates, both at a location and a patient level.

Acknowledgements. This work has been funded by the Netherlands Organization for Scientific Research under BRICKS/FOCUS grant number 642.066.605.

References

1. Velikova, M., Samulski, M., Lucas, P.J.F., Karssemeijer, N.: Improved mammographic CAD performance using multi-view information: A Bayesian network framework. Physics in Medicine and Biology 54, 1131–1147 (2009)
2. Samulski, M., Karssemeijer, N.: Matching mammographic regions in mediolateral oblique and cranio caudal views: A probabilistic approach. In: Proceedings of SPIE, Medical Imaging, vol. 6915 (2008)
3. Good, W., Zheng, B., Chang, Y., Wang, X., Maitz, G., Gur, D.: Multi-image cad employing features derived from ipsilateral mammographic views. In: Proceedings of SPIE, Medical Imaging, vol. 3661 (1999)
4. van Engeland, S., Karssemeijer, N.: Combining two mammographic projections in a computer aided mass detection method. Medical Physics 34, 898–905 (2007)
5. Paquerault, S., Petrick, N., Chan, H., Sahiner, B., Helvie, M.A.: Improvement of computerized mass detection on mammograms: Fusion of two-view information. Medical Physics 29, 238–247 (2002)
6. Qian, W., Song, D., Lei, M., Sankar, R., Eikman, E.: Computer-aided mass detection based on ipsilateral multiview mammograms. Acad. Rad. 14, 530–538 (2007)
7. Heckerman, D., Breese, J.S.: Causal independence for probability assessment and inference using Bayesian networks. IEEE Trans. on SMC–A 26, 826–831 (1996)
8. Visscher, S., Lucas, P.J.F., Schurink, C.A.M., Bonten, M.J.M.: Modelling treatment effects in a clinical Bayesian network using Boolean threshold functions. Artificial Intelligence in Medicine (2008)
9. Murphy, K.: Bayesian Network Toolbox (BNT) (2007), http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html

Modelling Screening Mammography Images: A Probabilistic Relational Approach

Nivea Ferreira and Peter J.F. Lucas

Institute for Computing and Information Sciences,
Radboud University Nijmegen
{nivea,peterl}@cs.ru.nl

Abstract. Computer-aided detection systems have as aim the increase of detection rates when analysing mammograms, by identifying features that are characteristic for breast cancer. In this research we aimed at using the features extracted from mammographic images in order to analyse the development of suspicious lesions. Different from other approaches, we based our data modelling on object orientation. This allowed not only for a description of domain entities and their intrinsic relationships, but also for the application of relational probabilistic techniques, which can handle heterogeneous data instances both in terms of learning and inference.

1 Introduction

There is considerable empirical evidence that the early detection of breast cancer has a positive effect on the prognosis of the disease, and, thus, breast cancer screening is widely seen as one of the cornerstones of breast cancer management. The screening process involves reading off the resulting mammograms, i.e., breast X-ray images, by trained radiologists. Some years ago, researchers have proposed the use of *computer-aided detection* (CAD) systems as a contribution to the reduction in the number of missed cases. A proper interpretation of mammograms requires that all the images are interpreted in relationship to each other. For example, if the mammogram of the left breast indicates low density of the glandular tissue, it is very likely that the image density of the right breast will be low as well. Models used in CAD systems have so far failed to take such interpretations into account, and this may explain lack of progress in the area.

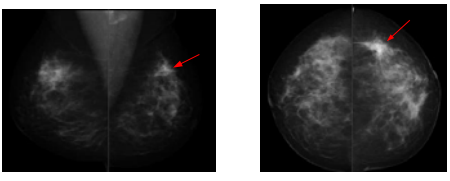
Recent research on relational probabilistic models offers new methods for expressing relationship, possibly uncertain, between objects in a domain [1]. Part of the ideas underlying this field come from object-oriented database theory. As all mammograms have identical features, yet with different values due to the fact that mammograms may be taken under different projections or from opposite breasts, it is quite natural to exploit object orientation in mammogram interpretation.

The aim of the presented research was, firstly, to show that object orientation offers a natural start for the design of pattern recognition techniques for breast cancer detection and, secondly, to study the use of relational probabilistic methods for the diagnosis of breast cancer.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 405–409, 2009.

2 Screening Mammography

Mammographic images are obtained from different projections, usually mediolateral oblique (MLO) and craniocaudal (CC). The CC and MLO images yield complementary information to the interpreting radiologist in the sense that an abnormality in one of the views of the breast is likely, although not always, present in the other view. An abnormality is usually called a *finding* or *lesion*. An example of a lesion on MLO and CC views is shown in Fig. 1.



(a) (b)

Fig. 1. Highlighted left-breast lesion in (a) MLO view and (b) CC view

The mammographic image is further distinguished into so called *regions of interest*, or region for short, characterised by means of *features*, e.g., density and location. These features may suggest a certain level of suspiciousness for cancer. Regions can be naturally seen as objects and their associated features as object attributes. We used output produced by the CAD system described in [2,3]; this system performs segmentation of a mammogram, and extracts regions of interest based on pixel information. The CAD system then determines the features of each identified region.

3 Relational Probabilistic Modelling

In relational learning, the data instances are neither recorded in homogeneous structures as commonly used in machine learning, nor is automatically assumed that instances are independent and identically distributed, the almost standard assumption in machine learning. In the context of relational models, random variables are the objects' attributes, and thus a relational probabilistic model always assumes the presence of an object-oriented data model.

Probabilistic relational models define a generic dependency structure at the level of *types*, which enables generalisation from a single instance [1]. Learning relational probability trees (RPTs) takes a set of subgraphs as input, where each subgraph contains a target object to be classified and a set of other objects with forms the target object's relational neighbourhood. It then constructs a probability estimation tree to predict the target label. The algorithm searches over the space of binary relational features in order to obtain a split of the data, taking into account feature scores and correlation among features. Within the features highly correlated with the target object, the feature showing the maximum correlation is selected and included in the tree.

4 Relational Mammography Modelling

4.1 The Mammogram Data Model and Database

A patient is associated to an exam. The exam is a collection of image projections of the patient's breast: MLO-right, MLO-left, CC-right and CC-left. Each of those images consists of regions. In a particular exam, from an object representing a region in CC and another object representing a region in MLO we obtain a link. All regions in CC are linked to all regions in MLO in the same breast projection. Fig. 2(a) depicts this database design.

Our data set contained 1,063 screening exams of which 383 were confirmed as being cancerous by pathology. All exams contained both MLO and CC views. The total number of breasts was 2,126. Lesion contours were marked by, or under supervision of, an experienced screening radiologist. For each exam, information referent to the at most first 5 detected regions was obtained. In this data set in particular, there were in total 10,478 MLO regions and 10,343 CC regions.

Each MLO (similarly, CC) region is an object of type *region*, and its corresponding features were directly incorporated as object attributes. In detail, the features used here included information such as *distance to skin*, *size* – the area of the region, *contrast*, *focal mass* – describing the existence of a circumscribed lesion, *spiculation* – pattern of straight lines directed toward the centre of a lesion (which is indicative for malignancy) and *linear structure* – indicating the presence of normal breast tissue. In total we had 60 features for each region.

The data graph given in Fig. 2(b) shows the different objects in the database labelled according to their type (patient, exam, breast, view, region). Notice that, even though not depicted in the figure, the CC regions are also linked to the regions in the MLO view of the same breast.

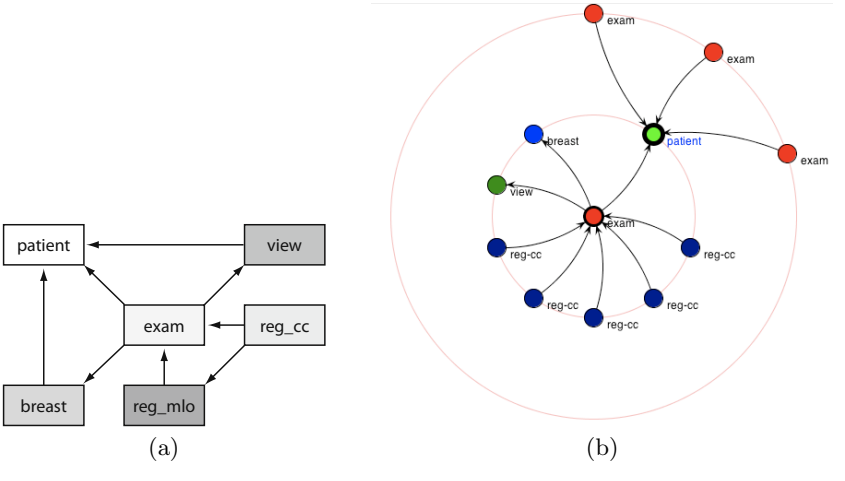


Fig. 2. Design of the breast cancer domain: (a) data scheme and (b) data graph

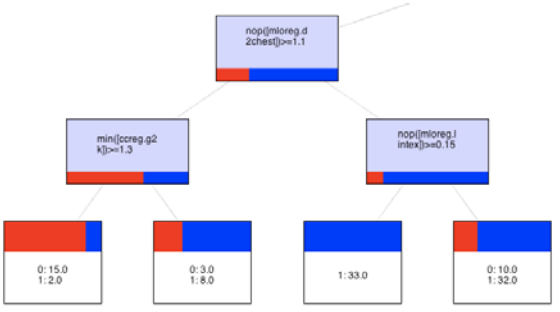


Fig. 3. A subpart of a relational probability tree; the grey regions at the top of the leaves correspond to probabilities

4.2 Mammogram Relational Probability Tree

Learning of an RPT looks not only into the object whose attribute is to be estimated, but also takes into account the effect of related objects. For instance, in order to predict the level of suspiciousness of a detected MLO region, the model takes into account the features of the given region, but also the attributes of the CC regions to which the MLO region is linked. In an RPT, root and internal nodes test the value of an object's attribute, and a leaf node establishes the probability estimation of an instance that reaches it.

As an example, consider the (sub)tree shown in Fig. 3. The first node shown asks for the value of attribute *d2chest*, i.e., distance to chest, of the considered MLO region: if this value is greater than or equal to 1.1, then the next node checks whether the value of the attribute *lintex* (linear texture) is greater than or equal to 0.15: if yes, with probability 0.76 this is classified as a cancerous regions; otherwise, the region is certainly cancerous. If *d2chest* is less than 1.1 then the minimum value of *g2k* (focal mass) is tested. Note that this refers to the value of linked CC regions.

5 Learning Relational Mammogram Models

We used the database described in Section 4.1. Here we report results on building relational probability trees taking into account some of the available features of an MLO region in relationship to its linked CC regions. Models were learned using the Proximity software [1].

Different models were build using different sets of features and different depth on the relational probability tree. Despite the different settings, model accuracy was usually high. This was due to the fact that the distribution of the data in terms of non-cancerous and cancerous patients was highly unbalanced. When we looked, however, at the classification results in terms of the area under the ROC curve (AUC) [4], we were able to better evaluate and compare different models.

Using a (sub)set of the 20 mammographic features of a region – for both MLO and its linked CC regions, we obtained a model partially such as shown in Fig. 3. This model of probability tree of maximal depth 5 had an AUC equal to 0.7536.

In previous models ([5,6]) we included the features *FPLevel* and *likelihood* – features calculated by the CAD system. A score for suspiciousness, i.e., the feature *likelihood*, for a region is computed based on the region's features, and converted into the false positive level (*FPlevel*), i.e., the average number of normal regions with the same or higher suspiciousness score. The inclusion of these features allowed us to obtain an AUC of 0.8728.

However, our goal was not only to obtain further improvement on the suspiciouness level of detected regions, but to use all the relational information available in order to better understand the process of image interpretation itself. By exploiting the information among features and among regions we aimed at achieving a better modelling of mammographic image interpretation. We compared the values obtained for each case, given the probability tree model against the level of suspiciousness calculated by the single-view CAD system. The average per case AUC value was 0.81 for the RPT model and 0.75 for the CAD system. When comparing the RPT predictions against the false-positive level of the CAD system, the performance of both in terms of accuracy was similar.

6 Final Remarks

In this paper, we presented ways to exploit relational probabilistic models of mammograms, a natural and comprehensible way of representing the various domain entities, as well as the uncertain relations among those. When analysing mammograms the ultimate goal is the interpretation of an exam, taking into account not only the characteristics of isolated regions, but also the breast image as whole as well as previous exams (if any) and the RPT models show a good performance in such analysis.

References

1. Neville, J., Jensen, D.: Relational dependency networks. Journal of Machine Learning Research (2007)
2. van Engeland, S.: Detection of Mass Lesions in Mammograms by Using Multiple Views. PhD thesis, Radboud University Nijmegen (2006)
3. Timp, S.: Analysis of Temporal Mammogram Pairs to Detect and Characterise Mass Lesions. PhD thesis, Radboud University Nijmegen (2006)
4. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36 (1982)
5. Velikova, M., Lucas, P., Ferreira, N., Samulski, M., Karssemeijer, N.: A decision support system for breast cancer detection in screening programs. In: Proceedings of the 18th European Conference on Artificial Intelligence (2008)
6. Ferreira, N., Velikova, M., Lucas, P.: Bayesian modelling of multi-view mammography. In: Proceedings of the ICML Workshop on Machine Learning for Health-Care Applications (2008)

Data-Efficient Information-Theoretic Test Selection

Marianne Mueller[1], Rómer Rosales[2], Harald Steck[2], Sriram Krishnan[2], Bharat Rao[2], and Stefan Kramer[1]

[1] Technische Universität München, Institut für Informatik, 85748 Garching, Germany
[2] IKM CAD and Knowledge Solutions, Siemens Healthcare, Malvern PA 19335, USA

Abstract. We use the concept of conditional mutual information (MI) to approach problems involving the selection of variables in the area of medical diagnosis. Computing MI requires estimates of joint distributions over collections of variables. However, in general computing accurate joint distributions conditioned on a large set of variables is expensive in terms of data and computing power. Therefore, one must seek alternative ways to calculate the relevant quantities and still use all the available observations. We describe and compare a basic approach consisting of averaging MI estimates conditioned on individual observations and another approach where it is possible to condition on all observations at once by making some conditional independence assumptions. This yields a data-efficient variant of information maximization for test selection. We present experimental results on public heart disease data and data from a controlled study in the area of breast cancer diagnosis.

1 Information Maximization for Medical Test Selection

Consider a collection of patient records containing data generally indicative of various clinical aspects associated to the patient, e.g., patient demographics, reported symptoms, results from laboratory or other tests, and patient disease/condition. Let us define each single element (source) of patient data as a random variable V_m, e.g., patient age (demographics), presence of headache (symptom), blood pressure (test result), or occurrence of diabetes (disease). In this paper, we consider the set of variables $V = \{V_1, \ldots, V_M\}$ and we assume each variable to be discrete with a finite domain. For a given patient, a few of these variables may have been observed while others may have not. In some cases, we know such values of the variables with some probability.

Let us now consider the time when a diagnosis needs to be made for a specific patient, for which some of these variables have been observed, denoted as background attributes $Z_i \in V$, for $i \in \{1, ..., k\}$; e.g., demographic information and patient symptoms. The remaining $M - k$ variables denoted X_j have not been observed yet; e.g., the result of a lab test. Finally, for the patient, let Y denote a variable in which we are ultimately interested, but that we have not observed, such as the occurrence of cancer. This variable may not be observable directly

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 410–415, 2009.

or it may be observable at a high cost. Thus, we prefer to rely on other sources of information and try to infer the value of this variable.

Formally, we want to optimize $X^* = \arg\max_j I(X_j, Y|Z_1 = z_1, \ldots, Z_k = z_k)$, where the quantity I is the mutual information (MI) between our variable of interest Y and the (test) variables whose values we could potentially obtain X_j, conditioned on the fact that we already know $Z_1 = z_1, \ldots, Z_k = z_k$. Using the definition of MI [1] and the shorthand $\mathbf{z} = (z_1, \ldots, z_k)$ to denote the assignment of multiple variables, the function of interest is:

$$I(X_j, Y|\mathbf{z}) = \sum_{x_j} \sum_{y} P(x_j, y|\mathbf{z}) \cdot \log \frac{P(x_j, y|\mathbf{z})}{P(x_j|\mathbf{z}) \cdot P(y|\mathbf{z})} \tag{1}$$

Assuming the variables follow a multinomial joint probability distribution that can be reliably estimated, this problem can be solved by testing each of the potential candidate variables individually to see which provides the most information.

In order to arrive at a diagnosis with high certainty, one observation may not be enough. In some cases, we may be allowed to observe more than one variable. Thus, this process can be repeated iteratively until a user-defined precision or confidence level is achieved. For example, this limit can be defined in terms of the amount of information that is left in the variable of interest (entropy). Once a variable X_i has been tested and observed, it can be incorporated as part of the background knowledge and the maximization problem is updated: $\arg\max_{j \neq i} I(X_j, Y|x_i, \mathbf{z})$.

Ideally, the quantity to optimize is MI conditioned on all available data or observed variables. However, in practice this may be difficult because in general the conditional joint probabilities $P(x_j, y|z_1, z_2, \ldots, z_k)$ cannot be properly estimated from limited data for large k. The data required to properly estimate the conditional joints may grow exponentially with k. In addition, the number of unobserved variables plays an important role in terms of computational complexity (which can also grow exponentially with the number of unobserved variables).

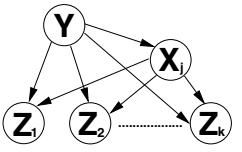
2 Combining All Available Background Information

Since with limited data our estimate of $P(x_j, y|z_1, z_2, \ldots, z_k)$ will not be accurate, we seek a *data-efficient* method – one whose data requirements do not grow exponentially with k – to compute or approximate P and thus be able to use all available background information to decide what test should be chosen.

2.1 Information Averaging

A simple heuristic consists of averaging the information conditioned on each background variable Z_i separately, $I(X_j, Y|\mathbf{z}) \approx \frac{1}{k} \cdot \sum_{i=1}^{k} \sum_{z_i} \alpha(z_i) I(X_j, Y|z_i)$, where $\alpha(z_i)$ is equal to the prior $P(z_i)$ if z_i is not observed and $\alpha(z_i) = 1(z_i)$ if observed. $1(z_i)$ is equal to one if z_i is the observed value for Z_i and zero otherwise. We will use this as our baseline method.

Table 1. Bayes net and approximations of probability distribution and entropy given $\mathbf{z}$



	Averaging	Joint Model (Q)
$P(y\|x_j,\mathbf{z})$	$\frac{1}{k}\cdot\sum_{i=1}^{k} P(y\|x_j,z_i)$	$\frac{P(x,y)\cdot\prod_{i=1}^{k} P(z_i\|x,y)}{\sum_{y\in Y} P(x,y)\cdot\prod_{i=1}^{k} P(z_i\|x,y)}$
$H(Y\|x_j,\mathbf{z})$	$\frac{1}{k}\sum_{i=1}^{k} H_P(Y\|x_j,z_i)$	$H_Q(Y\|x_j,\mathbf{z})$

One way to think about this heuristic is to consider k independent models involving Y, X_j, Z_i of the form $P(y, x_j, z_i) = P(z_i|x_j, y)P(x_j, y)$. MI can be computed for each model separately given the observed data. Similar to model averaging, the MI of X_j about Y can be found by averaging the individual information values, giving rise to the above equation.

2.2 Exploiting Conditional Independence Assumptions

We have so far assumed a very general model of the data, specifically a full multinomial distribution with a large number of degrees of freedom. However, if we relax this assumption, we can find a family of models for which the computation of the mutual information of interest is computationally and data efficient. These models make stronger conditional independence assumptions (simpler joint models), so that when the z_i's are observed, the computation of $I(X_j, Y|\mathbf{z})$ does not have large data requirements.

We consider the Bayes net in Table 1, where the background information Z_i depends on the test result X_j and the actual disease status Y. This network should not be viewed as reflecting causality, but simply as a statistical model. We obtain a new probability distribution Q:

$$Q(x_j, y, \mathbf{z}) = P(x_j, y) \cdot \prod_{i=1}^{k} P(Z_i = z_i|x_j, y). \tag{2}$$

The above Bayes network defines another multinomial model with the particular advantage that for the MI computations above, it only requires computing joint distributions of at most three variables, even if we do not know what variables Z_i will be observed beforehand. This is beneficial since we indeed do not know what variables will be observed for a particular case (patient).

For learning the model, we are interested in finding $P(z_i|x_j, y)\forall i \in \{1, ..., k\}$ and $P(x_j, y)$ since these distributions fully define the model. Using the maximum likelihood criterion we can see that: $P(z_i|x_j, y) = \frac{count(Z_i=z_i, X_j=x_j, Y=y)}{count(X_j=x_j, Y=y)}$ and $P(x_j, y) = \frac{count(X_j=x_j, Y=y)}{count(X_j=any, Y=any)}$. Inference can be performed efficiently also [2].

In summary, we have found a method that allows us to use all the available background information for deciding what test to perform. For this we have relaxed the model assumptions and use a class of models whose data requirements

are more practical. Inference and learning are computationally efficient in these models as well.

3 Validation and Results

First, we test our approaches on public heart disease data (HD) consisting of 920 patients and 14 attributes (two background attributes, 11 observable test attributes, and one binary class attribute).[1] On this dataset each patient is an instance. The variable of interest Y expresses if the patient has heart disease (411 patients) or not (509 patients). In a second step, we perform experiments on the breast cancer diagnosis data (BC) described in detail in the long version of this paper [2] (16 background attributes, 4 observable test attributes, and one binary class attribute). Here, we use lesions as instances. From 132 patients we analyze 216 biopsied lesions, 134 malignant and 82 benign. We use the biopsy result as the variable of interest Y. Note that the number of data points is small given the dimensionality of the data, which is quite common in controlled medical studies where the cost of data gathering is usually high.

We validate the two approaches considered above for each instance in a leave-one-out validation.[2] First, we determine for each instance which test X^* maximizes the information given the patient specific attributes $\mathbf{z}$. After observing the selected test $X^* = x_j$, we decide for the most likely diagnosis $y^* = arg\max_{y \in Y} P(y|x_j, \mathbf{z})$. We say the instance is *correctly assessed*, if y^* equals the actual state of disease of the instance, and *incorrectly assessed*, if the diagnosis was different. This accuracy measure could be easily improved by training a classifier on the features $\{X_j, Z_1, \cdots Z_k\}$ and predicting Y for each test instance.

The two approaches provide two different estimations of $I(X, Y|\mathbf{z})$. Our validation criteria (correctness, certainty, and information gain) additionally require the computation of $P(y|x_j, \mathbf{z})$ and the conditional entropy $H(Y|x_j, \mathbf{z})$. The discussion in Section 2 showed that it is not possible to compute this for large k. Therefore, we use the approximations shown in Table 1. H_P (H_Q) denotes the entropy of the probability distribution P (Q). To get a better grading of the results of our approaches, we compare it to two different ways of test selection. We evaluate: selecting a test *proposed* by information maximization, selecting one of the possible tests at *random* and selecting the *best*[3] test. Because of the different approximations (see Table 1), the performance of the *random* and *best* methods differs for each approach.

Correctness: Table 2 compares the ratio of instances that were assessed correctly: On both data sets it holds that if we select an examination at random,

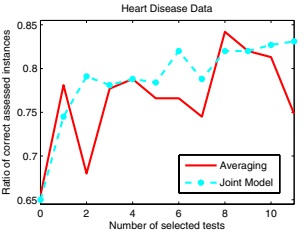
[1] http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[2] We only consider instances with complete information on the test attributes as test instances (278 for HD and 138 for BC) because it is not possible to evaluate the cases where the result value of the selected test is missing.

[3] We determine the performances of all tests beforehand and select the test with the best performance (note that this is only possible in this controlled experiment).

Table 2. Ratio of correctly assessed instances. On the left: when iterating the test selection

Ratio of correctly assessed instances						
	Heart Disease			Breast Cancer		
	proposed	*best*	*random*	*proposed*	*best*	*random*
Av.	0.781	0.996	0.681	0.657	0.846	0.696
J.M.	0.745	0.989	0.665	0.748	0.944	0.691



in two thirds of the cases the decision about the status of disease is correct. On the HD data, selecting a test by information maximization leads to an improvement of about one tenth. On the BC data, selecting a modality by information maximization leads to an improvement in the second approach where we assume a simple joint model. Hence, the latter appears to give a more useful estimate of $I(X, Y|\mathbf{z})$. The difference of the two approaches on the HD data only becomes apparent when iterating the test selection process and adding the previously obtained test results as background knowledge, otherwise we have only two background attributes (Table 2). The second approach performs more stable and mostly outperforms the first approach.

Certainty of Decision: Results [2] show that on both data sets the certainty for the correctly classified lesions is higher (lower entropy) than for the incorrectly classified lesions. On the BC data, the difference is large for the joint model approach, but for model averaging these quantities are very similar. This indicates that the proposed joint model is better suited at modeling the information content in the test/decision variables.

Actual Information Gain: We compare the entropy of $P(Y|\mathbf{z})$ before performing a test with the entropy of $P(Y|x_j, \mathbf{z})$ after observing the test result $X_j = x_j$ (see [2]). Comparing both approaches to random, the joint model approach achieves a better result than the baseline approach for both data sets.

4 Conclusion

We have employed the concept of MI to address the problem of choosing tests efficiently[4]. We applied this to a problem in medical diagnosis. While MI is a well-understood concept, it is hard to calculate accurately for general probability models in practice due to small datasets and the underlying computational complexity. We have experimentally shown how certain model assumptions can help circumventing these problems. Making these assumptions, we obtain a

[4] Due to lack of space, the reader is referred to [2] for a discussion of related work.

comparatively data-efficient variant of test selection based on information maximization. Results indicate that the proposed joint model outperforms the information averaging approach by comparing the performance of each approach relative to a *random* and a *best* selection.

References

1. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley Interscience, Hoboken (1991)
2. Mueller, M., Rosales, R., Steck, H., Krishnan, S., Rao, B., Kramer, S.: Data-Efficient Information-Theoretic Test Selection, Technical Report TUM-I0910, Institut für Informatik, TU München (2009)

Effect of Background Correction on Cancer Classification with Gene Expression Data

Adelaide Freitas*, Gladys Castillo**, and Ana São Marcos***

Department of Mathematics, University of Aveiro, Portugal
adelaide@ua.pt, gladys@ua.pt, anasaomarcos@ua.pt

Abstract. This paper empirically compares six background correction methods aimed at removing unspecific background noise of the overall signal level measured by a scanner across microarrays. Using three published cDNA microarray datasets we investigated the effect of background correction on cancer classification in terms of the predictive performance of two classifiers (k-NN and support vector machine with linear kernel) induced from microarray data where a particular background correction method is applied, individually and in combination with a single-bias or double-bias-removal normalization method.

1 Introduction

Microarray technology [6,7] allows simultaneous measurement of expression levels of thousands of genes in a single experiment. A microarray consists of a glass slide with spots (where probes are attached) arranged into several print-tip(PT) groups according to a particular layout. In cDNA microarrays experiments, two samples of mRNA labelled with distinct (*red* Cy5 and *green* Cy3) fluorescent dyes (*target*) are simultaneously hybridized to a microarray spotted with a particular DNA sequence (*probe*). The microarray is then scanned and the resulting image is processed to obtain a quantification of the transcript abundance at each spot, a amount proportional to the total fluorescence, i.e., the red and green fluorescence intensities (R, G). The *relative expression level* for each gene spotted is the log ratio $M = \log_2 R/G$. Given m number of tissue samples and n number of genes, the data generated by microarray experiments is a $m \times n$ matrix of gene expression levels. In addition, for each sample is given its classification (e.g. presence or absence of a tumor). The task is to build a classifier from microarray data that provides the best classification for future tissue samples.

Each step in microarray experiments can introduce variability in the measured intensities that affects the quality of the raw data. Background correction (BC) and normalization (NM) are two pre-processing steps aimed at cleaning raw data at undesirable variations due to technical factors, but trying to retain the

* Supported by FCT, POCI (EC fund FEDER) through Research Unit Mathematics and Applications.
** Supported by FCT, POCI (EC fund FEDER) through CEOC.
*** Grant PTDC/MAT/72974/2006 (FCT).

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 416–420, 2009.

intrinsic biological variations[8]. In this work we aim to investigate the effects that BC has on the predictive performance of classifiers induced from microarray data. Previous related work [1,4] have been more focused on evaluating the effects of BC on the detection of differentially expressed genes or the precision of point estimates. Instead, our study is more related to the study given in [8] that evaluates NM methods for supervised classification. The rest of the paper is organized as follows. Section 2 briefly reviews the BC methods. The experimental study to analyze the effect of BC on cancer classification is discussed in Section 3. Finally, Section 4 contains the conclusions and future work.

2 Background Correction

Background correction aims to remove unspecific background noise of the overall signal measured by a scanner. Intensities measured for each spot can include a contribution not specifically due to the hybridization of the target to the probe, but due to other causes (e.g. the presence of other chemicals on the glass) [7]. Since any observed signal is affected by a background signal, we can obtain more accurate gene expression levels by measuring and removing it. All the datasets used herein were generated by the GenePix Pro software. A local method to estimate the bakcground signal is implemented. For each spot the pixels are classified as *foreground intensities* or *background signals* according to whether they are inside the spot or in the surrounding area. The foreground intensities (R_f, G_f) and the background signals (R_b,G_b) are estimated (e.g. using the mean values). Next, to obtain an accurate measure of true intensities (R, G), a BC on the observed values (R_f, G_f) is performed. A complete overview of BC methods is given in [4]. Here we briefly describe the methods used in our study[1]:

1. **noBC:** No BC is performed, i.e., $R = R_f$ and $G = G_f$.
2. **Subtraction(sub):** It subtracts the background from the foreground values, i.e., $R = R_f - R_b$ and $G = G_f - G_b$.
3. **Half:** Any intensity less than 0.5 after subtraction is set to 0.5, i.e., $R = 0.5$, if $R_f - R_b < 0.5$; $R = R_f - R_b$, otherwise.
4. **Minimum(min):** Any intensity, zero or negative after subtraction, is set to the minimum of the positive corrected intensities for that array, i.e., $R = R_f - R_b$ if $R_f - R_b \geq 0$; $R = \min_{1 \leq i \leq N}(R_{fi} - R_{bi} : R_{fi} - R_{bi} \geq 0)$, otherwise.
5. **Edwards(edw):** A smoothing function is used if the difference between the foreground and background is less than a given threshold value [2], that is, $R = R_f - R_b$ if $R_f - R_b \geq \delta$, $R = \delta e^{1-(R_b+\delta)/R_f}$, otherwise.
6. **Normexp(nexp):** Assuming $R_f = R + R_b$ and $G_f = G + G_b$, the intensities (R, G) are calculated as the expected values by considering that true signals follow an exponential distribution and background signals are normally distributed.

[1] Due to lack of space, in some methods we limit only to present the formulae for the red fluorescence intensities. The green intensities are calculated similarly.

3 An Experimental Study with Microarray Data

We evaluated the effect of BC on the performance of two classifiers: k-NN and support vector machines(SVM) using leave-one-out cross validation (LOO-CV). We used three microarray datasets[5]: *lymphoma* (3 classes), *liver cancer* (2 classes) and *lung cancer* (5 classes). For each dataset we assessed 36 different pre-processing strategies combining a BC method followed by a NM method. All the NM and BC methods are summarized in Table 1. We implemented three single-bias removal (*one-step*) and two double-bias-removal (*two-steps*) NM strategies. These methods use robust locally weighted regression (*loess*) for adjusting the intensity and/or spatial effect due to different sources of dye biases. In addition, NM methods are applied at a local or global level depending on whether only the spots in a PT group or the entire microarray is considered for adjustment. Each pre-preprocessed data associated to a pair of methods (BC, NM) is obtained as follows: *i)* for each spot, BC is applied to estimate (R, G) and obtain the M-values; *ii)* M-values are normalized; *iii)* only those probes that are present in all the microarrays are retained; the M-values for a same probe are averaged; *iv)* M-values are centered and scaled to a unit norm; *v)* missing values are imputed. As a result, the resulting numbers of samples, m, and features, n, are: $m = 108(68/31/9)$ and $n = 7079$ for Lymphoma; $m = 207(131/76)$ and $n = 21901$ for Liver and $m = 65(39/13/4/4/5)$ and $n = 22646$ for Lung. The LOO-CV procedure for k-NN was implemented in R using class/knn.cv(k = 3,...,10) and class/knn(k = k*) where the optimal k* was also selected via LOO-CV. We used the SVM learner implemented in the operator LibSVMLerner with a linear kernel in RapidMiner[3].

Table 2 depicts the LOO-CV error rates for the 36 pre-processing methods for each dataset. Regarding SVM, sub seems to be the best BC method followed by

Table 1. BC and NM methods used in the comparative study

Methods	Bioconductor R package/function(parameters)	Corrected Intensity based on
NB	*limma*/**backgroundCorrect**(RGlist,method="none")	No BC
Subtraction	*limma*/**backgroundCorrect**(RGlist,method="sub")	Subtraction
Minimum	*limma*/**backgroundCorrect**(RGlist,method="min")	Truncated Subtraction
Half	*limma*/**backgroundCorrect**(RGlist,method="half")	Truncated Subtraction
Edwards	*limma*/**backgroundCorrect**(RGlist,method="edwards")	Model
Normexp	*limma*/**backgroundCorrect**(RGlist,method="normexp")	Model
NN	*marray*/**maNorm**(data,norm="none")	NN
IGloess	*marray*/**maNorm**(data, norm = "loess", subset = TRUE, span = 0.4)	**I**ntensity **G**lobal **loess** (IG)
ILloess	*marray*/**maNorm**(data, norm = "printTipLoess", subset = T, span = 0.4)	**I**ntensity **L**ocal **loess** (IL)
SLloess	*marray*/**maNormMain**(data, f.loc = list(maNorm2D(g = "maPrintTip", subset = T, span = 0.4)))	**S**patial **l**ocal **loess**(SL)
IGloessSLloess	*marray*/d=**maNorm**(data,norm="loess",subset=TRUE, span=0.4)/**maNormMain**(d,f.loc=list(maNorm2D(g= "maPrintTip",subset=T,span = 0.4)))	**I**ntensity **G**lobal **loess** followed by **S**patial **L**ocal **loess** (IG-SL)
ILloessSLloess	*marray*/d=**maNorm**(data, norm="printTipLoess",subset= T,span=0.4)/**maNormMain**(d,f.loc=list(maNorm2D(g= "maPrintTip",subset=T,span=0.4)))	**I**ntensity **L**ocal **loess** followed by **S**patial **L**ocal **loess** (IL-SL)

Table 2. LOO-CV error rates (%) for the 36 (BC, NM) strategies per dataset

Data set	Method	k-NN						SVM					
	BC	NB	sub	half	min	edw	nexp	none	sub	half	min	edw	nexp
Lymphoma	NN	26.85	21.29	23.14	21.29	25.92	19.44	16.67	14.81	14.91	14.81	14.81	13.89
	IG	12.96	16.66	20.37	16.66	22.22	11.1	7.41	5.56	12.96	5.56	11.11	7.41
	IL	12.03	15.74	17.59	15.74	21.29	11.1	5.56	5.56	12.96	5.56	9.26	7.41
	SI	16.66	18.51	17.59	18.51	20.37	12.96	12.96	6.48	12.96	6.48	12.04	9.26
	IG-SL	12.03	16.6	20.37	15.74	21.29	9.25	5.56	6.48	12.96	6.48	11.11	8.33
	IL-SL	9.25	16.6	17.59	16.6	21.29	11.11	5.56	6.48	11.11	6.48	10.19	6.48
Lung	NN	35.38	26.15	36.92	36.92	35.38	35.38	29.23	23.08	32.31	32.31	23.08	27.69
	IG	23.07	29.23	41.53	41.53	41.53	32.3	21.54	20	27.69	26.15	23.08	21.54
	IL	20	27.69	38.46	38.46	41.53	29.23	21.54	18.46	24.62	24.62	23.08	21.54
	SL	32.3	27.69	29.23	29.23	36.92	30.76	23.08	18.46	24.62	24.62	23.08	23.08
	IG-SL	24.61	30.76	35.38	36.92	44.61	29.23	20	18.46	24.62	26.15	23.08	21.54
	IL-SL	21.53	24.61	41.53	41.53	35.38	27.69	20	18.46	24.62	24.62	23.08	21.54
Liver	NN	16.9	16.42	17.39	17.39	17.87	16.90	3.86	3.86	3.86	3.86	3.86	3.86
	IG	13.52	15.94	14.97	12.56	11.59	13.04	3.86	3.86	3.38	3.86	3.38	3.38
	IL	13.04	11.59	15.94	15.94	14.00	13.52	4.83	3.86	3.86	3.86	4.35	3.38
	SL	17.39	14.49	14.00	14.00	16.42	14.49	4.83	4.35	4.35	4.35	4.35	3.38
	IG-SL	19.80	11.59	14.97	15.94	15.94	10.14	4.35	3.86	3.86	4.35	3.86	3.38
	IL-SL	15.45	12.07	14.00	14.00	14.00	9.66	4.83	3.86	3.86	3.86	3.86	2.90
	p/n/t		8/10/0	5/13/0	6/12/0	5/12/1	10/6/2		14/2/2	6/10/2	7/8/3	8/8/2	8/5/5

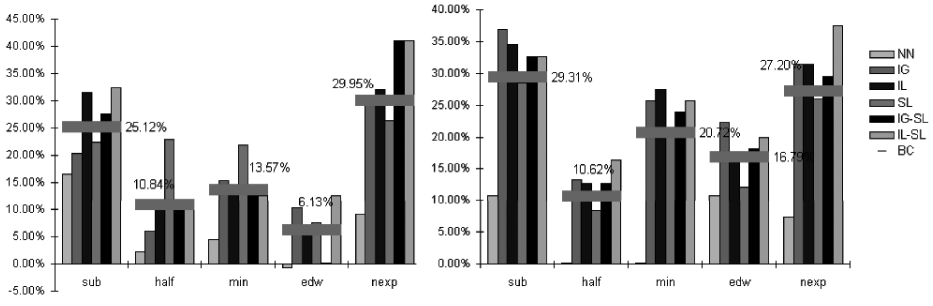


Fig. 1. Bar plots of ARRs per classifier: k-NN (left), SVM (right). Vertical bars represent the ARR per (BC,NM) grouped by BC methods. Horizontal bars represent the ARR for each BC method.

half and nexp. For k-NN, instead, similar results for nexp and sub are observed. As shown in the last line, for k-NN there are more positive differences for nexp (10) and sub (8). For SVM, sub presents more gains (14). In order to assess the significance of the contribution to the improvements of the performance due to the application of BC and NM methods, each of them separately and also due to the interaction between both methods, BC↔NM, we assumed a two way additive model for the LOO-CV error $e(i, j)$ of a particular combination BC $= i$, NM $= j$. By applying a parametric ANOVA we have not found significative difference on the performance for the five BC methods in comparison to the baseline NB ($p - values > 0.8$). However, residual values of the adjusted model were not normally distributed, which suggests no warrantee from these conclusions. To assess how much gain in performance could bring the combination (BC $= i$, NM $= j$) we employed the *reduction rate*. From the additive model of the error, the reduction rate can be defined as $RR(i, j) = RR(i, \cdot) + RR(\cdot, j) - RR(i \leftrightarrow j)$ where the first term represents the reduction due to BC, the second term due

to NM and the last term due to the interaction between them. Taking $e_{bs} \equiv e(\mathsf{NB}, \mathsf{NN})$ as the baseline error (neither BC nor NM were performed) we obtain:

$$RR(i,j) = \frac{e_{bs} - e(\cdot, j)}{e_{bs}} + \frac{e_{bs} - e(i, \cdot)}{e_{bs}} - \frac{e_{bs} - (e(\cdot, j) + e(i, \cdot) - e(i, j))}{e_{bs}}$$

Figure 1 illustrates the *average reduction rate* (ARR) for each combination of methods (i, j) (the averages of $RR(i, j)$) against the ARR for each $\mathsf{BC} = i$ (the averages of $RR(i, \cdot)$) over the three datasets. For both classifiers, by analyzing the vertical bars, the combined methods (i, j), for $i \in \{\mathsf{sub}, \mathsf{nexp}\}$, $j \neq \mathsf{NN}$, show higher ARR, while by observing the horizontal bars, the BC methods that present higher ARR are sub, nexp and min, thus indicating a better contribution to the predictive performance.

4 Conclusions and Future Work

We compared six BC methods in combination with six NM methods in the context of cancer classification in order to evaluate the effect of BC on the performance of classifiers induced from microarray data. Results show that sub and nexp seems to be the best methods. The reduction rate associated to the additive model, as herein proposed, allows quantifying the effect of BC methods on the performance of classifiers. However, the results obtained from ANOVA give us some indications that the application of BC methods might not bring significant gains on the classifier performance. To obtain more reliable conclusions, we plan to extend this study with more datasets, classifier models and suitable statistical tests for result analysis.

References

1. Bolstad, B.: Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization. PhD Dissertation. University of California, Berkeley (2004)
2. Edwards, D.: Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics 19(7), 825–833 (2003)
3. Mierswa, I., Wurst, M., Klinkenberg, R., et al.: YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Proc. of the 12th ACM SIGKDD (2006)
4. Ritchie, M.E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., Smyth, G.K., et al.: A comparison of background correction methods for two colour microarrays. Bioinformatics 23(20) (2007)
5. Stanford Microarray Database, http://genome-www5.stanford.edu/
6. Schena, M., Shaon, D., Heller, R., Chai, A., Brown, P., Davis, R., et al.: Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proceedings of the National Academy of Sciences of USA 93 (1996)
7. Yang, Y.H., Buckley, M.J., Speed, T.P.: Analysis of cDNA microarray images. Brief Bioinform 2, 341–349 (2001)
8. Wu, W., Xing, E., Myers, C., Mian, I.S., Bissel, M.J., et al.: Evaluation of normalization methods for cDNA microarray data by k-NN classification. BMC Bioinformatics 6, 191 (2005)

On Quality of Different Annotation Sources for Gene Expression Analysis

Francesca Mulas[1,2], Tomaz Curk[3], Riccardo Bellazzi[1,2], and Blaz Zupan[2,3,4]

[1] Dipartimento di Informatica e Sistemistica, University of Pavia, Italy
[2] Centro Interdipartimentale di Ingegneria dei Tessuti, Pavia, Italy
[3] Faculty of Computer and Information Science, University of Ljubljana, Slovenia
[4] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, USA
fra.mulas@gmail.com, tomaz.curk@fri.uni-lj.si
riccardo.bellazzi@unipv.it, blaz.zupan@fri.uni-lj.si

Abstract. Mining of biomedical data increasingly relies on utility of knowledge repositories. In gene expression analysis, these are often used for gene labeling with an assumption that similarly annotated genes have similar expression profiles. In the paper we use this assumption to craft a method with which we scored six different annotation sources (*e.g.*, Gene Ontology, PubMed, and MeSH annotations) for their utility in gene expression data analysis. Experiments show that the sources that include manual curation perform well and, for instance, score better than automatic annotation from gene-related PubMed abstracts. We also show that there is no clear winner, pointing at the need for methods that could successfully integrate annotations from different sources.

1 Introduction

Recently, the field of data analysis in bioinformatics is shifting from data-centric to integrative, where the findings from different experimental data sets are combined with any potentially useful assertions from available knowledge-sources. In gene expression analysis, one of the currently most popular approaches is that of gene set enrichment analysis [1], where instead of ranking the genes by differential expression the method ranks gene sets. In principle, this increases the number of measurements in ranked item and with this the robustness of predictions. Genes sets come from different sources, and are most often those that share annotations in Gene Ontology or are referenced in the same KEGG pathway. With increasing number of such knowledge-bases one of the questions is related to their quality. How similar are these knowledge-based gene annotation sources in terms of their predictive quality? Should we rely more on semi-manual annotations in Gene Ontology, or trust textual sources and automatic inference of text-based annotations?

In this paper, we lay out the methodology that can be used to assess the predictive quality of knowledge sources for gene labeling and annotation. We focus on the utility of these labels in gene expression analysis. The principal idea

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 421–425, 2009.

of the proposed approach is to combine the gene labels coming from a knowledge-source into a so-called gene annotation profile. Given a specific knowledge source and some gene expression data sets, genes with similar annotation profiles should have a similar expression profile. We used our method to rank six different types of annotation and test them on 19 gene expression data sets. The results expose the difference in potential utility of these sources, but also point to their diversity, thus encouraging further efforts in crafting methods for their integration.

2 Data and Methods

2.1 Data Sets

The study uses 19 *Mus musculus* gene expression data sets from Gene Expression Omnibus (GEO)[1] data repository[2] that include measurements at exactly two different experimental conditions and for each report the results on at least eight different samples. On average, these data sets include about 50.000 "genes" (either real *Mus musculus* genes or yet-to-be-mapped open reading frames), and measure their expression in anywhere from 8 to 23 samples.

2.2 Gene Annotation Sources

We have considered two knowledge repositories and four literature-based annotation sources to describe the gene annotation profile. *Gene Ontology* (GO) annotations include all GO terms related to a gene, excluding those inferred from expression patterns (IEP evidence code) [2]. *Gene Ontology Slim terms* (GO-Slim) include a small subset of most biologically-relevant concepts. Generic GO slim set as proposed by S. Mundodi and A. Ireland was used. *PubMed IDs* (Articles) include the IDs of all the papers related to a gene, obtained from Entrez Gene data base[3] at the National Center for Biotehnology Information (NCBI). Here we test the assumption that co-cited genes have similar expression patterns. *MeSH headings* (MeSH) annotation includes a set of Medical Subject Headings (MeSH)[4] from papers referencing a gene[5]. For each gene, we have constructed an annotation vector that counts the occurrence of MeSH terms in its related PubMed entries. For *Major MeSH headings* (MeSH-M) only major headings (main topics of the article) were considered. The last annotation included *words in the title and abstracts of related publications* (Words). From each gene-related PubMed entry we have extracted a set of words and compiled the annotation vector reporting on the proportion of related articles in which a specific word occurred.

[1] http://www.ncbi.nlm.nih.gov/sites/entrez?db=geo

[2] The data sets included are GDS1213, GDS1366, GDS1635, GDS1939, GDS1978, GDS2082, GDS2092, GDS2309, GDS2433, GDS2511, GDS2552, GDS2703, GDS2748, GDS2765, GDS2766, GDS2823, GDS2903, GDS3162, GDS3210.

[3] ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz

[4] http://www.nlm.nih.gov/mesh

[5] www.ncbi.nlm.nih.gov/entrez

2.3 Preprocessing and Gene Similarity Scoring

As a number of annotations were coming from NCBI's data bases, it was crucial to associate data's gene names to NCBI's gene IDs. The mapping was successful for about 40% of gene names, which were kept in the analysis. For Words, we additionally removed the stop words, numbers and short words and lemmatized the remaining words [3]. For each data set, we removed all annotations that appear in more than 30% of genes and those that appear in less than five genes. From each data set we then removed genes with less than five annotations of specific type. For each data set, the annotation data was then subject to a TF-IDF (Term Frequency, Inverse Document Frequency) transformation. TF-IDF ensures that terms, which are not specific to a gene and are not frequently used to describe a gene, have low weight in the annotation profile. Each annotation vector was normalized.

We used cosine similarity to compute the distance between two distinct gene annotation profiles. Euclidean distance was used to assess similarity between two expression profiles.

2.4 Scoring of Annotation Sources

Our goal was to test if genes with similar annotation profiles have similar expression profiles. To perform this analysis we carried out the following procedure. For each data set and for each gene annotation source we first built a *sampling null distribution*. We randomly selected 1.000 genes and for each computed the mean expression profile similarity to four other randomly selected genes. Next, we sampled 1.000 *seed genes* and for each selected four most annotation-similar genes. From these, we selected 20 seeds and their corresponding neighbors with highest average similarity scores, and we computed the average similarity between the seed expression profile and profiles of its four neighbors. This similarity was then compared to the null distribution, obtaining the score equal to the proportion of null-distribution seeds with smaller expression-based similarities to their set of random neighbors. The scores over 20 seeds are averaged, thus obtaining the *score for the annotation source*.

3 Results and Discussion

Table 1 shows average scores that were obtained with the procedure described above. All annotation scores were lower than 0.5, a threshold where the annotation source does not provide any useful information for expression analysis.

Table 1. Averages of the annotation scores and ranks obtained in the 19 data sets

	GO	GO-Slim	Articles	MeSH	MeSH-M	Words
$\bar{s_a}$	0.199	0.283	0.350	0.245	0.266	0.261
rank	2.053	3.684	5.105	2.947	3.474	3.737

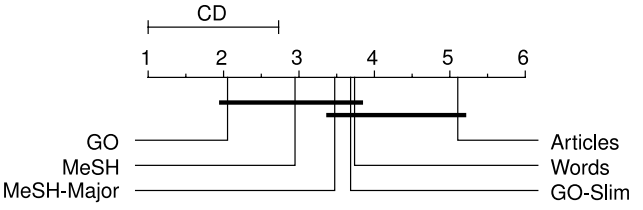


Fig. 1. Average ranks of annotation sources with critical distance-analysis [4] grouping the sources with no significant difference in ranking ($p = 0.05$)

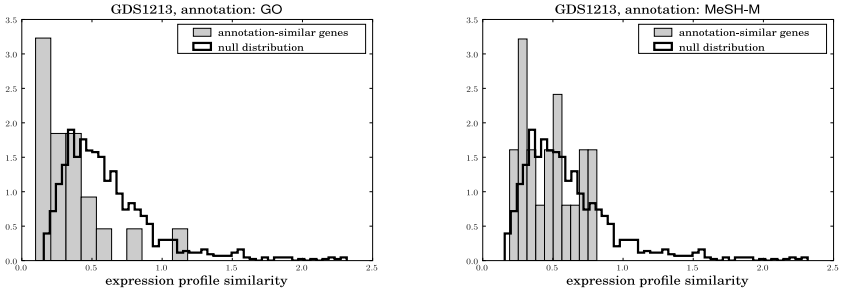


Fig. 2. Expression-based similarity distributions for 20 seed genes for GO and MeSH-M compared to null-distribution

According to Demsar [4], we ranked the annotation sources for each data set and report the average annotation source ranks. Average ranks, together with a critical rank distance (p=0.05) are displayed in Fig. 1. They indicate that there are some differences between the utility of annotation sources. GO had the highest average score and its performance is significantly better than the one provided by Articles. Moreover, GO had the best performance in 7 out of 19 data set. Rather interestingly, GO scored better than GO-Slim, which had an average score similar to Words. GO-Slim have low scores probably because the annotation is performed with less, usually more general, terms than in GO, which weakens the correspondence between expression and annotation. MeSH annotation scored significantly better than Words, showing that keywords that are manually assigned to publications represent useful information about genes. On the contrary, automatic extraction of words from abstracts can result in terms that do not really characterize the genes, this explaining the lower score of Words. Articles annotation showed the lowest average score.

A very interesting point that we observed from the results is that, except for Articles, each annotation score won for at least one data set. This suggests that an integration of annotations from different sources could result in a very good predictive quality.

The average scores reported in Table 1 are not to be confused with statistical comparison of the two distributions, that of the null distribution and the distribution of expression profile distances obtained from annotation-similar gene sets. Namely, testing the null hypothesis that the mean of the annotation-similar distribution is the same as our null distribution, the obtained p-values are much lower that the annotation scores. For instance, observe the two distributions for annotations GO and MeSH-M on the data set GDS1213 (Figure 2). There, the scores are $s_{\mathsf{GO}} = 0.207$ and $s_{\mathsf{MeSH-M}} = 0.386$, while corresponding p values are $p_{\mathsf{GO}} = 0.0003$ and $p_{\mathsf{MeSH-M}} = 0.0703$.

4 Conclusions

The paper investigated the utility of different gene annotation systems to predict gene co-expression, and for that proposed an original method for gene source annotation scoring and used it on a set of gene expression data sets from *Mus musculus*. The results confirm that knowledge sources that incorporate manual curation tend to be more reliable. In our experiments, for instance, GO was consistently one of the best predictors in the considered data sets. The ranks of annotation sources, however, vary from one data set to the other. This clearly points to the need for methods that would integrate the knowledge coming from different annotation sources, which is also a direction of our further research.

Acknowledgements. This work was generously supported by the "Fondazione Cariplo". RB was also supported by the FIRB project "ITALBIONET - Rete Italiana di Bioinformatica", and TC and BZ by the grants from Slovenian Research Agency (L2-1112, J2-9699 and P2-0209).

References

1. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci., USA 102(43), 15545–15550 (2005)
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25(1), 25–29 (2000)
3. Petrovic, S., Kolar, M., Saric, F., Demsar, J., Dalbelo Basic, B., Zupan, B.: orngTxt, Orange data mining tool, Add-On for Text mining (2008), http://www.ailab.si/orange/downloads.asp
4. Demsar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. J. Mach. Learn. Res. 7, 1–30 (2006)

An Architecture for Automated Reasoning Systems for Genome-Wide Studies

Angelo Nuzzo[1], Alberto Riva[2], Mario Stefanelli[3], and Riccardo Bellazzi[3]

[1] Centre for Tissue Engineering, University of Pavia, Pavia, Italy
[2] Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL, USA
[3] Department of Computer Science and Systems, University of Pavia, Pavia, Italy
{angelo.nuzzo,mario.stefanelli,riccardo.bellazzi}@unipv.it, ariva@ufl.edu

Abstract. The massive amounts of data generated by high-throughput experiments makes modern biomedical research a data-intensive discipline, shifting the research methodology from a hypothesis-based approach to a hypothesis-free one. A formal procedure should be defined to properly design a study, understand the outcomes and plan improvements for each task performed during the experiments. Such formal approach needs the identification of a high-level conceptual model of the knowledge discovery process occurring in genome-wide studies: this is what existing computational tools lack. Starting from an epistemological model of the discovery process proposed for diagnostic reasoning, we describe how the design and execution of modern genome-wide studies can be modelled using the same framework. We show the general validity of the model, how it can be instantiated to model typical scenarios of genome-wide studies, and how we use it to develop tools aimed at building semi-automated reasoning systems.

Keywords: Genome-wide studies, decision support system, reasoning models.

1 Introduction

Thanks to high-throughput technologies, a great number of measurements (hundreds to hundreds of thousands) can be simultaneously performed over the genome, so that a genome-wide scan is nowadays feasible. Genome-wide (GW) studies provide the unprecedented opportunity to obtain a large scale picture of what is contained (in case of genotyping experiments) or what is going on (in case of gene expression experiments) in many different loci in the genome at the same time. Thus, data-driven approaches for information (and then knowledge) extraction are enabled, so that research methodology shifts from a *hypothesis based* approach to a *hypothesis-free* one. Two challenging issues arise in this scenario: i) to develop software tools that properly and easily handle the huge amount of heterogeneous data produced, and ii) to properly interpret and understand results coming from the analysis of such data.

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 426–430, 2009.

The first of the two issues highlighted above has been addressed by developing several tools, the purpose of which is mainly focused on providing an efficient way of handling, integrating, manipulating and making all data easy to be explored by researchers [1] [2] [3] [4].

As concern results interpretation, a standard, formal procedure should be defined to properly design a study, understand the outcomes and plan improvements to improve each task performed during the experiment. Such formal approach needs the identification of a high-level conceptual model of the knowledge discovery process occurring in GW studies. This is what existing computational methods and tools lack. An epistemological model of the discovery process has already been formalized in the Artificial Intelligence in Medicine field, in order to provide a conceptual framework for the implementation of medical knowledge-based systems [5]. Following that model, in this paper we describe how the design and execution of modern GW studies can be modelled using the same framework, and we implemented a tool for each of these steps in order to perform a genome-wide association study.

2 The Model

Cognitive science studies pointed out that in a problem solving process experts use first to select a set of hypotheses and then they focus their efforts on testing and refining these hypotheses, braking the process in two phases: i) a hypotheses selection phase, in which an initial information is used to select possible hypotheses, and ii) a hypotheses testing phase, in which hypotheses selected in the previous step are used to forecast expected consequences, that should be matched with other (or new) available information in order to confirm the hypotheses from which they come [5].

Different particular inference types are involved in the general so-called Select and Test Model (ST model; a schema is shown in Fig. 1a). The first step of the process is *abstraction*: a set of solution feature is extracted starting from the initial data and information. Then *abduction* allows the hypotheses generation phase starting form the problem's features. The result of abduction is the definition of the *spaces of hypothetical solutions* of the problem, which have to be tested. These hypotheses have to be *ranked* to define and ordering of the following testing steps (ranking is based on some preference criteria which can be application-dependent or defined using prior knowledge). After hypotheses have been abducted and ranked, the testing phase starts to explore their consequences. *Deduction* allows us to derive from each candidate hypothesis what one expects to be true if that hypothesis is true. Once predictions have been derived from hypotheses, they need to be matched in order to choose the best hypothesis. Induction is able to match single statement to single statement and, therefore, to match a single statement derived as a prediction from a hypothesis with a single statement describing a portion of the available information. During this phase, induction corroborates those hypotheses whose expected consequences turn out to be in agreement with available information and refuses those which failed this test. The overall process is cyclic: once a set of expected consequences has been deducted, their evaluation may need new information to be acquired in order to either restrict the hypotheses set (induction) or refining existing hypotheses or generating new ones (new abstraction, thus starting another loop).

This description of the model is a high-level representation of the process in terms of the concept involved in it. If we consider a specific task, it has to be specialized, identifying which of the blocks remain valid and which is the specific content of each block. In the same work [5], authors showed model instances in three typical medical tasks: diagnosis, therapy and monitoring activities. In our opinion, modern post-genomic era is going through the same evolution, due to the same nature of the two processes: both of them are problem solving tasks, regarding two different domains. Thus, in the following paragraph we describe how the ST Model can be instantiated in the case of genome-wide studies.

3 Application on Genome-Wide Association Studies

Association studies aim at finding statistically significant differences in the distribution of a set of markers between a group of individuals showing a trait of interest (the cases) and a group of unrelated individuals who do not exhibit the trait (the controls) [6]. Among the several different kinds of association studies, genome-wide association studies (GWAS) rely on a set of genetic markers covering the whole genome [6]. This strategy is motivated when there is little or no *a priori* information about the location of the genetic cause of the phenotype being studied. GWAS studies typically rely on two kinds of datasets, one collecting clinical (i.e. phenotypic) measurements, and the other one storing the individuals' genotypic markers values. Starting from these dataset, a statistical procedure is applied to identify which SNPs are more likely associated with a chosen phenotypic trait, usually in term of a p-value. As the biological interest is related to genes potentially involved in the trait manifestation, SNPs are usually annotated (i.e., they are mapped on the genome to find the genes they belongs to). The biological significance of the genes found to be statistically related to the trait is the final aim of the study. A first evaluation can be done exploiting the existing biological knowledge, for example making a Gene Ontology enrichment analysis, or identifying metabolic pathways containing those genes, or other sets of genes which are known to be related to the traits, etc. This search of new information is a way of deducting a biological evaluation for the statistical associations found.

Let's see how the Select and Test Model fits the tasks involved in a GWAS (Fig. 1b), as well as how we can execute each step of the process using the tools we implemented to address challenges (described in the relative references).

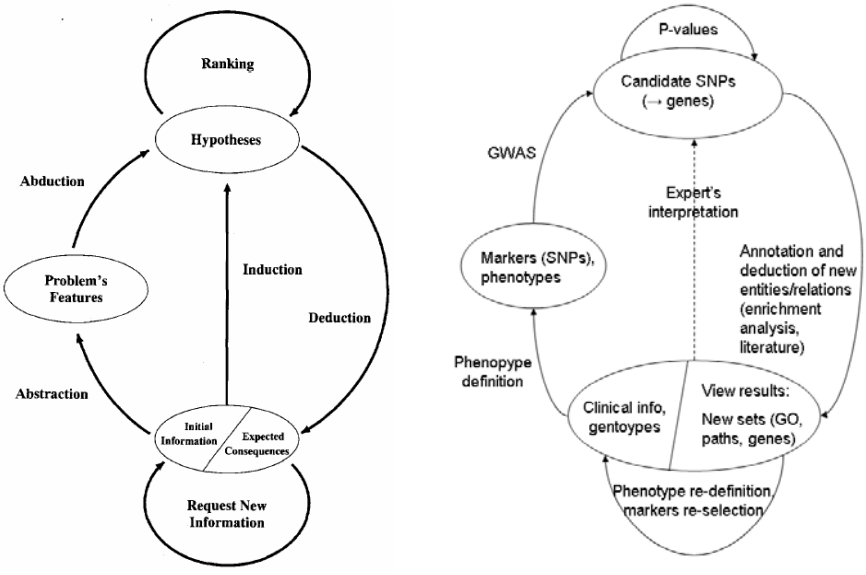
Abstraction. The first step is to select which of the clinical measurements in the initial dataset are useful for a proper phenotype definition. We developed the Phenotype Miner [1] to make a dynamic inspection of the data to identify the most suitable definition of the interesting phenotypes in the population under investigation. The result of this step is the identification of a set of individuals having the same phenotype.

Abduction. The association test is the hypotheses generator. Whatever tool a user chooses to compute the statistical association, the result is a set of candidate SNPs, so that the hypothesis to be tested is "SNP x is associated with the phenotype", for each SNP of the set. In our case, we used a freely available statistical package, pLink [7], to compute the association, so that the Phenotype Miner used in the previous step provides data formatted to run pLink commands.

Ranking. The candidate SNPs list is ordered according to the p-values which measure the statistical significance of the association between each marker and the phenotype.

Deduction. We used the functionality provided by Genephony [2] to manipulate the SNP list and make deduction. A candidate gene set is generated by annotating the SNPs set, then for example we can retrieve the metabolic pathways containing those genes, the Gene Ontology classes represented or find other genes ore phenotype that are already known to be associated with the phenotype under investigation.

At this point, all consequences derived with the deduction phased are proposed to the user (researcher, biologist, or more in general the domain expert), so that he/she can either make an eliminative induction to reduce the hypotheses space, or re-define the phenotype or reduce the initial marker set to filter the initial individual's set and start the analysis again.



a) b)

Fig. 1. a) A schematic representation of the epistemological model of hypothetical reasoning [5]. b) Model instance the case of Genome Wide Association Studies.

The possible scenarios, in fact, may be: i) a candidate gene is already known to be associated with the phenotype, so nothing new has been discovered, ii) there is no knowledge of association, so the hypothesis cannot be refused, iii) a candidate gene is also associated to another phenotype, so another analysis can be started defining and searching that phenotype in the initial clinical dataset, in order to corroborate the possible new finding.

4 Conclusion

The analysis of the very large volumes of heterogeneous data produced by modern genome-wide (GW) studies need the development of ad-hoc new methods and approaches, both at computational and conceptual level. Computational issues have been addressed by in the last few years with the development of several tools, which are mainly focused on providing an efficient way of handling, integrating, manipulating and making all data easy to be explored by researchers [2] [3] [4]. In this paper we focused on the conceptual issues, with the definition of an epidemiological model on which the design of next generation knowledge management tool could be based. We described the general validity of the model, giving a description of how it can be instantiated for GW association studies. The model can effectively catch the cycling tasks characterizing the scientific discovery processes involved in modern hypotheses-free biological research, so that other typical GW experiments can be modelled as well (e.g. transcription factor binding sites identification or knock-out gene experiments). A computational validation of a more comprehensive framework based on this model is now in progress, in order to build an automated reasoning systems aimed at supporting GW studies.

Acknowledgments. This work is a part of the project "Bioinformatics for Tissue Engineering: Creation of an International Research Group", funded by the "Fondazione Cariplo", and NIH grant R01 HL87681-01 “Genome-Wide Association Studies in Sickle Cell Anemia and in Centenarians”. RB was also supported by the FIRB project "ITALBIONET - Rete Italiana di Bioinformatica" funded by MIUR.

References

1. Nuzzo, A., Segagni, D., Milani, G., Rognoni, C., Bellazzi, R.: A Dynamic Query System for Supporting Phenotype Mining in Genetic Studies. Stud. Health Tech. Inf. (Medinfo 2007) 129(Pt 2), 1275–1279 (2007)
2. Nuzzo, A., Riva, A.: A Knowledge Management Tool for Translational Research. In: 1st AMIA Summit on Translational Bioinformatics, San Francisco, CA, p. 171 (2008)
3. Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W.J., Nekrutenko, A.: Galaxy: a platform for interactive large-scale genome analysis. Genome. Res. 15, 1451–1455 (2005)
4. Huang, W., Sherman, B.T., Tan, Q., Collins, J.R., Alvord, W.G., Roayaei, J., Stephens, R., Baseler, M.W., Lane, H.C., Lempicki, R.A.: The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 8(9) R183 (2007)
5. Ramoni, M., Stefanelli, M., Magnani, L., Barosi, G.: An Epistemological Framework for Medical Knowledge-Based Systems. IEEE Trans. on Systems, Man and Cybernetics 22(6), 1361–1375 (1992)
6. Balding, J.D.: A tutorial on statistical methods for population association studies. Nature Rev. Gen. 7(10), 781–791 (2002)
7. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C.: PLINK: a toolset for whole genome association and population-based linkage analysis. Am. J. Hum. Genet. 81, 559–575 (2007)

A Mutual Information Approach to Data Integration for Alzheimer's Disease Patients

Italo Zoppis[1,⋆], Erica Gianazza[2,⋆], Clizia Chinello[2], Veronica Mainini[2], Carmen Galbusera[2], Carlo Ferrarese[3,4], Gloria Galimberti[3], Alessandro Sorbi[5], Barbara Borroni[6], Fulvio Magni[2], and Giancarlo Mauri[1]

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano, Italy
[2] Department of Experimental Medicine, University of Milano-Bicocca, Monza, Italy
[3] Department of Neuroscience and Biomedical Technology, University of Milano-Bicocca, Monza (MI), Italy
[4] Department of Neurology, San Gerardo Hospital, Monza (MI), Italy
[5] Department of Neurological and Psychiatric Sciences, University of Florence, Florence, Italy
[6] Department of Neurology, University of Brescia, Center for Aging Brain and Dementia, Brescia, Italy

Abstract. Clinical data alignment plays a critical role in identifying important features for significant experiments. A central problem is data fusion i.e., how to correctly integrate data provided by different labs. This integration is done in order to increase ability of inferring target classes of controls and patients. Our paper proposes an approach based both on a information theoretic perspective, generally used in a feature construction problem [3] and on the approximated solution for a mathematical programming task (i.e. the weighted bipartite matching problem [6]). Numerical evaluations with two competitive approaches show the improved performance of the proposed method. For this evaluation we used data sets from plasma / ethylenediaminetetraacetic acid (EDTA) of controls and Alzheimer patients collected in three different hospitals.

Alzheimer disease (AD) represents one of the most common neurodegenerative disorder in the elderly. Alzheimer is often discovered late, so it is urgent to define biomarkers for an early detection, for a differential diagnosis from other neurodegenerative diseases and to monitor the course of the disease [7]. One of the emerging mass spectrometry (MS)-based screening method allowing high-throughput analysis of peripheral fluids with a simple and automated process is the ClinProt technique. A successful discovery of a proteomic profile related to an altered state has been obtained in different human diseases with this methodology (e.g. [1]). Since MALDI-TOF mass spectrometer resolution working in linear mode has a mass accuracy in the range of about +/- 8 Daltons, the measured m/z of the same

⋆ The present work has been supported by grants FIRB Italian Human ProteomeNet n. RBRN07BMCT of the Italian Ministry of Research (both these authors contributed equally to this study).

C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 431–435, 2009.

peptide can be different in each spectrum; consequently, it is necessary to align the mass spectra according to the sorted union of the m/z values or in a more general setting by the comparison of attributes (features) sharing common qualities (commonalities). The quality of sharing common attributes in data sources has been studied by methods that search for statistical dependencies between them. The earliest method was the classical linear Canonical Correlation Analysis [4] which has been extended to non linear variants (e.g. [5]) and more general methods that maximize mutual information [8]. In the same line of thought, our alignment is based on establishing peptide correspondences between data sets by concatenating intensity values (features) from different profiles, which are most informative for the respective target class. More specifically, the main source of inspiration for our proposal is the Information-Theoretic formalism used for the feature construction and extraction processes [3]. We describe formally this approach in section 1. Numerical evaluation and some comments are presented in section 2.

1 Materials and Methods

Samples were collected from three different hospitals using a standardized protocol. Plasma was obtained from blood collected in EDTA. Sample purification was performed with ClinProt MB-HIC8 (Magnetic Beads based Hydrophobic Interaction Chromatography) kit. Samples were then analysed by MALDI-TOF MS as previously described [1]. Mass spectra were acquired in positive linear mode in the m/z range of 1000-10000 Daltons with a resolution power of about 500-800. Accumulation of 420 laser shots resulted in a total averaging spectrum containing about 80-100 features.

1.1 Features and Model Construction

In order to solve such a kind of problems one generally applies a mutual information approach. One method uses two major components [3]. First a "relevance mechanism", which given a set of variables evaluates the relevance of the set; then a "construction mechanism" to properly define new variables. We formulate here below the problem by defining for each lab k the following objects:

1. Let $\mathcal{P}^{(k)}$ be the set of helpful peptides population for k. Each peptide $p_j \in \mathcal{P}^{(k)}$ has an associated random variable $I_{p_j}^{(k)}$, distributed as $f_{I_{p_j}^{(k)}}(i_{p_j}^{(k)})$ which describes the process of intensity value measurement[1].
2. Let $D^{(k)} : \Omega \to \{0, 1\}$ be a r.v. with distribution $f_{D^{(k)}}(d^{(k)})$ which gives the patient class membership for the specific disease. Ω represents the sample space for all patient population.
3. Finally let $M_{p_i}^{(k)} \sim f_{M_{p_i}^{(k)}}(m_{p_i}^{(k)})$ a r.v. which describes the mass weight values distribution for each peptide.

[1] We use the superscript to annotate the associated lab indexes.

The construction and relevance mechanisms are specifically formulated as follows:

Definition 1 (Construction mechanism)

Let $Z_{p_j}^{(k)} = \log_2 \frac{f_{I_{p_j}^{(k)},D^{(k)}}(I_{p_j}^{(k)},D^{(k)})}{f_{I_{p_j}^{(k)}}(I_{p_j}^{(k)}) f_{D^{(k)}}(D^{(k)})}$ *for each* k*. Then, for each pair* (k,s) *of labs and pairs of peptides* (p_j,p_t) *satisfying* $|M_{p_j}^{(k)} - M_{p_t}^{(s)}| \leq 8$ *we define a new feature* $Z_{p_j,p_t}^{(k,s)} = Z_{p_j}^{(k)} + Z_{p_t}^{(s)}$ *which gives the dependency of intensity values and target class events for different labs whenever the mass values are supposed to describe the same peptides entityes.*

Definition 2 (Relevance mechanism)
We consider the expectation

$$< Z_{p_i,p_j}^{(k,s)} > = < \log_2 \frac{f_{I_{p_j}^{(k)},D^{(k)}}(I_{p_j}^{(k)},D^{(k)})}{f_{I_{p_j}^{(k)}}(I_{p_j}^{(k)}) f_{D^{(k)}}(D^{(k)})} > + < \log_2 \frac{f_{I_{p_t}^{(s)},D^{(s)}}(I_{p_t}^{(s)},D^{(s)})}{f_{I_{p_t}^{(s)}}(I_{p_t}^{(s)}) f_{D^{(s)}}(D^{(s)})} > \tag{1}$$

that is, the sum of mutal information shared by $I_{p_j}^{(k)}$ *and* $D^{(k)}$ *in each lab i.e.,* $\mathcal{I}(I_{p_j}^{(k)},D^{(k)}) + \mathcal{I}(I_{p_t}^{(s)},D^{(s)})$*. Finally taking these peptides for which (1) has the highest values:*

$$(\tilde{p_j},\tilde{p_t}) = \underset{p_j \in \mathcal{P}^{(k)},\, p_t \in \mathcal{P}^{(s)}}{\operatorname{argmax}} \{\mathcal{I}(I_{p_j}^{(k)},D^{(k)}) + \mathcal{I}(I_{p_t}^{(s)},D^{(s)})\} \tag{2}$$

Each pair $(\tilde{p_j},\tilde{p_t})$ in (2) gives the sequence of the same peptides whose intensity values share most of the informations with the patient target classes[2].

1.2 Maximum Weight Bipartite Matching and Data Integration

In order to give an approximation for (2) we use an algorithmic solution for the Maximum Weight Bipartite Matching problem [6]. In this case, it implies reformulating the problem through bipartite graphs[3]. When it comes to considering weighted bipartite graphs (i.e. a function $w : E \to \Re$ exists), one can

[2] Since we are considering only two labs an estimation of (2) does not cause computational problems. Estimating mutual information in this case is straightforward because both the joint and marginal probability table can be obtained by discretizing and tallying, for each peptide and lab pairs the samples from $f_{I_{p_j}^{(k)},D^{(k)}}(i_{p_j}^{(k)},d^{(k)})$, $f_{I_{p_j}^{(k)}}(i_{p_j}^{(k)})$ and $f_{D^{(k)}}(d^{(k)})$ respectively.

[3] A graph $G = (V,E)$ is bipartite if there exists partition $V = A \cup B$ with $A \cup B = \varnothing$ and $E \subseteq A \times B$. A matching is a subset $\mathcal{M} \subseteq E$ so that $\forall v \in V$ at the most one edge in $\mathcal{M}$ is incident upon v. The size of a matching is $|\mathcal{M}|$, the number of edges in $\mathcal{M}$. A maximum matching $\mathcal{M}$ such that every other matching $\mathcal{M}'$ satisfies $|\mathcal{M}'| \leq |\mathcal{M}|$.

define the weights of a matching $\mathcal{M}$ as the sum of the weights of edges in $\mathcal{M}$: $s(\mathcal{M}) = \sum_{e \in \mathcal{M}} w(e)$. It is therefore possible to consider the following:

Problem: Given a bipartite weighted graph G, find a maximum weight matching.

For each (k, s) we first construct the samples $S^{(k)}_{p_j} = \{M^{(k)}_{1,p_j}, M^{(k)}_{2,p_j}, \ldots, M^{(k)}_{n,p_j}\}$ from $f_{M^{(k)}_{p_j}}(m^{(k)}_{p_j})$ and $S^{(s)}_{p_t} = \{M^{(s)}_{1,p_t}, M^{(s)}_{2,p_t}, \ldots, M^{(s)}_{n,p_t}\}$ from $f_{M^{(s)}_{p_t}}(m^{(s)}_{p_t})$. Then let the set $A = \bigcup_{p_j \in \mathcal{P}^{(k)}} S^{(k)}_{p_j}$ and $B = \bigcup_{p_t \in \mathcal{P}^{(s)}} S^{(s)}_{p_t}$. Therefore, we can think the problem has been structured through these two sets which characterize the vertex partition $V = A \cup B$ of $G(V, E)$. We proceed in this construction by adding the set of edges $E = \{(M^{(k)}_{\alpha,p}, M^{(s)}_{\beta,q}) : M^{(k)}_{\alpha,p} \in A, M^{(s)}_{\beta,q} \in B \wedge |M^{(k)}_{\alpha,p} - M^{(s)}_{\beta,q}| \leq \delta\}$ and for each edge, the weight:

$$w((M^{(k)}_{\alpha,p}, M^{(s)}_{\beta,q})) = \sum_{x \in \{k,s\}} \sum_{i^{(x)}_{p_i}, d^{(x)}} \tilde{f}_{I^{(x)}_{p_i}, D^{(x)}}(i^{(x)}_{p_i}, d^{(x)}) log \frac{\tilde{f}_{I^{(x)}_{p_i}, D^{(x)}}(i^{(x)}_{p_i}, d^{(x)})}{\tilde{f}_{I^{(x)}_{p_i}}(i^{(x)}_{p_i}) \tilde{f}_{D^{(x)}}(d^{(x)})}$$

where each $\tilde{f}$ is a histogram estimation for the respective distribution functions. This way we have a weighted bipartite graph. Hence, (2) can be obtained with one of the many general applied techniques [6]. The data combination process is performed by the concatenation of each pair of features linked by each edge in the matching.

2 Numerical Evaluations and Conclusions

A cohort of 6 control subjects and 9 AD patients was recruited from different clinics related to the University of Florence (Florence, Italy), 23 controls and 18 AD patients from San Gerardo Hospital (Monza, Italy) and a total of 6 controls and 15 AD patients from the Center for Aging Brain and Dementia (Brescia, Italy). Our intent is to evaluate the application of the Mutual Information based data fusion (labeled as MI based) by comparing the inference results with other two different methods. This evaluation has been obtained first of all by integrating the following combinations of data sets: Monza + Florence (MF data), Monza + Brescia (MB data), Florence + Brescia (FB data) and Monza + Florence + Brescia data (MFB data). Finally providing two competitive approaches, from now on called respectively Equal Mass Fusion, labeled as EM (the features from different labs have been unified whenever the associated mass values were equal) and T-Test based, labeled as TT (for all pair of features whose mass difference ranges in an interval of $+/-8$ units we compared the means from two different samples by a statistical t-Test. Then we unified these pairs of features with the maximum value of significance).

Comparisons have been computed on the basis of Area Under Roc curve (AUC), Recall and Precision [2]. In Table 1 we show the correspective average values of these performances. Clearly, the MI approach is complessively better

Table 1. Average performance scores

	AUC			Recall			Precision		
	EM	TT	MI	EM	TT	MI	EM	TT	MI
MF	0.4809	0.6079	0.6644	38.4600	57.7760	70.588	46.2980	58.0560	61.3760
MB	0.5846	0.5221	0.8272	36.5200	32.1740	65.2180	48.7760	52.7780	78.0020
FB	0.5556	0.7056	0.7222	87.1400	92.8560	95.7140	61.1520	76.5240	76.9620
MFB	0.4501	0.7056	0.5983	48.1260	92.8560	49.3760	45.3500	76.5240	57.3880

then the competitive methods, outperforming, on average, the other combined methods in 3 out of 4 cases. This happens for all the employed indexes. On the contrary, TTest based alignment seems to have a better performance for the MBF data, scoring the higher values in all the three performance indexes.

This study was realized on a small casistic, thus it is necessary to validate our results with a wider independent number of Alzheimer's patients and other techniques. In addiction, it is important to verify the diagnostic efficacy of these predictive models in a blinded manner on samples from subjects with different neurodegenerative pathologies.

References

1. Bosso, N., Chinello, C., Picozzi, S.C.M., Gianazza, E., Mainini, V., Galbusera, C., Raimondo, F., Perego, R., Casellato, S., Rocco, F., Ferrero, S., Bosari, S., Mocarelli, P., Kienle, M.G., Magni, F.: Human urine biomarkers of renal cell carcinoma evaluated by clinprot. Proteomics - Clinical Applications 2(7-8), 1036–1046 (2008)
2. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: ICML 2006, pp. 233–240. ACM, New York (2006)
3. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A. (eds.): Feature Extraction: Foundations and Applications. Springer, Heidelberg (2006)
4. Hotelling, H.: Relation between two sets of variates. Biometrika 28, 321–377 (1936)
5. Hsieh, W.W.: Nonlinear canonical correlation analysis by neural networks. Neural Networks 13, 1095–1105 (2000)
6. McHugh, J.A.: Algorithmic Graph Theory. Prentice-Hall, Englewood Cliffs (1990)
7. Ray, S., Britschgi, M., Herbert, C., Takeda-Uchimura, Y., Boxer, A., Blennow, K., Friedman, L.F., Galasko, D.R., Jutel, M., Karydas, A., Kaye, J.A., Leszek, J., Miller, B.L., Minthon, L., Quinn, J.F., Rabinovici, G.D., Robinson, W.H., Sabbagh, M.N., So, Y.T., Sparks, D.L., Tabaton, M., Tinklenberg, J., Yesavage, J.A., Tibshirani, R., Wyss-Coray, T.: Classification and prediction of clinical alzheimer's diagnosis based on plasma signaling proteins. Nat. Med. 13(11), 1359–1362 (2007)
8. Viola, P.A.: Alignment by maximization of mutual information. Int. Jour. of Computer Vision, 16–23 (1995)

Author Index