Riccardo Bellazzi
Ameen Abu-Hanna
Jim Hunter (Eds.)

LNAI 4594

Artificial Intelligence in Medicine

11th Conference on Artificial Intelligence in Medicine, AIME 2007
Amsterdam, The Netherlands, July 2007, Proceedings

Springer

Lecture Notes in Artificial Intelligence 4594

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Riccardo Bellazzi Ameen Abu-Hanna
Jim Hunter (Eds.)

Artificial Intelligence in Medicine

11th Conference on Artificial Intelligence
in Medicine, AIME 2007
Amsterdam, The Netherlands, July 7-11, 2007
Proceedings

Springer

Volume Editors

Riccardo Bellazzi
University of Pavia
Dipartimento di Informatica e Sistemistica
via Ferrata 1, 27100 Pavia,Italy
E-mail: riccardo.bellazzi@unipv.it

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

Jim Hunter
University of Aberdeen
King's College, Department of Computing Science
273422 Aberdeen AB24 3UE, UK
E-mail: jhunter@csd.abdn.ac.uk

Library of Congress Control Number: Applied for

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 3-540-73598-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-73598-4 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12089543 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), and Aberdeen, UK (2005).

This volume contains the proceedings of AIME 2007, the 11th Conference on Artificial Intelligence in Medicine, held in Amsterdam, The Netherlands, July 7-11, 2007. The AIME 2007 goals were to present and consolidate the international state of the art of AI in biomedical research from the perspectives of methodology and application. The conference included invited lectures, a panel discussion, full and short papers, tutorials, workshops, and a doctoral consortium. In the conference announcement, authors were solicited to submit original contributions on the development of theory, systems, and applications of AI in medicine, including the exploitation of AI approaches to molecular medicine and biomedical informatics. Authors of papers addressing theory were requested to describe the development or the extension of AI methods and to discuss the novelty 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, application papers were required to describe the implementation of AI systems in solving significant medical problems, and to present sufficient information to allow an evaluation of the practical benefits of such systems.

AIME 2007 received the second highest number of submissions ever (137). Submissions came from 31 different countries, including 12 outside Europe. All papers were carefully peer-reviewed by at least two experts from the Program Committee with the support of additional reviewers. The reviewers judged the quality and originality of the submitted papers, together with their relevance to the AIME conference. Four criteria were taken into consideration in judging submissions: the overall reviewers' recommendation, the suitability of the paper for an oral or poster presentation, the reviewers' detailed comments and the reviewers' confidence about the subject area. In a meeting held in Amsterdam during March, 24–25 a small committee consisting of the AIME 2007 Organizing Committee Chair, Ameen Abu-Hanna, the AIME President, Jim Hunter, and the AIME 2007 Program Chair, Riccardo Bellazzi, took the final decisions on the

AIME 2007 scientific program. As a result, 28 long papers (with an acceptance rate of about 20%) and 38 short papers were accepted. Each long paper was presented as an oral presentation, and was allocated a time of 25 minutes during the conference. Each short paper was presented as a poster. One of the novelties of AIME 2007 was a separate evening session combining poster presentations and dinner. Each poster was discussed with the help of two Poster Session Chairs.

The papers were organized according to their topics in eight main themes: 1) Agent-based systems; 2) Temporal data mining; 3) Machine learning and knowledge discovery; 4) Text mining, natural language processing and generation; 5) Ontologies; 6) Decision support systems; 7) Applications of AI-based image processing techniques; 8) Protocols and guidelines and 9) Workflow systems.

As another novelty, AIME 2007 had the privilege of hosting a panel discussion on "The Coming of Age of AI in Medicine," organized and moderated by Vimla Patel (Arizona State University, USA). The distinguished panellists were Edward Shortliffe (University of Arizona College of Medicine, USA), Mario Stefanelli (University of Pavia, Italy), Peter Szolovits (Massachusetts Institute of Technology, Cambridge, MA, USA) and Michael Berthold (University of Konstanz, Konstanz, Germany). Peter Szolovits and Michael Berthold were also the invited speakers at AIME 2007. Peter Szolovits gave a talk on "Rationalism and Empiricism in Medical AI" and Michael Berthold on "The Fog of Data: Data Exploration in the Life Sciences." The choice of these topics was partly related to the recent broadening of the field of AI in medicine and biomedical informatics, which now spans themes from clinical decision support to supporting research in genomics, proteomics and computational biology as applied to medicine and health care.

Following its first appearance at AIME 2005, a doctoral consortium, organized on this occasion by Jim Hunter, was held again this year. A scientific panel consisting of Carlo Combi, Michel Dojat, Frank van Harmelen, Elpida Keravnou, Peter Lucas, Silvia Miksch, Silvana Quaglini, and Yuval Shahar supported the selection of PhD student contributions and discussed the contents of the students' doctoral theses.

As at previous AIME meetings, two full-day workshops were organized prior to the AIME 2007 conference: a workshop entitled "From Knowledge to Global Care," organized by David Riaño, Rovira i Virgili University, Spain and Fabio Campana, CAD RMB, Rome, Italy, and the 12th workshop on Intelligent Data Analysis in bioMedicine and Pharmacology (IDAMAP 2007), organized by Allan Tucker, Brunel University, UK, and Carlo Combi, University of Verona, Italy, and sponsored by the IMIA working group on Intelligent Data Analysis and Data Mining.

Two half-day tutorials were given by John H. Holmes, University of Pennsylvania, USA: Introduction to Applied Clinical Data Mining and Advanced Applied Clinical Data Mining.

We would like to thank everyone who contributed to AIME 2007. 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, panellists, and the organizers of the workshops, tutorials and doctoral consortium. We would like to thank the Department of Medical Informatics at the Academic Medical Centre of the University of Amsterdam, which hosted and sponsored the conference. Finally, we thank the Netherlands Organization for Scientific Research (NWO), Medecs Decision Support Systems, and the BAZIS foundation for their sponsorship of the conference and the European Coordinating Committee for Artificial Intelligence (ECCAI) and the International Medical Informatics Association (IMIA) for their support.

May 2007

Riccardo Bellazzi
Ameen Abu-Hanna
Jim Hunter

Organization

Program Committee

Ameen Abu-Hanna, The Netherlands
Klaus-Peter Adlassnig, Austria
Steen Andreassen, Denmark
Pedro Barahona, Portugal
Robert Baud, Switzerland
Riccardo Bellzzi, Italy
Petr Berka, Czech Republic
Isabelle Bichindaritz, USA
Paul de Clercq, The Netherlands
Enrico Coiera, Australia
Carlo Combi, Italy
Michel Dojat, France
Henrik Eriksson, Sweden
John Fox, UK
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
Pedro Larranaga, Spain
Nada Lavrac, Slovenia
Johan van der Lei, The Netherlands
Xiaohui Liu, UK
Peter Lucas, The Netherlands
Silvia Miksch, Austria
Stefania Montani, Italy
Mark Musen, USA
Niels Peek, The Netherlands
Mor Peleg, Israel
Christian Popow, Austria
Silvana Quaglini, Italy
Alan Rector, UK
Stephen Rees, Denmark
Rainer Schmidt, Germany
Brigitte Seroussi, France
Yuval Shahar, Israel
Basilio Sierra, Spain
Costas Spyropoulos, Greece
Mario Stefanelli, Italy
Paolo Terenziani, Italy
Samson Tu, USA
Frans Voorbraak, The Netherlands
Dongwen Wang, USA
Thomas Wetter, Germany
Blaz Zupan, Slovenia
Pierre Zweigenbaum, France

Organizing Committee

Ameen Abu-Hanna
Gita Guggenheim
Ellen Lustenhouwer-Werring
Niels Peek
Winston Gary Tjon Sjoe Sjoe
Riccardo Bellazzi
Marco Tornielli
Giorgio Leonardi
Lucia Sacchi

Additional Reviewers

Alessio Bottrighi
Alexander Artikis
Alime Öztürk
Anand Kumar
Arjen Hommersom
Barbara Di Camillo
Barbara Oliboni
Elias Zavitsanos
Florence Forbes
Francesca Demichelis
Francois Portet
Fulvia Ferrazzi
Georgios Petasis
Giorgio Leonardi
Hameedullah Kazi
Igor Trajkovski
Jacques Bouaud
Janez Brank
Konstantinos Stamatakis
Laura Giordano
Lina Rojas
Linda Peelen
Luca Anselma
Lucia Sacchi
Marcel van Gerven
Marion Verduijn
Maurice Prijs
Nicola Barbarini
Paola Cerchiello
Paolo Magni
Patrick Martini
Perry Groot
Petra Kralj
Phattanapon Rhienmora
Rattapoom Waranusast
Rick Goud
Sergios Petridis
Silvia Panzarasa
Stefan Visscher
Theodore Dalamagas

Panel

The Coming of Age of AI in Medicine

Organizer: Vimla L. Patel, Arizona State University, USA

Workshops

From Knowledge to Global Health Care

Co-chairs: David Riaño, Rovira i Virgili University, Spain, Fabio Campana, CAD RMB, Italy

IDAMAP 2007: Intelligent Data Analysis in bioMedicine and Pharmacology

Co-chairs: Carlo Combi, University of Verona, Italy, Allan Tucker, Brunel University, UK

Tutorial

Introduction to Applied Clinical Data Mining

John H. Holmes, University of Pennsylvania, USA

Advanced Applied Clinical Data Mining

John H. Holmes, University of Pennsylvania, USA

Table of Contents

Agent-Based Systems

Temporal Data Mining

Machine Learning and Knowledge Discovery

Text Mining, Natural Language Processing and Generation

Ontologies

Decision Support Systems

Applications of AI-Based Image Processing Techninques

Protocols and Guidelines

Workflow Systems

Part I

Agent-Based Systems

A Human-Machine Cooperative Approach for Time Series Data Interpretation

Thomas Guyet[1], Catherine Garbay[2], and Michel Dojat[3]

[1] TIMC Laboratory, Domaine de la Merci F-38706 La Tronche, France
Thomas.Guyet@imag.fr
[2] CNRS/LIG Laboratoire d'Informatique de Grenoble, France
[3] UMR-S 836 Inserm/UJF/CEA "Institut des Neurosciences", Grenoble, France

Abstract. This paper deals with the interpretation of biomedical multivariate time series for extracting typical scenarios. This task is known to be difficult, due to the temporal nature of the data at hand, and to the context-sensitive aspect of data interpretation, which hamper the formulation of *a priori* knowledge about the kind of patterns to detect and their interrelations. A new way to tackle this problem is proposed, based on a collaborative approach between a human and a machine by means of specific annotations. Two grounding principles, namely autonomy and knowledge discovery, support the co-construction of successive abstraction levels for data interpretation. A multi-agent system is proposed to implement effectively these two principles. Respiratory time series data (Flow, Paw) have been explored with our system for patient/ventilator asynchronies characterization studies.

1 Introduction

In the mass of monitoring data, the detection of specific patterns relative to a significant event requires a time-consuming visual data examination by an expert. The relations between events occurring in various physiological parameters are meaningful but difficult to identify. In addition, the potential generic character of the detected patterns is rarely pointed out. We assume that without the use of a computerized assistant, this mass of data can not be fully exploited. However, the design of such an assistant is difficult because the temporal and multivariate nature of the data at hand and the context-sensitive aspect of data interpretation hamper the formulation of *a priori* knowledge about the kind of patterns to detect and their interrelations. The number of patterns and their combination to form meaningful situations is potentially very high, making their systematic enumeration and description by a human impossible.

For instance, in the context of patients hospitalized in Intensive Care Unit (ICU) and mechanically ventilated, asynchrony, *i.e.* a frequent mismatch between the ventilator assistance and the patient demand, has recently been reported [1]. Automatic asynchrony detection could be used to adjust ventilator settings and then improve mechanical assistance efficiency. However, to model asynchrony patterns for a robust automatic detection, essential information is still missing.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 3–12, 2007.

In this paper, we propose a collaborative human-machine approach to gradually define from the data the significant patterns to detect, and to discover specific temporal relations between events.

Following [2], we advocate the use of annotations as an elegant and efficient way for communicating between a human and a machine. Starting from expert annotations, the computer builds its own abstractions and in turn annotates the signals. The human-machine interactions occur at three abstraction levels (*e.g.* segmented time series, symbolic time series data and scenarios), around which a collaborative approach to the interpretation of time series data is proposed. We implemented our approach using the multi-agent (MA) paradigm. The MA system (MAS) design is based on two grounding principles, namely autonomy and knowledge discovery, which support the construction by the system of data interpretation at successive abstraction levels.

2 State of the Art

Knowledge-Based Temporal Abstraction. Basically, abstraction transforms initial numerical time series to symbolic time series, which is known to result in a so-called semantic gap. The larger is the gap the more difficult is the abstraction. To fill this gap, as showed by Shahar [3], the abstraction process relies on several mechanisms and uses various types of *a priori* knowledge. There is a wide literature regarding the way to realize several temporal abstraction levels in medical domains [4].

For instance, in the Calicot system [5], dedicated to cardiac arrhythmias detection, *a priori* knowledge is introduced to support the discovering of new rules. Neural networks are trained to discover normal or abnormal P-waves and QRS-complexes and translate physiological time series into symbolic time series. Then, the system learns rules with an Inductive Logic Programming method where *a priori* knowledge and bias can be introduced. The computed rules show complex temporal relations between the symbols attached to the various arrhythmias, which are difficult to understand by an expert. However, ICU data appears unfortunately insufficiently formalized to envisage such a procedure.

Guimarães et al. [6] propose the TCon method for automated temporal knowledge acquisition applied to sleep-related respiratory disorders. TCon uses machine learning techniques to discover new elementary patterns (rules) from the time series. These rules are then translated, in order to be easily interpretable for the clinician, a central motivation that we share. But we are then facing a trade-off: too much *a priori* knowledge hampers knowledge discovery, whilst not enough *a priori* knowledge produces non-coherent or difficult to interpret abstractions. Step by step validation of the abstraction process by the human expert, based on well-adapted visual representations, has been shown as a way to build robust abstraction process [7]. In the same vein, we propose that during the knowledge discovery process, the expert continuously interacts with the learning system to drive the process and extract useful chunks of knowledge.

Collaborative Knowledge Discovery. The development of mutual learning capabilities is a way to solve the previously mentioned trade-off between autonomy and intelligibility. In [8], the authors present the desired cycle of knowledge where clinicians are central actors in knowledge production with a positive impact on the final system. Classical approaches to collaborative knowledge discovery consist either in human-driven machine learning, the clinician acting as an expert, or in computer assisted human learning, the machine acting as a teacher (Intelligent Tutoring Systems). TeachMed [9] for instance is a recent system to support clinical learning. The student is observed and when, for TeachMed, he/she needs assistance a dialog is engaged to help him/her to find the solution. Computers can also serve as a medium for a collaborative mutual learning (Computer-Supported Collaborative Learning) in facilitating interaction between humans [10]. In [11] human and machine are both embedded in a learning cycle. On this basis, they automatically generate in a faster way more reliable guidelines compared to classical methods.

Multi-agent systems exhibit several properties that make them a “natural” framework for the design of such collaborative systems, particularly due to the open character of the MA architecture and to the autonomy of its components. They facilitate the combination of several approaches for distributed problem solving, support the autonomous emergence of new knowledge and may naturally integrate a human user as a specific agent. According to this view, the MASKS environment [12] introduces agent interactions based on cooperative behaviors to facilitate knowledge discovery. Agents, that encapsulated machine learning algorithms, respectively evaluate and may merge their generated rules. In [13], a MAS is used to collaboratively annotate artistic artifacts. The system opens a share space to enable experts to insert their own annotations and learn mutually from each other.

Interactive knowledge construction literature indicates that human/machine interaction enables knowledge discovery both for the machine and the human. Whilst the human helps to focus the machine learning process to the discovery of useful chunk of knowledge, the machine helps to focus the human learning process to a deep exploration of the data at hand.

3 The Methodological Concepts

Signal Interpretation. We consider interpretation as a complex process involving complementary exploration, description and annotation tasks, which are operated by a collection of autonomous agents (including humans), working to progressively refine the interpretation [13]. The role of the exploration task is to focus the attention, at each abstraction level, on relevant information, in terms of time series segments, event classes or typical event relations. In our system, several steps are necessary before reaching the upper notion of scenario: 1) Segmentation, 2) Symbolic time series and 3) Scenario construction. The description and annotation steps are situated in a dynamic context which includes the agent past experience and additional information on data. The description task aims to build numerical and symbolical models of the information under interest, possibly by fusion. Finally, the annotation task role is to attach symbolic labels to newly provided raw data using the constructed models.

Human-Machine Collaboration. We propose to approach the interpretation of time series data as performed by two cooperating agents - a man and a machine - operating across three successive abstraction levels (segments, symbolic time series and scenarios). These agents share a common environment (*e.g.* annotations, time series data, segments ...) and mutually interact via annotations to progressively refine their own interpretation. In the absence of consistent *a priori* knowledge and considering the difficulty of this processing, an active partnership between man and machine is sought. It does not come down to a fixed request-answer interaction scheme in which each actor is meant to compensate for the lack of knowledge of its partner, and therefore supposed to share its world of meaning. This partnership is rather meant to allow a co-construction of annotations, in which the interpretation of facts is not defined beforehand by one of the actors, but co-constructed in the course of their interaction. According to this principle, each partner is in turn given the possibility to observe and interpret annotations provided by its partner, and/or to propose annotations judged as appropriate according to a given interpretation context. In such framework, each partner is meant to "play" in its own world of meaning, thus preserving its autonomy, with no prevalence of one on the other. In consequence, there is a possibility of learning and discovery for both partners. To be noticed is the fact that any information - be it a case example or a conceptual interpretation - is provided within its context of appearance, so that interaction develops in a situated way.

4 A Multi-Agent System for Time Series Interpretation

Central to our design, is the distinction between three main tasks, namely segmentation (localization of relevant segments), classification (symbolic labeling of the segments), and scenario construction (computation of inter-symbol relationships). Dedicated agents are conceived to support these various tasks; they make use of specific models, which drive the way to explore and describe, via annotations, the information at hand (forward application). These models are autonomously computed (backward application), based on knowledge discovery methods, and according to previously computed annotations, which may be provided by an external user, or by other agents. Interaction and feedback loops occur all along the interpretation process to ensure its progressive refinement.

To be noticed is the fact that all these processes, as the human agent himself, run in parallel and may interact at any time, thanks to the annotation process which results in modification propagated within the information at hand, and thus altering but not disrupting the interpretation process.

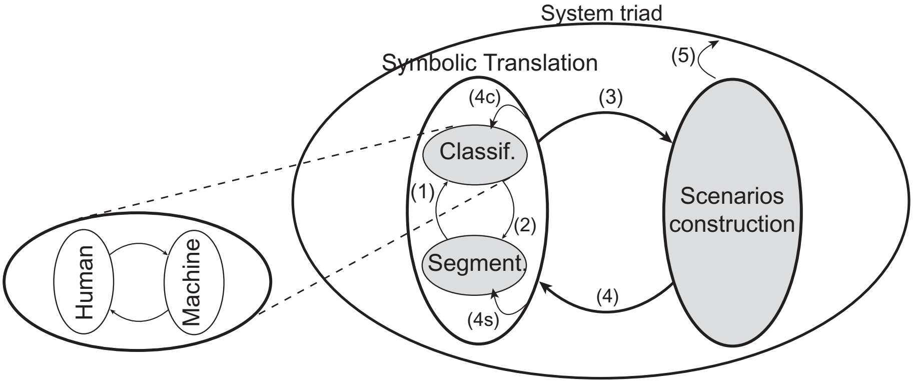


Fig. 1. MAS entities organization according to recursive triads. The classification triad is expanded to show the man-machine interaction.

4.1 System Design

A global view of the system architecture is presented in Figure 1. A triadic organization is proposed as a conceptual framework to support the needed inter-agent interactions and feedback loops. A triad is made of two processing entities working in reciprocal interaction that constitutes a more abstract processing whole. Each abstraction level is organized into a triad, giving rise to new annotation elements to be processed at a higher abstraction level (upper triad). Conversely, each upper level triad may launch lower-level triadic agents (feedback loop). Time series data are processed within the segmentation and classification triads, which mutually interact (links 1 and 2, Figure 1) within the symbolic translation triad to build symbolic time series. Symbolic time series are processed within the scenario construction triad. The symbolic translation and scenario construction triads mutually interact within the system triad (links 3 and 4). Scenarios are the final outcome to the system triad (link 5). When entering a feedback loop (link 4, 4s and 4c), the constructed scenarios may conversely be used to propagate modifications to the symbolic time series, and from this triad to the lower level classification and segmentation agents.

4.2 Agentification

Classification Agents. For the sake of simplification, we first of all describe the system style of working in front of univariate time series. The general control cycle of a classification agent is (i) to collect segments (ii) to compute segment descriptors (iii) to classify them and (iv) to build segment class models. The segments to analyze may be pointed out by a human, or delimited by the segmentation agents. At system start, some initial annotations or models are provided to some agent, to launch the whole interpretation process.

Each segment is described by a feature vector (typically 20 sampled values and the length of the segment), which are classified. A weighted Euclidean distance is used to compare them. The k-mean algorithm has been selected because of its capacity to work under our specific constraints: lack of *a priori* knowledge (in particular no *a priori* knowledge of the number of classes to observe), and dynamicity (evolution of the class characterization along the interpretation in progress). We use an extended version of the algorithm, in which different distances (*i.e.* adaptive weights in the Euclidean computation) are selected and computed, depending on the class under interest. For each class, the mean vector and the distances weights constitute the model of the segments of this class, attached to a symbolic name with no proper semantic.

The next issue to consider is then how to match this symbolic name with an intelligible denomination for the clinician (*e.g.* asynchrony events). This is performed by means of triadic man-machine interaction, according to which any modification (*e.g.* annotation) performed by any agent is merely a perturbation of the information under interpretation. According to this view, the system interpretation may be shifted toward human interpretation in a transparent way.

Segmentation Agents. For each class of segments constructed by the classification agent, a new segmentation agent is created, and provided with the corresponding segment model. Consequently, the number of segmentation agents changes in the course of the interpretation. Each segmentation agent performs a pattern matching algorithm : It calculates the distance between the agent's segment model and a sliding window of variable length, that delineates a segment, moving on the time series. Then, the segments closest to the model are selected. A non-overlapping constraint is used to eliminate further segment hypotheses. Distances are computed in the model parameter space (the previously described vectorial description of a segment), rather than in the time series space, to cope with differences of pattern size. The segmentation of a time series is the concatenation of the segments constructed by all these agents.

Here again, annotations are exchanged between the machine and the human via the interface. Segments are shown to the human through the system visualization interface. Such visualization allows focusing the human attention in two ways: whilst annotated signal elements point out "recognizable" events, and submit them to further examination, non annotated events conversely either reveal a lack of knowledge, or suggest further exploration of the data.

To be noticed is the fact that any result, newly obtained at the segmentation level, is transmitted forward to the classification agent, which may in turn drive a further launching of segmentation agents. This process continues until some stabilization is being reached, which is controlled by human interventions.

Symbolic Translation Triad. The role of the symbolic translation triadic agent is to translate numerical time series into symbolic time series. It collects to this end the segments and models of segments computed by the segmentation and classification agents. For each segment (S), the triadic agent creates a time-stamped symbol with the date and duration of S and the symbolic name (N) of the most accurate models of segments. Concatenated time-stamped symbols for a given time series constitute the symbolic time series of a patient record. The whole process results in concatenated time-stamped symbols which characterize the patient record at the symbolic level.

For multivariate time series, one classification agent is dedicated to each *type* of time series (Flow, Paw, ...), and defines the corresponding segmentation agents. The symbolic time series merges time-stamped symbols from all types of time series.

Scenario Agents. The scenario construction is considered as the process of finding temporal relations between symbols. This process is driven by the assumption of some symbols, selected by the clinician, that are to be "explained" through their causal links to some ancestor "events" (symbols). The task of the scenario agent, associated to a symbol E, is to collect the symbols preceding each occurrence of E in a temporal window with a fixed length. Each set of such symbols is then considered as an example for the learning process at hand. From these examples, the agent finally constructs the time-stamped pattern, *i.e.* the scenario that "explains" E. To this end, we use an extended version of the `A Priori` algorithm [14], which builds the biggest frequent time-stamped pattern. The resulting scenario may of course be submitted to the user, which may result in modifications in the proposed pattern, or collecting other example patterns to support or contradict the current interpretation. In a complementary feedback loop (link 4 in Figure 1), the proposed scenario may be considered as a model to drive further analysis. Any deviation from the current model is considered as a potential for improvement of the annotation process. The system will then focus its attention on the corresponding deviation locations.

5 Patient-Ventilator Asynchronies Exploration Results

5.1 Rationale

Patients suffering from respiratory disorders and hospitalized in Intensive Care Units are mechanically supported with assist-control ventilation. An adequate synchronization between patient and ventilator is then required to improve patient's comfort and optimize work of breathing. A recent study [1] shows that 24% of 60 patients mechanically ventilated exhibited an asynchrony index higher than 10% of respiratory efforts. Ineffective triggering, *i.e.* when patient's efforts do not result in ventilator triggering, was the major cause (85%) of asynchrony. Asynchrony was associated with a longer duration of mechanical ventilation. Therefore, it is important to identify factors increasing the incidence of asynchrony. Some major asynchronies can be detected by an experienced clinician. This remains a difficult extra workload and no physiological models exist yet to allow for their automatic detection. In order to detect, from the physiological data, specific patterns, regularities or sequences of events (scenarios), associated to the occurrence of asynchrony, we provided a clinician with our system to annotate and explore respiratory time series. They were constituted of flow and airway pressure (Paw) signals continuously recorded during 30 minutes and sampled at 200 Hz (see [1] for details).

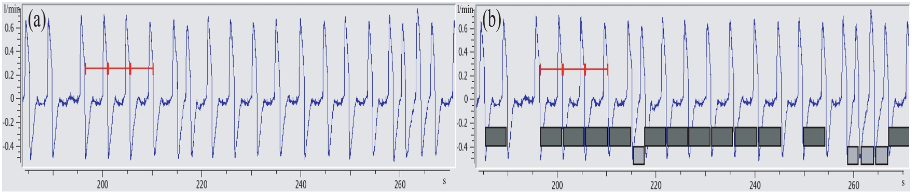


Fig. 2. Model construction (Flow signals). 2a (left): Two annotations (gray horizontal bar) are inserted on the flow signal by the clinician to indicate asynchrony periods. 2b (right): Annotation completion by the system: dark-gray boxes indicate retrieved asynchronies periods. Medium-gray boxes indicate retrieved non-asynchronies periods.

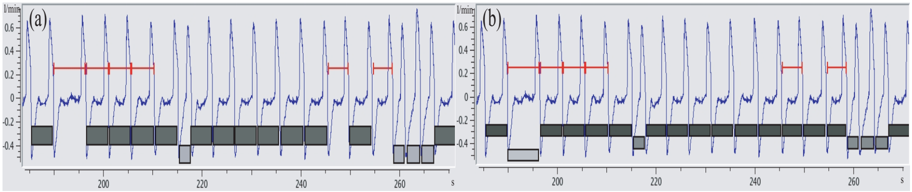


Fig. 3. Model discovery (Flow signals). 3a (left): The clinician has annotated the unrecognized segments. 3b (right): A new model for double triggering periods has been discovered (single light-gray box), the "ineffective triggering" model has been refined to retrieve all ineffective triggering periods.

5.2 Time Series Data Exploration Case Study

We first of all illustrate the capacity of our system to build annotations in an autonomous way, thereby assisting the clinician in the elaboration of new interpretation models; we illustrate in a second step its capacity to validate previously constructed models, assisting the clinician in the model-driven investigation of unknown data. Alternating successive steps of model construction/model validation would finally allow the clinician to build a progressively refined interpretation of data, together with more robust models.

Models Construction and Automatic Asynchrony Detection. Three patient's records were fully interpreted via our system. Using the system interface, the clinician by a visual inspection, annotated some inefficient triggering (see Figure 2a) on a specific part of the signal. Then, the clinician launched the processing step *i.e.* the execution of the classification agents, the segmentation agents and finally the symbolic translation triadic agent. Based on the initial partial annotation by the expert, the system symbolically labeled the complete time series (see Figure 2b). On the Figure 2, all asynchronies considered as similar to those annotated by the clinician as ineffective triggering, were retrieved (dark-gray boxes). The three complete time series were automatically annotated based on clinician's annotations that represented about 30% of the total asynchronie events. The mean sensitivity and the mean specificity were respectively egal to 0.64 and 0.93. Finally, the system constructed scenarios that bring new information about asynchrony occurrence conditions. For this purpose, Paw and Flow signals were automatically explored in a temporal windows (fixed to $10s$ *i.e.* approximatively three periods) around asynchrony events. Our system detected that asynchrony on the Flow signal was associated in 90% of cases with asynchrony on the Paw signal and that they were preceeded by another asynchrony in 90% of cases. This extracted knowledge was then used to automatically detect deviant patterns, such as a single asynchrony (may be artefactual) or mismatch between flow and paw asynchronies. New annotations (artefacts, trends) should be inserted by the clinician to finalize the validation and enrich these first results.

Models Comparison and Extraction of New Patterns. Our system may also support the clinician in the model-driven investigation of unknown data, to validate the genericity of a previously constructed model. To illustrate this point, we used seven new data sets. For three of these data sets, the "ineffective triggering" model that was previously built from one patient data set, appeared clearly not relevant. However, for the four remaining data sets, the model revealed adapted and the clinician was able to identify new patterns: two ineffective triggering in the same period pattern (see Figure 3) and double triggering patterns. Extensive data exploration should be realized to confirm these preliminary results.

6 Discussion and Perspectives

We have presented an original approach to support clinicians in the difficult task of multivariate time series data interpretation. Our approach is centered on the collaboration between a clinician and an autonomous system, both embedded in a learning cycle. We advocate for the design of computerized tools that really support the clinician in his/her decision making process, rather than provide him/her with final results. The presented system participates to this medical computerized tools design evolution [8]. We are aware of the computing complexity of the processes we have presented. But we assume that data interpretation is a prospective task with no critical time constraints.

The undergoing experiments on patient ventilator asynchrony exploration strongly support the interest of our approach. We expect that the global system implementation in progress, including the feedback from scenario construction, will improve the detection of known events and consequently its sensitivity. Several extensions of the present implementation may be envisaged. The fusion of several physiological parameters would enable the construction of more elaborate asynchrony models, as well as taking into account contextual information such as pathology, treatments or the mode of ventilation used. The objective is to integrate contextual information on patient as a interpretation element and maybe enable clinician to construct typical scenarios for groups of patients. Finally, the preliminary experiments on scenario learning indicate that a more sophisticated time representation should be introduced (such as interval relations for example). Further experimentations in close collaboration with clinicians will allow us to fully evaluate the real impact of such as interactive approach for knowledge discovery and knowledge formalization.

Acknowledgments. We thank Dr. Thille and Pr. Brochard from H. Mondor Hospital (Créteil, Fr) for providing the clinical data and for comments.

References

1. Thille, A., Rodriguez, P., Cabello, B., Lellouche, F., Brochard, L.: Patient-Ventilator Asynchrony During Assisted Mechanical Ventilation. Intens. Care Med. 32(10), 1515–1522 (2006)
2. Salatian, A., Hunter, J.: Deriving Trends in Historical and Real-Time Continuously Sampled Medical Data. J. Intell. Inf. Syst. 13(1-2), 47–71 (1999)
3. Shahar, Y.: A Framework for Knowledge-Based Temporal Abstraction. Artif. Intell. 90(1-2), 79–133 (1997)
4. Augusto, J.C.: Temporal Reasonning for Decision Support in Medicine. Artif. Intell. Med. 33(2), 1–24 (2005)
5. Fromont, É., Quiniou, R., Cordier, M.O.: Learning Rules from Multisource Data for Cardiac Monitoring. In: Miksch, S., Hunter, J., Keravnou, E.T. (eds.) AIME 2005. LNCS (LNAI), vol. 3581, pp. 484–493. Springer, Heidelberg (2005)
6. Guimarães, G., Peter, J.H., Penzel, T., Ultsch, A.: A Method for Automated Temporal Knowledge Acquisition Applied to Sleep-Related Breathing Disorders. Artif. Intell. Med. 23(3), 211–237 (2001)
7. Silvent, A.S., Dojat, M., Garbay, C.: Multi-Level Temporal Abstraction for Medical Scenarios Construction. Int. J. Adapt. Control. 19(5), 377–394 (2005)
8. Zupan, B., Holmes, J., Bellazzi, R.: Knowledge Based Data Analysis and Interpretation. Artif. Intell. Med. 37(1), 163–165 (2006)
9. Kabanza, F., Bisson, G., Charneau, A., Jang, T.S.: Implementing Tutoring Strategies Into a Patient Simulator for Clinical Reasoning Learning. Artif. Intell. Med. 38(1), 79–96 (2006)
10. Lee, E., Chan, C., Aalst, J.: Students Assessing their Own Collaborative Knowledge Building. Int. J. of Computer-Supported Collaborative Learning 1(1), 57–87 (2006)
11. Morik, K., Imhoff, M., Brockhausen, P., Joachims, T., Gather, U.: Knowledge Discovery and Knowledge Validation in Intensive Care. Artif. Intell. Med. 19(3), 225–249 (2000)
12. Schroeder, L., Bazzan, A.: A Multi-Agent System to Facilitate Knowledge Discovery: An Application to Bioinformatics. In: Proceedings of the Workshop on Bioinformatics and Multiagent Systems (2002)
13. Bottoni, P., Garbay, C., Lecca, F., Mussio, P., Rizzo, P.: Collaborative Indexing and Retrieval by Annotation: the Case of Artistic Artifacts. In: Proceedings of the 2nd International Workshop on Content-based Multimedia Indexing, pp. 315–322 (2001)
14. Dousson, C., Duong, T.: Discovering Chronicles with Numerical Time Constraints from Alarm Logs for Monitoring Dynamic Systems. In: Dean, T. (ed.) Proceedings of the 16th International Joint Conference on Artificial Intelligence, pp. 620–626. Morgan Kaufmann, San Francisco (1999)

MRF Agent Based Segmentation: Application to MRI Brain Scans

B. Scherrer[1,2], M. Dojat[1], F. Forbes[3], and C. Garbay[2]

[1] INSERM U836-UJF-CEA-CHU (Grenoble Institute of Neuroscience)
[2] CNRS UMR 5217, LIG (Laboratoire d'Informatique de Grenoble) MAGMA
[3] INRIA, Laboratoire Jean Kuntzmann, Universite de Grenoble (MISTIS)

Abstract. The Markov Random Field (MRF) probabilistic framework is classically introduced for a robust segmentation of Magnetic Resonance Imaging (MRI) brain scans. Most MRF approaches handle tissues segmentation via global model estimation. Structure segmentation is then carried out as a separate task. We propose in this paper to consider MRF segmentation of tissues and structures as two local and cooperative procedures immersed in a multiagent framework. Tissue segmentation is performed by partitionning the volume in subvolumes where agents estimate local MRF models in cooperation with their neighbours to ensure consistency of local models. These models better reflect local intensity distributions. Structure segmentation is performed via dynamically localized agents that integrate anatomical spatial constraints provided by an *a priori* fuzzy description of brain anatomy. Structure segmentation is not reduced to a postprocessing step: rather, structure agents cooperate with tissue agents to render models gradually more accurate. We report several experiments that illustrate the working of our multiagent framework. The evaluation was performed using both phantoms and real 3T brain scans and showed a robustness to nonuniformity and noise together with a low computational time. This MRF agent based approach appears as a very promising new tool for complex image segmentation.

1 Introduction

MRI Brain Scan Segmentation is a challenging task and has been widely addressed in the last 15 years. Difficulties arise from various sources including the size of the data, the high level of noise with strong field images (3T or higher), the intensity nonuniformity or the low contrast between tissues. The Markov Random Field (MRF) probabilistic framework is classically used to introduce spatial dependencies between voxels, providing a labeling regularization and a robust to noise segmentation [1],[2]. In addition to noise, the intensity nonuniformity is a critical issue. Most approaches estimate global tissue intensity models through the entire volume and require the estimation of an explicit "bias field" model to account for the intensity nonuniformity [3],[4]. These models are based on underlying assumptions that are not always valid and requires additional computational burden for their estimation. Local segmentation approaches are

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 13–23, 2007.

attractive alternatives [5],[6],[7]. The principle is to locally compute the tissue intensity models in various volume partitions. These models fit better to local image properties. In particular local segmentation intrinsically handles the intensity nonuniformity without any bias field modelization. Existing approaches either use local estimation as a preprocessing step to estimate a bias field model [5], or use redundant information to ensure consistency and smoothnesss between local estimated models [7], greedily increasing computation. We consider that these approaches do not fully take advantage of locality. We propose to embed a local MRF segmentation approach into a distributed and cooperative framework based on the multiagent (MA) paradigm. This paradigm provides convenient mechanisms of cooperation and coordination, allowing to ensure consistency of local models in an elegant way. In addition, we achieve segmentation of tissues and structures in a unified and cooperative way. We show how to design a MRF agent-based segmentation approach and report some experiments on MRI brain scans that exhibit the properties of the system. These efforts extend a first tentative of agent based region growing tissue segmentation [6] to a unified MRF tissue and structure segmentation.

This paper is organized as follows. We present in Section 2 how MRF processes are turned into cooperating agent entities. Section 3 reports evaluation results and presents several experiments to exhibit some interesting properties of such a local approach. Section 4 is devoted to the discussion and the conclusion.

2 MRF Agent Based Segmentation

2.1 MRF Framework

We consider a finite set of N voxels $V = \{1, ...N\}$ on a regular three-dimensional (3-D) grid. Our aim is to assign each voxel i to one of K classes considering the observed greylevel intensity y_i at voxel i. Both observed intensities and unknown classes are considered to be random fields denoted respectively by $\mathbf{Y} = \{Y_1, ..., Y_N\}$ and $\mathbf{Z} = \{Z_1, ..., Z_N\}$. Each random variable Z_i takes its value in $\{e_1, ..., e_K\}$ where e_k is a K-dimensional binary vector corresponding to class k. Only the k^{th} component of this vector is non zero and is set to 1. In a traditionnal Markov model based segmentation framework, it is assumed that the conditional field $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ is a Markov random field, *ie.*

$$p\left(\mathbf{z}\left|\mathbf{Y} = \mathbf{y}, \boldsymbol{\Phi}\right.\right) = W_{y,\Phi}^{-1} \exp\left(-H\left(\mathbf{z}\left|\mathbf{y}, \Phi\right.\right)\right)$$

where $H\left(\mathbf{z}\left|\mathbf{y}, \Phi\right.\right)$ is an energy function depending on some parameters $\Phi = (\Phi_y, \Phi_z)$ and given by:

$$H\left(\mathbf{z}\left|\mathbf{y}, \Phi\right.\right) = H\left(\mathbf{z}\left|\Phi_z\right.\right) - \sum_{i \in V} \log p\left(y_i\left|z_i, \Phi_y\right.\right). \tag{1}$$

This energy is a combination of two terms: the first term in (1) is a regularization term that accounts for spatial dependencies between voxels. Denoting by $\mathcal{N}(i)$ the neighbors of voxel i, we will consider a Potts model with external field:

$$H(\mathbf{z}\,|\varPhi_z) = \sum_{i\in V}\left[{}^t z_i v_i - \frac{\beta}{2}\sum_{j\in\mathcal{N}(i)} {}^t z_i z_j\right]. \tag{2}$$

The second summation above tends to favor neighbors that are in the same class when parameter β is positive. This β parameter accounts for the strengh of spatial interaction. Other parameters are the v_i´s that are K-dimensional vectors defining the so-called external field. In this case $\varPhi_z = \{v_1, ..., v_N, \beta\}$. The v_i's can be related to *a priori* weights accounting for the relative importance of the K classes at site i. The introduction of these extra parameters in the standard Potts model enables us to integrate *a priori* knowledge on classes. The second term in (1) is a data-driven term based on intensities. For MRI we generally consider a Gaussian probability density function for the observed intensity y_i when the tissue class is z_i. It follows that $p\,(y_i\,|z_i = e_k, \varPhi_y) = g_{\mu_k,\sigma_k}(y_i)$ with $\varPhi_y = \{\mu_k, \sigma_k, k = 1...K\}$. Segmentation is then performed according to the Maximum A Posteriori principle (MAP) by maximizing over $\mathbf{z}$ the probability $p\,(\mathbf{z}\,|\mathbf{y}, \varPhi)$. This requires the evaluation of an intractable normalizing constant $W_{\mathbf{y},\varPhi}$ and the estimation of the unknown parameters $\varPhi$. A standard approach is to use the ICM algorithm that alternates between parameter estimation and segmentation but results in biased estimates. EM-based algorithms and variants proposed by [8] can be rather considered. They are based on Mean-field like approximations which make the MRF models case tractable. In all these approaches, MRF estimation is performed globally through the entire volume.

2.2 Local Approach of Tissue Segmentation

We agentify the global MRF process by distributing in the volume several local MRF processes in a MA paradigm.

We consider a decentralized and memory shared MA framework based on the classical Agent/Group/Behaviour conceptual model. At the system starting point, a global tissue agent $\mathcal{A}_G^T$ is responsible for partitioning the volume into C non-overlapping territories $\{\mathcal{T}_c^T, c = 1...C\}$ and instantiating one situated tissue agent $\mathcal{A}_c^T$ per territory. We consider $K = 3$ tissue classes: CSF (Cephalo-Spinal Fluid), GM (Grey Matter) and WM (White Matter). The hidden tissue classes t_i's take their values in $\{e_1, e_2, e_3\}$ respectively for classes $\{e_{CSF}, e_{GM}, e_{WM}\}$. The tissue agent $\mathcal{A}_c^T$ segments its territory with the MRF model defined by the energy (see Equations (1) and (2)):

$$H^c(\mathbf{t}\,|\mathbf{y}, \varPhi^c) = \sum_{i\in\mathcal{T}_c^T}\left[{}^t t_i \lambda_i^c - \frac{\beta^c}{2}\sum_{j\in\mathcal{N}(i)} {}^t t_i t_j - \log p\left(y_i\,\middle|t_i, \varPhi_y^c\right)\right], \tag{3}$$

where the parameters $\varPhi^c = \left\{\varPhi_t^c, \varPhi_y^c\right\}$ have to be estimated. The external field denoted by $\{\lambda_i^c, i \in \mathcal{T}_c^T\}$ is not estimated but used to incorporate information coming from structure segmentation agents (see Section 2.3). Each tissue agent owns three behaviours to segment its territory: 1) *Tissue_Init* which initializes the EM algorithm by computing initial intensity models via a Fuzzy-C Mean algorithm (FCM), 2) *Tissue_EM* which computes the MRF model parameters in cooperation with neighbouring agents and 3) *Tissue_Stabilized* when the agent is in the idle state.

Cooperation between tissue agents: The agent $\mathcal{A}_c^T$ cooperates with its neighbors group denoted by $\mathcal{G}_N(\mathcal{A}_c^T)$ to ensure a global consistency of the local estimated model. It performs:

Model Checking: for each tissue class k we compute a local mean model from the models of $\mathcal{G}_N(\mathcal{A}_c^T)$. The KullBack-Leibler distance $\mathcal{D}_k^c$ provides a measure of dissimilarity between intensity models of each class that allows to apply model correction if required.

Model Correction: we compute the corrected mean and variance of class k from a linear combination of estimated and mean models according to $\mathcal{D}_k^c$.

Model Interpolation: we use cubic splines interpolation between corrected models of $\mathcal{A}_c^T$ and of $\mathcal{G}_N(\mathcal{A}_c^T)$ to compute one intensity model $\{g_{\mu^c_{i,k},\sigma^c_{i,k}}, i \in \mathcal{T}_c^T, k \in \{e_{CSF}, e_{GM}, e_{WM}\}\}$ per voxel. There are two benefits from interpolating models. First, this ensures smooth model variation between neighboring agents. Second, this intrinsically handles nonuniformity of intensity inside each agent.

Coordination between tissue agents: Each agent enters into an idle mode after its initialization. The global tissue agent $\mathcal{A}_G^T$ then computes a global tissue intensity model using the FCM algorithm and wakes up the 20 percents agents whose local model is nearest from the global model. Once each $\mathcal{A}_c^T$ has finished its local model estimation, it wakes up its neighbors. First, it allows them to perform estimation in turn. Second, it allows already stabilized agent to perform *model checking*, restarting their estimation if needed to take into account updated models of $\mathcal{A}_c^T$ in their *model correction* and *model interpolation*.

2.3 Cooperative Approach of Tissue and Structure Segmentation

Automatic structures segmentation can not rely only on radiometry information because intensity distributions of grey nucleus are largely overlapping. *A priori* knowledge should be introduced. Classical approaches rely on an *a priori* known atlas describing anatomical structures. Atlas warping methods are however time consuming and limited due to inter-subject variability. A recent different way to introduce *a priori* anatomical knowledge is to describe brain anatomy with generic fuzzy spatial relations [9],[10]. We generally consider three kind of spatial relations: distance, symmetry and orientation relations. They are expressed as 3D fuzzy maps to take into account the general nature of the provided knowledge. We describe each subcortical structure by a set of generic fuzzy spatial relations provided by a brain anatomist. Fusion operators between fuzzy sets then permits to combine the knowledge provided by each spatial relation and provides a Fuzzy Localization Map (FLM) of the structure in the volume.

We define one structure agent $\mathcal{A}_l^S$ per structure and currently consider $L = 9$ subcortical structures: the ventricular System, the two Frontal Horns, the two Caudate Nucleus, the two Thalamus, and the two Putamens. Each structure agent segments its territory with a MRF model of $K = 2$ classes referred as *structure* and *background*. The FLM f^l of structure l is used in two ways: first it dynamically provides the structure territory $\mathcal{T}_l^S$ containing the structure l by a simple thresholding. Second we propose to integrate it as an *a priori* anatomical knowledge in the MRF framework. Denoting by $\mathbf{s} = \{s_i, i \in \mathcal{T}_l^S\}$ the hidden classes, the energy function of the MRF model of $\mathcal{A}_l^S$ is given by:

$$H^l\left(\mathbf{s}\left|\mathbf{y}, \varPsi^l\right.\right) = \sum_{i \in \mathcal{T}_l^S} \left[{}^t s_i \alpha_i^l - \frac{\beta^l}{2} \sum_{j \in \mathcal{N}(i)} {}^t s_i s_j - \log p\left(y_i \left| s_i, \varPsi_y^l\right.\right) \right], \qquad (4)$$

with $\varPsi^l = \left\{\varPsi_s^l, \varPsi_y^l\right\}$ and $s_i \in \{e_1, e_2\} = \{e_B, e_S\}$ for a voxel of the background or a voxel belonging to structure l. The external field denoted by $\{\alpha_i, i \in \mathcal{T}_l^S\}$, where α_i^l is a 2-dimensional vector, allows to incorporate the *a priori* knowledge contained in the FLM. We denote by f_i^l the value of f^l at voxel i and propose to introduce the prior fuzzy knowledge of the FLM as relative prior weights for each voxel i, by setting:

$$\alpha_i^l = \gamma \begin{bmatrix} -\log\left(1 - f_i^l\right) \\ -\log f_i^l \end{bmatrix}, \qquad (5)$$

where γ adjusts the influence of the external field. When $f_i^l \approx 0$, the voxel i is unlikely to belong to the structure. If γ was null, the segmentation would be performed only from the intensity models. Else, according to (5), $\alpha_i^l(1) < \alpha_i^l(2)$ which favours in (4) the class *background*. When $f_i^l \approx 1$, the voxel i is likely to belong to the structure. In that case $\alpha_i^l(1) > \alpha_i^l(2)$ and the class *structure* is favored.

Each structure owns four behaviours: 1) *Struct_Init* that initializes the structure agent, 2) *Struct_ComputeFuzzyMap* that computes or updates the FLM from fuzzy spatial relations, 3) *Struct_EM* that computes MRF model parameters and 4) *Struct_Stabilized* when the agent is in the idle state. As we define a neighborhood group for each tissue agent, we define the group $\mathcal{G}_{T\to S}\left(\mathcal{A}_l^S\right)$ of tissue agents cooperating with a structure agent $\mathcal{A}_l^S$, the group $\mathcal{G}_{S\to T}\left(\mathcal{A}_c^T\right)$ of structure agents cooperating with a tissue agent $\mathcal{A}_c^T$, and the group $\mathcal{G}_{S\to S}\left(\mathcal{A}_l^S\right)$ of structure agents using $\mathcal{A}_l^S$ as a reference in a spatial relation. These groups allow to detail cooperation and coordination mechanisms between agents.

Updating structure models via tissue models: each structure being composed of a single tissue $T^l \in \{e_{CSF}, e_{GM}, e_{WM}\}$, we do not estimate intensity model Ψ_y^l of class *structure* and class *background*. We rather compute them from tissue intensity models of $\mathcal{G}_{T\rightarrow S}\left(\mathcal{A}_l^S\right)$ estimated by tissue agents, by setting:

$$\begin{cases} p\left(y_i \left| s_i = e_S, \Psi_y^l\right.\right) = p\left(y_i \left| t_i = T^l, \Psi_y\right.\right) \\ p\left(y_i \left| s_i = e_B, \Psi_y^l\right.\right) = \max\limits_{t \in \{e_{CSF}, e_{GM}, e_{WM}\}} p\left(y_i \left| t_i = t, \Psi_y\right.\right) \end{cases} ,$$

so that improvements in tissue intensity models estimation will be dynamically taken into account by structure agents.

Feedback of Structure Segmentation on Tissue Segmentation: conversely, results from structure agents are integrated in the tissue segmentation model via the external field λ^c of tissue agents. We express it as the disjunctive fusion of a posteriori probabilities $p\left(\mathbf{s} \left| \mathbf{y}, \Psi^l\right.\right)$ coming from structures of $\mathcal{G}_{S\rightarrow T}\left(\mathcal{A}_c^T\right)$. It follows that structure segmentation is not reduced to a second step but is combined with tissue segmentation to improve its performance.

Update Fuzzy Maps: when the segmentation of structure l is updated the structure models of $\mathcal{G}_{S\rightarrow S}\left(\mathcal{A}_l^S\right)$ take it into account by re-computing their spatial relations with respect to l, making the knowledge gradually more accurate.

3 Evaluation

The evaluation was performed using both phantoms and real 3T brain scans. We first evaluated the tissue segmentation performances provided by tissue agents only. We quantitatively compared our approach to two well known approaches, FSL [2] and SPM5 [4], with the Jaccard similarity[1] on the BrainWeb[2] [11] database (see Figure 2). It shows comparable results and particularly more robustness to noise than SPM5, whereas the computational time was approximately 4min with our approach and respectively 8min and 14min with FSL and SPM5 on a 4Ghz Pentium, 1Go RAM. The Figure 3 shows visual evaluation on a very high bias field brain scan[3]. SPM5, which uses an a priori atlas, failed in the segmentation while FSL did not estimate a correct bias field. Our local approach clearly appears to be more robust to very high intensity inhomogeneities. Figure 4 shows results of cooperative tissue and structure segmentation. In addition to these competitive performances, our platform allows us to demonstrate some interesting properties of this local approach. Figure 5 shows that large territory sizes result in poor performance. Smaller territory sizes allow to better model local intensity distributions, but need to be large enough to correctly

[1] The Jaccard similarity between two sets is defined by the size of their intersection divided by the size of their union. A value of 1.0 represents a complete agreement.

[2] BrainWeb provides MR volumes whose the segmentation is known (semi-manually segmented by experts) and to which different values of noise and nonuniformity can be added.

[3] This image was acquired with a surface coil which provides a high sensitivity in a small region (here the occipital lobe) and is useful for functional imaging.

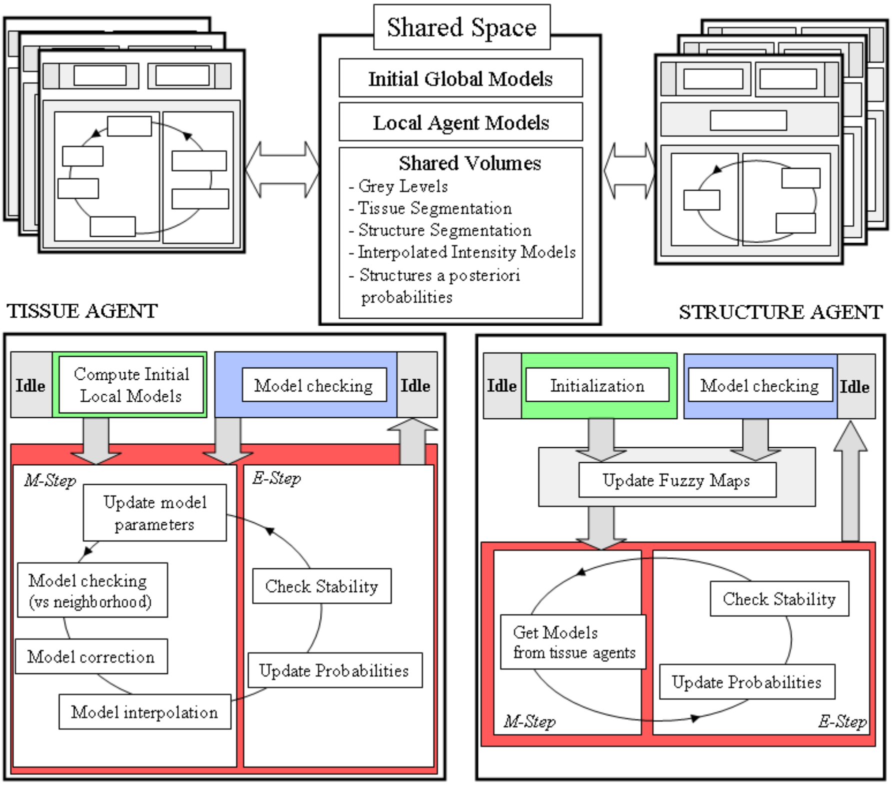


Fig. 1. Synthetic view of our agent based approach

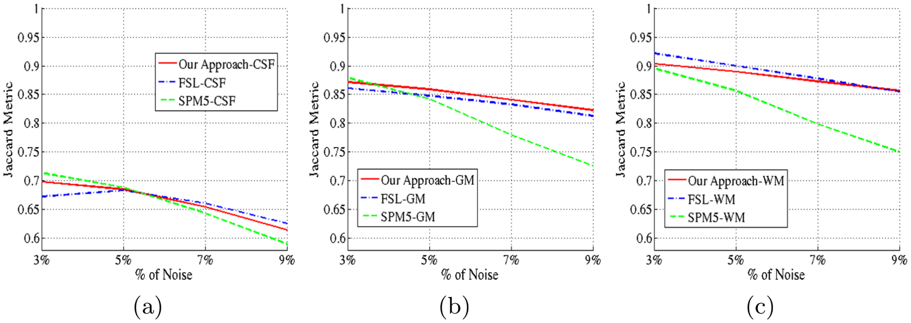


Fig. 2. Comparison of our approach to FSL and SPM5 for tissue segmentation on the BrainWeb phantoms with 40% of nonuniformity and different noise values. Evaluation for class CSF (a), GM (b) and WM (c) classes.

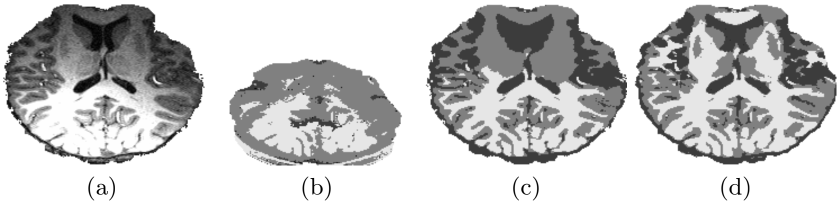


Fig. 3. Evaluation on a MRI brain scan with very high intensity nonunformity (a). Tissue segmentation provided by SPM5 (b), FSL (c) and our approach (d).

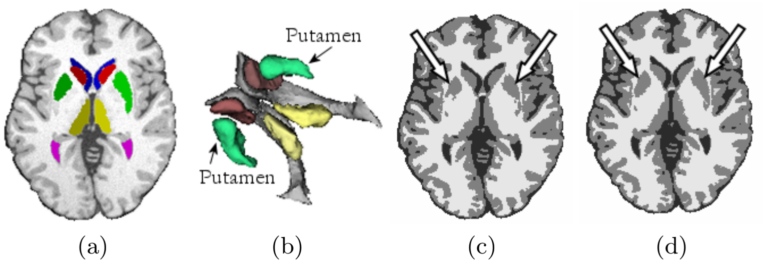
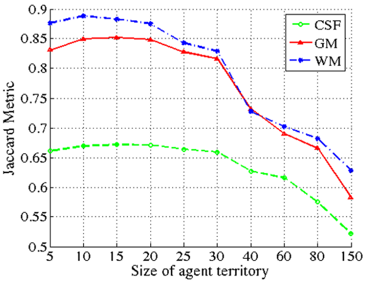


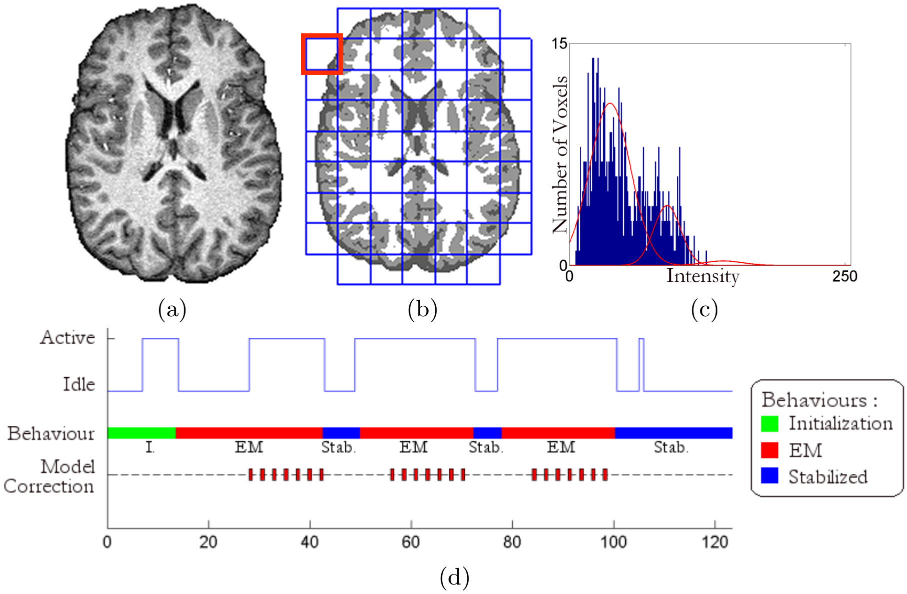
Fig. 4. Cooperative tissue and structure segmentation: structure segmentation (a) with 3D rendering (b) shows good results with computational time less than 15 minutes. Image (d) shows visual improvement in tissue segmentation compared to the segmentation produced by tissue agents only (c) (see putamens).



Size of agent territory	Number of agents
5	17422
10	2474
15	807
20	379
25	212
30	137
40	76
60	18
80	8
150	1

Fig. 5. Influence of tissue agent territory size on the robustness to intensity nonuniformity, using the BrainWeb phantom with 3% of noise and 100% of nonuniformity

estimate models. In practice we use territory size of 20x20x20 voxels. Figure 6 illustrates the activity of a tissue agent immersed in the system. It allows to understand the behaviour of an agent that would have poor estimated models without cooperation mechanisms.



(a) (b) (c)

(d)

Fig. 6. Evaluation of tissue segmentation (b) on a real 3T brain scan (a). Our implementation allows to observe the activity of each agent: (c) shows the local histogram of the red-outlined agent in (b). All tissue classes are not represented, leading to bad models estimation without cooperation. Figure (d) shows the agent behaviour: after initialization it waits for the neighbouring agents to wake it up. Then it performs model corrections in its EM behaviour. After convergence (Stabilized behaviour), model checking restarts the estimation twice. Then, the agent stay in idle mode.

4 Discussion and Conclusion

We propose an original MRF agent based approach for MRI brain scan segmentation. Markovian segmentation is an efficient probabilistic framework which introduces a local regularization via spatial dependancies between voxels. It makes it robust to local pertubations such as data noise. In addition, accurate segmentation of MRI brain scan must take into account the intensity nonuniformity. These global perturbations require to spatially adapt models over the volume. Instead of estimating a spatial bias field model or using non-tractable non-stationnary MRFs, we propose an original agent based approach: we integrate the local level of regularization via Markov models and the global level via distributed and cooperative agents which estimate local models. Coordination mechanisms are introduced (1) to ensure the spreading of knowledge between neighbouring agents and (2) to prevent agents from starting with poor knowledge. The main advantage over other existing local approaches [5],[6],[7] is that using this appropriate framework we are able to make full use of local estimation.

In addition, it allows to integrate *a priori* information in an elegant way. Indeed, we can consider two levels of knowledge as regards MRI brain scans: the tissue knowledge at a local level and the brain structure knowledge at a global level. In general these two levels are processed independantly [9],[10]. We rather consider that they are linked and must be used in a common setting. We show how to introduce *a priori* anatomical knowledge expressed by fuzzy spatial relations in the MRF framework to segment several subcortical structures. We also show how to combine the MRF models of tissues and structures. Models are mutually constrained in their convergence, making both models gradually more accurate and providing optimal results for both levels. We thus show how the Markovian approach can be extended by introducing mechanisms (1) to handle multiple levels of regularizations (multi-scale regularization) and (2) to consider conjointly multiple level of heterogeneous knowledge (multi-scale knowledge processing). Our implementation provides interesting visualization tools which allow us to track specific agents and explore local knowledge such as local segmentation errors or local intensity models. The evaluation shows competitive results compared to other algorithms with lower computational time. Indeed, local easy-to-segment territories converge quickly, allowing the system to focus on other areas. In addition, it appears to be robust to the intensity nonuniformity without any bias field assumption and estimation. Note that we currently consider a regular cubic partitioning for tissue segmentation but should in future work evaluate the contribution of more particular partitionning, like adaptive or spherical partitionings. The locality may allow to integrate brain extraction as an additional level of knowledge in our processing. We also plan to extend our model to new structures to study specific pathologies. Finally, MRF agent based computing appears as an interesting and modular tool for complex image segmentation.

References

1. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imag. 18(10), 897–908 (1999)
2. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximisation algorithm. IEEE Trans. Med. Imag. 20(1), 45–47 (2001)
3. Wells, W.M., Grimson, W.E.L., Kikinis, R., Jolesz, F.A.: Adaptative segmentation of MRI data. IEEE Trans. Med. Imag. 15(4), 429–442 (1996)
4. Ashburner, J., Friston, K.: Unified segmentation. NeuroImage 26, 839–851 (2005)
5. Shattuck, D.W, Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M.: Magnetic resonance image tissue classification using a partial volume model. NeuroImage 13(5), 856–876 (2001)
6. Richard, N., Dojat, M., Garbay, C.: Automated segmentation of human brain MR images using a multi-agent approach. Artificial Intelligence in Medicine 30, 153–175 (2004)
7. Zhu, C., Jiang, T.: Multicontextual fuzzy clustering for separation of brain tissues in magnetic resonance images. NeuroImage 18(3), 685–696 (2003)

8. Celeux, G., Forbes, F., Peyrard, N.: EM procedures using mean field-like approximations for model-based image segmentation. Pattern Recognition 36(1), 131–144 (2003)
9. Barra, V., Boire, J.: Automatic segmentation of subcortical brain structures in MR images using information fusion. IEEE Trans. Med. Imag. 20(7), 549–558 (2001)
10. Colliot, O., Bloch, I., Camara, O.: Description of brain internal structures by means of spatial relations for MR image segmentation. In: SPIE Medical Imaging, San Diego, vol. 5370, pp. 444–445 (2004)
11. Collins, D., Zijdenbos, A., Kollokian, V., Sled, J., Kabani, N., Holmes, C., Evans, A.: Design and construction of a realistic digital brain phantom. IEEE Trans. Med. Imag. 17(3), 463–468 (1998)

R-CAST-MED: Applying Intelligent Agents to Support Emergency Medical Decision-Making Teams

Shizhuo Zhu[1], Joanna Abraham[1], Sharoda A. Paul[1], Madhu Reddy[1],
John Yen[1], Mark Pfaff[1], and Christopher DeFlitch[2]

[1] College of Information Sciences and Technology, The Pennsylvania State University,
Unversity Park, PA 16802, USA
{szhu,jabraham,spaul,mreddy,jyen,mpfaff}@ist.psu.edu
[2] Department of Emergency Medicine, Penn State Milton S. Hershey Medical Center,
Penn State College of Medicine, Hershey, PA 17033, USA
cdeflitch@psu.edu

Abstract. Decision-making is a crucial aspect of emergency response during mass casualty incidents (MCIs). MCIs require rapid decisions to be taken by geographically-dispersed teams in an environment characterized by insufficient information, ineffective collaboration and inadequate resources. Despite the increasing adoption of decision support systems in healthcare, there is limited evidence of their value in large-scale disasters. We conducted focus groups with emergency medical services and emergency department personnel who revealed that one of the main challenges in emergency response during MCIs is information management. Therefore, to alleviate the issues arising from ineffective information management, we propose R-CAST-MED, an intelligent agent architecture built on Recognition-Primed Decision-making (RPD) and Shared Mental Models (SMMs). A simulation of R-CAST-MED showed that this tool enabled efficient information management by identifying relevant information, inferring missing information and sharing information with other agents, which led to effective collaboration and coordination of tasks across teams.

Keywords: Intelligent Agent, Simulation, Mass Casualty Incidents, Decision Support Systems, Information Management.

1 Introduction

Emergency medical decision-making is complex, especially during mass casualty incidents. During a *Mass Casualty Incident* (MCI), hospitals are required to deal with a large influx of patients with various levels of trauma in a short period of time [1]. The key goals for healthcare providers during such MCIs include rapid evacuation of patients from the incident site, and quick provision of critical medical care to a large number of patients [2], etc. These activities require a coordinated effort on the part of pre-hospital and hospital-based teams. Medical decisions taken by patient-care personnel during such crises, both individually and collectively, significantly impact the mortality rate of critically injured patients.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 24–33, 2007.

Pre-hospital services such as police, fire, EMS (Emergency Medical Services) and HAZMAT (Hazardous Materials) teams are responsible for ensuring that patients are stabilized and transported rapidly. They need to resolve such issues as: how many patients need to be transported, how to transport the patients, and which facilities are best suited to handle the patients. The decision on where to transport patients is usually based on the trauma levels of the patients and on particular emergency department (ED) capabilities. The EDs receiving patients are required to make decisions such as, how to triage the large number of incoming patients, whether to seek assistance from other departments, and when to alert other EDs or request additional resources. Though the pre-hospital and ED teams make different kinds of decisions, there is a *decision dependency* between these different teams.

In order to gain insight into the difficulties with MCI decision-making, we conducted focus groups with both EMS and ED teams at a major teaching hospital. Participants were presented with the scenario of a train derailment involving hazardous materials and asked to describe how they would react to events and the prominent challenges they would face. Based on their responses, we discovered that timely access to relevant information is not only a major requirement but also a major challenge for decision-making during a MCI. For instance, information required by ED team to make decisions, such as how many beds to prepare for incoming patients, depends on information available to and provided by EMS team, such as how many patients are en-route to that ED. This *information dependency* plays a key role in decision dependency.

Computer-based decision-support systems have been used for clinical and administrative purposes in a variety of settings; however they have rarely been applied to decision-making during MCIs. Existing clinical decision support systems (CDSSs) are primarily used to facilitate decisions regarding a *single patient* and by a *single team* of healthcare providers, and thereby limited in their ability to deal with MCIs where decisions are made about multiple patients by multiple inter-professional teams. Based on our fieldwork, we propose an agent-based emergency medical decision support system, R-CAST-MED, to help healthcare providers deal with the challenge of information management during MCIs.

The following section provides background on decision-making in MCIs and the use of decision support systems in healthcare. Section 3 and 4 describe the architecture of R-CAST-MED and the simulation of a particular MCI scenario. In section 5, we discuss the significance of the simulation and the design recommendations to better support decision-making during MCIs. Finally, we conclude with some thoughts on role of agents in medical decision-making and future work in section 6.

2 Background

A mass casualty incident is any situation or event that places a significant demand on medical equipment and personnel [3]. Healthcare providers involved in dealing with patients of a MCI have to deal with a variety of challenges including organizational, logistical, and patient-care related [1]. Their response to these challenges will effect the mortality rate of critically injured patients. The post incident analyses of major MCIs such as the World Trade Center attacks in 2001 [3], the London bombings in

2005 [2], etc. have performed careful assessment of the response to these MCIs. The analysis found that decisions made during such incidents played a crucial role in the outcome of the incidents. For instance, during the World Trade Center attacks, the decision on where to locate the emergency management command post was a mistake and decisions taken by emergency responders to transport all the initial patients to the three nearest hospitals overwhelmed those institutions [3]. These examples highlight the complexity of the decision-making process during MCIs because of the large number of people involved, time pressure, and the uncertainty inherent in dealing with a new situation.

Decision Support Systems (DSSs) have been employed in healthcare to serve different purposes. For example, IDEAS for ICUs [5] makes use of case-based reasoning tools for disease diagnoses of patients. Despite their value in areas of healthcare, there has been limited research on DSS applications for crisis management [6]. Some decision support applications for emergency response during a crisis situation are “iRevive”, a mobile pre-hospital database system that supports point-of-care electronic patient data capture that assists in triage decision-making [7]; “Automated Triage Management (ATM)”, a decision support model that assists healthcare practitioners to find patients’ chief complaints [8]; and “Mobile Emergency Triage (MET)”, a DSS model designed for pediatric population [9].

Though these DSSs accelerate the clinical diagnosis process during MCIs, they do not explicitly support dependencies in work, such as filtering and sharing appropriate information among multiple professional teams.

3 R-CAST-MED

3.1 Focus Groups

To better understand the challenges associated with MCI decision-making as well as to examine ways to support and improve the same, we conducted 7 focus groups with EMS and ED personnel associated with a 500-bed teaching hospital. We presented participants with the following scenario of a train derailment incident.

> ***Scenario:*** *A 76-car Norfolk Southern freight train carrying hazardous materials derailed in Derry Township, Dauphin County. The track where the derailment occurred runs parallel to E. Hershey Park Drive and is close to the golf course of the Country Club of Hershey. Patients of this derailment are being brought into the ED while the ED is operating at capacity.*

The 21 participants included air and ground EMS, attending and resident physicians, and communication center personnel. We presented participants with the scenario, and asked them questions regarding their decision-making process during the MCI.

We discovered that the presence of geographically distributed teams of pre-hospital and hospital personnel with varying goals, training levels, priorities, and information requirements increased the complexity of the situation. We also found that the decisions made by one team depended on the decisions made by other teams. This decision dependency arose primarily out of information dependency, i.e. decisions made by teams during MCIs required up-to-date, accurate and relevant information to be exchanged between them. When presented with the train derailment scenario, some questions asked by ED physicians included “How many patients are involved?”, “What is the acuity level of patients coming to the ED?” This incoming information helped the ED team make decisions on how many ED beds and trauma bays to prepare, whether to set up decontamination tents, etc. The primary source of information for them was the communication center of the hospital which originally received information from the on-site first responders.

The complex and dynamic nature of a MCI necessitates the need for a decision support system that is user-friendly with the flexibility to choose how information is sent, received, filtered and shared, depending on the context of the crisis environment. To address some of the challenges identified in the focus groups, we develop R-CAST-MED on the basis of R-CAST (RPD-enabled Collaborative Agents Simulating Teamwork) [10] to support healthcare providers in decision-making tasks by filtering, proactively gathering, providing and sharing relevant information.

3.2 R-CAST

Cognitive Foundations. R-CAST is a collaborative agent architecture built on cognitive models, *Recognition-Primed Decision-making (RPD)* model and *Shared Mental Models (SMMs). RPD* model [11] describes how experienced decision makers make decisions under time pressure in real situations. It argues that human experts usually make decisions based on their past experiences. They select an experience that worked before for similar situations, instead of calculating and comparing expected utility for each decision choice. *SMM* is a hypothetical cognitive construct that refers to a common understanding among team members regarding their objectives, roles, knowledge etc. SMM attempts to explain many of the human behaviors in high performance teams [12].

R-CAST Agent Architecture. R-CAST is a RPD-enabled collaborative agent architecture extended from CAST (Collaborative Agents Simulating Teamwork) [10]. From a software engineering perspective, R-CAST is a component-based configurable agent architecture, i.e. each agent is configured by enabling/disabling components depending on particular applications. This adaptive feature allows R-CAST to be well-suited for the medical domain.

Fig. 1 depicts the basic architecture of a R-CAST agent. The knowledge base manager, information manager, communication manager, RPD-based decision making manager, and process manager are the key components. The knowledge base, experience base, and plan library are the repositories that contain inferential knowledge, experiential knowledge, and procedural knowledge respectively. The knowledge base defines *fact types* that the agent is able to understand, *rules* that the agent uses to infer new information, and primitive *facts* that the agent has already known. The experience base comprises of tree-like *experience spaces*, where every single experience encapsulates *cues*, *expectancies*, *goals*, and *course of actions (COA)*. The plan library specifies how the agent executes the COA in the form of *plans* and *operators*. The *domain adapter* is the interface between an agent and its surrounding environment, specifying domain-dependent functions and capabilities.

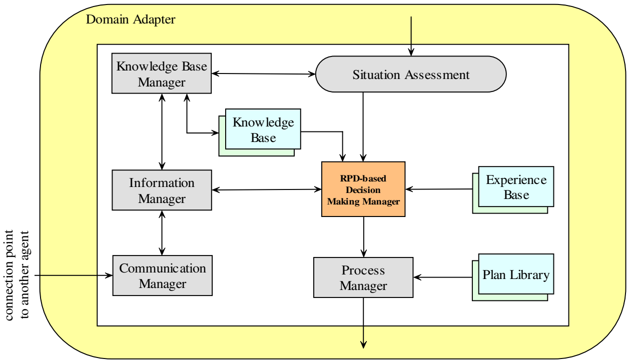


Fig. 1. R-CAST Agent Components

In an application, a R-CAST agent updates the knowledge base with newly obtained information through *knowledge base manager* by constant observation and assessment of the situation. Meanwhile, feature matching is performed by comparing the current situation with the cues of existing experiences by *RPD-based decision making manager*. If an experience is matched, the COA corresponded to this experience is captured and executed by the *process manager*. In cases of no match, the *information manager* identifies missing information requirements, and submits information inquiry requests to the *communication manager*. Upon receiving an information inquiry request, the communication manager tries to connect and exchange information with other agents that are potential information sources.

3.3 R-CAST-MED: Adapting R-CAST to Emergency Medical Domain

The cognitive basis of R-CAST makes it well-suited for decision-making under emergency medical situations. R-CAST-MED is a collaborative human-agent team architecture that involves human teams (such as EMS and ED teams) and their supporting agents, shown in Fig. 2.

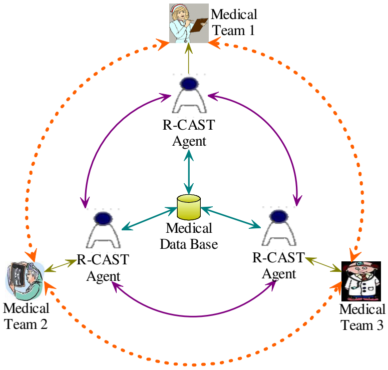


Fig. 2. R-CAST-MED Architecture

Each decision-making team is supported by an R-CAST agent. Decision makers read and write information from and to the computers. The agent behind the computer receives and analyzes information for the decision maker. If there is a decision in need, the agent makes a recommendation to its user according to its knowledge and experiences. If necessary, the information will be proactively exchanged between any two R-CAST agents, with no need of explicit request from decision makers. The agents may need to access the medical data base for general information.

R-CAST-MED utilizes and formalizes the information dependency feature of MCIs to support better decision-making. This information dependency feature should be understood within certain *context*, which is formalized in R-CAST-MED as *inferential context*, *experiential context*, and *procedural context*. This feature enables effective information management of R-CAST-MED. The issue of information overload is alleviated by filtering irrelevant information. The information sharing requirement is supported by identifying missing information through inferring lower level information from higher level information. The information dependency feature allows distributed information to be appropriately exchanged and used across teams. Therefore, R-CAST-MED makes it possible for healthcare teams including pre-hospital and hospital services to quickly process and fuse information from multiple sources to make decisions in crisis management

4 Simulation

The goal of this simulation was to examine how information was appropriately filtered, sought, and shared among agents, and how decision recommendations were made by agents depending on this information. We primarily focused on agents' abilities to interact with other agents. The train derailment scenario provided in section 3.1 was used as an input to the simulation. It was performed on a GUI (Graphical User Interface) platform adapted from NeoCITIES [13].

Based on the scenario, we built four agents corresponding to four teams (Fig. 3): 911 county communication center (911CCC), hospital communication center (HCC), EMS, and ED. Each team was equipped with a GUI and a supporting agent. The R-CAST-MED agents were created by configuring their knowledge bases, experience bases, and plan libraries based on the scenario data. We employed a server to periodically generate events and send reports to agents based on a predefined scenario text file. The server continuously sent various types of information, such as event location and number of patients, to relevant agents and teams. For instance, in the scenario, the information that a patient is calling 911 for help would be sent to the 911CCC agent. Therefore, based on the incident information received from the server or from other agents, the responsible agent made decision recommendations which were later displayed on the GUI.

The agents share some common knowledge, but differ in others specific to their respective context. To carry out certain types of tasks, an agent is required to know who would be the potential information source. For instance, the fact that EMS agent

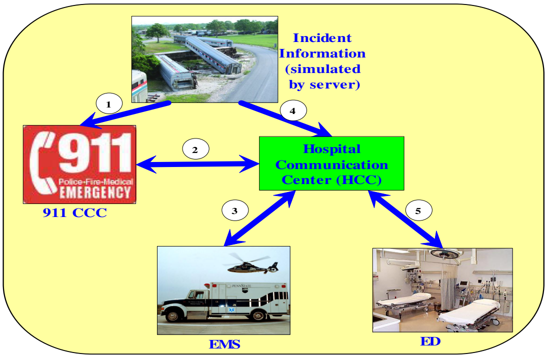


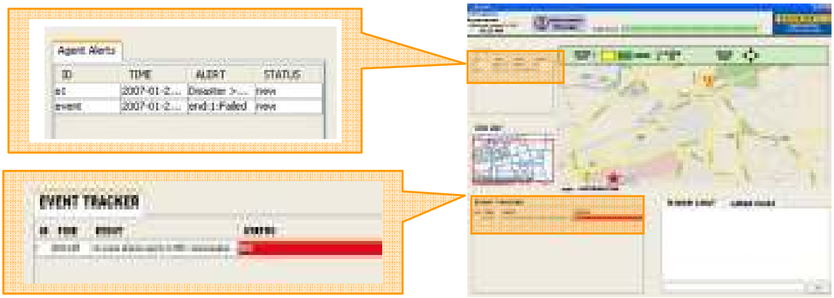
Fig. 3. Information Flow among Teams during a Train Derailment Scenario

sought information regarding available ED resources from HCC agent was indicated by the following representation:

```
(FactType ED_resource(?type ?amount)
                (template "ED has ?amount much ?type type resource")
                (source
                                (HCC plan_inform)
                )
)
```

Similarly, the HCC agent sought information from the ED agent. Thus a chain of information seeking was created to capture the information dependency across different agents. The required information was delivered back to the requesting agent as soon as one of the requested agents in the chain had obtained it.

Agent recommendations were displayed on the GUI (Fig. 4), which is composed of four main panels: a map that locates the incident (upper right); a chat box for domain experts to exchange information (bottom right); an event tracker panel that provides the event description (bottom left); and an agent alert panel that displays agent decision recommendations (upper left). Features such as information seeking and sharing between agents are not depicted on the GUI.



ID	TIME	ALERT	STATUS
a1	2007-01-2...	Disaster >...	new
event	2007-01-2...	end:1:Failed	new

Fig. 4. Graphic User Interface for Displaying Disaster Scenario

As soon as the agent received an event report, the corresponding information was displayed on the event tracker panel. The agent recommendation was displayed on the agent alert panel. Fig. 4 shows an example where a MCI event with information on the number and severity of patients is reported and the agent recommends the ED agent "to activate the disaster plan" because of the high number of potential patients.

The simulation demonstrated that agents can make decisions by effectively sharing and managing information. The results of the simulation showed that relevant and accurate information was exchanged between agents, and the appropriate decision recommendations were made. These decision recommendations were consistent with the data provided by the focus groups on what decisions they would make. We believe that this simulation highlights the potential for the R-CAST-MED agents to provide support for multiple teams to effectively collaborate and share appropriate and relevant crisis-related information without placing excess cognitive and affective constraints on the decision maker. For instance, in the simulation, the HCC agent responded to particular cues and delivered the information about the event location to the EMS agent without overloading the EMS agent with other irrelevant and extraneous information.

5 Discussion

5.1 Supporting Decision Dependency and Multi-team Decision-Making

For intelligent agent systems to play a useful role during a MCI, they must be able to facilitate and support decision dependencies and multi-team decision-making. In the following paragraphs, we use real-world examples to illustrate how R-CAST-MED supports these two key features.

The EMS after arriving at the scene, assesses the situation at the incident site to decide "whether to transport patients to ED" and "how many patients the ED can accommodate". In order to accomplish this goal, the *decision making* component of the EMS agent compares the current situation (e.g. number of patients) with the experiences in the experience base (e.g. how many patients should be transported to the ED, whether the patients need immediate trauma care). This component chooses one of the two paths: 1. if there is a match, the decision choice will be made and its corresponding COA would be selected from the plan library and executed by the process manager; 2. if there is no match, it will request the information manager for missing information (e.g. ED resource availability); in cases where the information manager cannot find such information in its local knowledge base, it requests the communication manager to contact another agent (which is affiliated to another team) for this missing information critical for decision-making.

As illustrated in the above example, decisions are interrelated because a decision regarding the transportation of patients to ED is dependent on the information provided by the ED agent (through communication center agent) to the communication manager of the EMS agent. The decision dependency feature is reflected in R-CAST-MED in several forms including contextual information dependency (derived from situation); inferential information dependency (based on rules built in knowledge base); and team-across information dependency (arises from communication across teams).

The second distinguishing feature of R-CAST-MED is its ability to support multi-team decision-making. R-CAST-MED can be used by teams composed of different professionals with varying skill levels that provide integrated care during a dynamic situation resulting in an influx of multiple patients. For example, The EMS agent furnishes the ED agent with details about patients' medical history, vitals and also, performs initial triage at the incident site prior to transport. Upon receiving this information from EMS agent, the ED agent can make necessary arrangements for patients that can be directly assigned to beds without repeating the triage process. The coordination support among multiple teams provided by R-CAST-MED leads to better quality of patient care given the rapid nature of the situation.

5.2 Designing Decision Support Systems to Support MCI

The chaotic and dynamic nature of MCIs causes inadequate access to relevant information, ineffective inter-team collaboration, isolated and redundant activities, communication breakdowns, and other affective and cognitive overload. To be effective in these environments, we need to design decision support systems (DSSs) that have (1) better contextualization features and (2) more proactive and rapid learning capabilities.

First, context is gaining increased attention as we are moving towards a more dynamic and integrated health system. Understanding the context of the information need based on the complexity of the situation is an important requirement for a DSS. Some DSSs such as R-CAST-MED have incorporated some contextual features. However, they still lack robust temporal and spatial contextual features that allow them to adapt to varying dynamic situations. Therefore, we must develop DSS that incorporate context in a more meaningful way.

Second, by improving the learning ability of DSSs, we can support decision-making in varying environments. Supporting the learning feature helps in identifying hidden associations in both explicit and implicit information that could be temporally and spatially distributed. This learning requirement necessitates DSSs to proactively synthesize new knowledge based on their ability to retain and recollect from past experiences. To support learning of DSSs, we need to understand human learning processes. In addition, learning algorithms such as Bayesian learning and case-based learning should be examined in order to verify its applicability in dynamic situations.

6 Conclusion

The decision-making process during MCIs is complex in nature. There are multiple factors that influence the decisions made by emergency responders including the dynamic nature of the incident, the need to access relevant information rapidly, sharing of accurate information, resource constraints, and coordination among teams. In this paper, we investigated a prominent challenge of a MCI that deals with effective information management. To address this challenge, we developed R-CAST-MED, a decision support system that achieved effective agent-agent interaction. Based on our simulation of R-CAST-MED, we confirmed that it helps in supporting effective information management therefore leading to better coordination of care.

Although our simulation highlighted the agent-agent interaction in R-CAST-MED, we did not evaluate the human-agent interaction. In our future research, we plan on incorporating human decision makers into our evaluation to verify whether the system can assist humans in improving situation awareness and decision-making effectiveness.

Acknowledgments. The authors would like to thank Victor Pilewski for his support during data collection process.

References

1. Hirshberg, A., Holcomb, J.B., Mattox, K.L.: Hospital Trauma Care in Multiple-Casualty Incidents. A Critical View Annals of Emergency Medicine 37(6), 647–652 (2001)
2. Aylwin, C.J., Konig, T.C., Brennan, N.W., Shirley, P.J., Davies, G., Walsh, M.S., Brohi, K.: Reduction in Critical Mortality in Mass Casualty Incidents: Analysis of Triage, Surge, and Resource Use After the London Bombings on July 7, 2005. In: Lancet, vol. 368(9554), pp. 2219–2225 (2006)
3. Asaeda, G.: The Day that the START Triage System Came to a STOP: Observations from the World Trade Center Disaster. Academic Emergency Medicine 9(3), 255–256 (2002)
4. Oster, N., Nierenberg, R., Menlove, S., Chason, K., Pruden, J.: Reflections: September 11, 2001 – What We Learned. Academic Emergency Medicine 9(3), 216 (2002)
5. Frize, M., Trigg, H.C.E., Solven, F.G., Stevenson, M., Nickerson, B.G.: Decision-Support Systems Designed for Critical Care. AMIA Fall Meeting (1997)
6. Stephenson, R., Anderson, P.S.: Disasters and the information technology revolution. Disasters 21, 305–334 (1997)
7. Gaynor, M., Seltzer, M., Moulton, S., Freedman, J.: A Dynamic, Data-Driven, Decision Support System for Emergency Medical Services. In: Sunderam, V.S., van Albada, G.D., Sloot, P.M.A., Dongarra, J.J. (eds.) ICCS 2005. LNCS, vol. 3515, pp. 703–711. Springer, Heidelberg (2005)
8. Guterman, J.J., Mankovich, N.J., Hiller, J.: Assessing the effectiveness of a computer-based decision support system for emergency department triage. IEEE (1993)
9. Michalowski, W., Rubin, S., Slowinski, R., Wilk, S.: Mobile Clinical Support System for Pediatric Emergencies. Decision Support Systems 36, 161–176 (2003)
10. Yen, J., Yin, J., Ioerger, T.R., Miller, M., Xu, D., Volz, R.A.: CAST: Collaborative Agents for Simulating Teamwork. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence, pp. 1135–1142 (2001)
11. Klein, G.A.: Recognition-primed decisions. In: Advances in man-machine systems research, vol. 5, pp. 47–92. JAI Press, W. B. Rouse. Greenwich, CT (1989)
12. Cannon-Bowers, J.A., Salas, E., Converse, S.A.: Cognitive psychology and team training: Training shared mental models and complex systems. Human Factors Society Bulletin 33, 1–4 (1990)
13. McNeese, M.D., Bains, P., Brewer, I., Brown, C., Connors, E.S., Jefferson, T., et al.: The NeoCITIES simulation: Understanding the design and experimental methodology used to develop a team emergency management simulation. In: Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society, pp. 591–594 (2005)

Knowledge-Based Modeling and Simulation of Diseases with Highly Differentiated Clinical Manifestations

Marjorie McShane[1], Sergei Nirenburg[1], Stephen Beale[1],
Bruce Jarrell[2], and George Fantry[2]

[1] University of Maryland Baltimore County,
1000 Hilltop Circle, Baltimore Maryland, 21250 USA
{marge,sergei,sbeale}@cs.umbc.edu
[2] University of Maryland School of Medicine,
655 West Baltimore Street Baltimore, Maryland 21201-1559, USA
BJarrell@som.umaryland.edu,
GFantry@medicine.umaryland.edu

Abstract. This paper presents the cognitive model of gastroesophageal reflux disease (GERD) developed for the Maryland Virtual Patient simulation and mentoring environment. GERD represents a class of diseases that have a large number of clinical manifestations. Our model at once manages that complexity while offering robust automatic function in response to open-ended user actions. This ontologically grounded model is largely based on script-oriented representations of causal chains reflecting the actual physiological processes in virtual patients. A detailed description of the GERD model is presented along with a high-level description of the environment for which it was developed.

Keywords: cognitive model, simulation, gastroesophageal reflux disease, virtual patient.

1 The Maryland Virtual Patient Environment

The Maryland Virtual Patient[1] (MVP) project is developing an agent-oriented environment for automating certain facets of medical education and certification. This environment is effectively a network of human and software agents, at whose core is a virtual patient – a knowledge-based model of a person with a disease. This model is implemented in a computer simulation. The virtual patient is a "double agent" that displays both physiological and cognitive function. Physiologically, it undergoes both normal and pathological processes in response to internal and external stimuli. Cognitively, it experiences symptoms, has lifestyle preferences, has memory (many of whose details fade with time), and communicates with the human user about its personal history and symptoms.

[1] Patent pending.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 34–43, 2007.

Other software agents in the MVP environment include consulting physicians, lab technicians and a virtual mentor (tutor). What makes virtual patient modeling feasible – considering that comprehensively modeling human physiology would be a boundless endeavor – is our goal-oriented approach: we are not trying to recreate the human organism in all its details, we are modeling it to the extent necessary to support its realistic autonomous functioning in applications aimed at training and testing the diagnostic and treatment skills of medical personnel. Trainees can use the MVP simulation environment to interview a virtual patient; order lab tests; receive the results of lab tests from technician agents; receive interpretations of lab tests from consulting physician agents; posit hypotheses, clinical diagnoses and definitive diagnoses; prescribe treatments; follow-up after those treatments to judge their efficacy; follow a patient's condition over an extended period of time, with the trainee having control over the speed of simulation (i.e., the clock); and, if desired, receive mentoring from the automatic mentor.

The virtual patient simulation is grounded in an ontologically-defined model of human anatomy and physiology. Disease processes and treatments are modeled with a clear separation between direct and indirect effects. For example, if a trainee performs a Heller myotomy (a surgical procedure that cuts the lower esophageal sphincter (LES)) on *any* patient, whether or not his condition suggests the need for one, his basal LES pressure will decrease substantially, in most cases permitting an excessive amount of acid reflux per day and, over time, giving him gastroesophageal reflux disease (GERD).[2] To emphasize, there is no rule that says that a Heller myotomy turns a virtual patient into a GERD patient: the anatomical result of a Heller myotomy sets off a chain of events that turns most virtual patients into GERD patients. This level of anatomical and physiological automaticity permits the virtual patient to respond realistically even to completely unexpected and clinically incorrect interventions by the trainee. If the trainee launches such interventions, he must subsequently manage a more complex patient for the duration of the simulation.

Instances of virtual patients with particular diseases and particular physiological peculiarities are generated from core ontological knowledge about human physiology and anatomy by grafting a disease process onto a generic instance of human. Disease processes themselves are described as complex events in the underlying ontology.

This paper details the script-based modeling of diseases that are marked by distinctly different clinical manifestations – what we call *tracks*. We focus on the example of gastroesophageal reflux disease (GERD), which has six different tracks and a range of patient parameterization within each track. Disease modeling lies at the core of the MVP system, since the models must be both robust enough to support realistic function and constrained enough to be readily implemented. The models must also be generalizable enough to permit approaches and content modules to be reused in the modeling of new diseases. Our current disease models, which have been implemented and tested, fulfill all of these requirements.

[2] GERD would not arise if the Heller myotomy were incomplete or if the patient had achalasia, which could cause LES pressure to increase over time (see [1] for details of achalasia modeling).

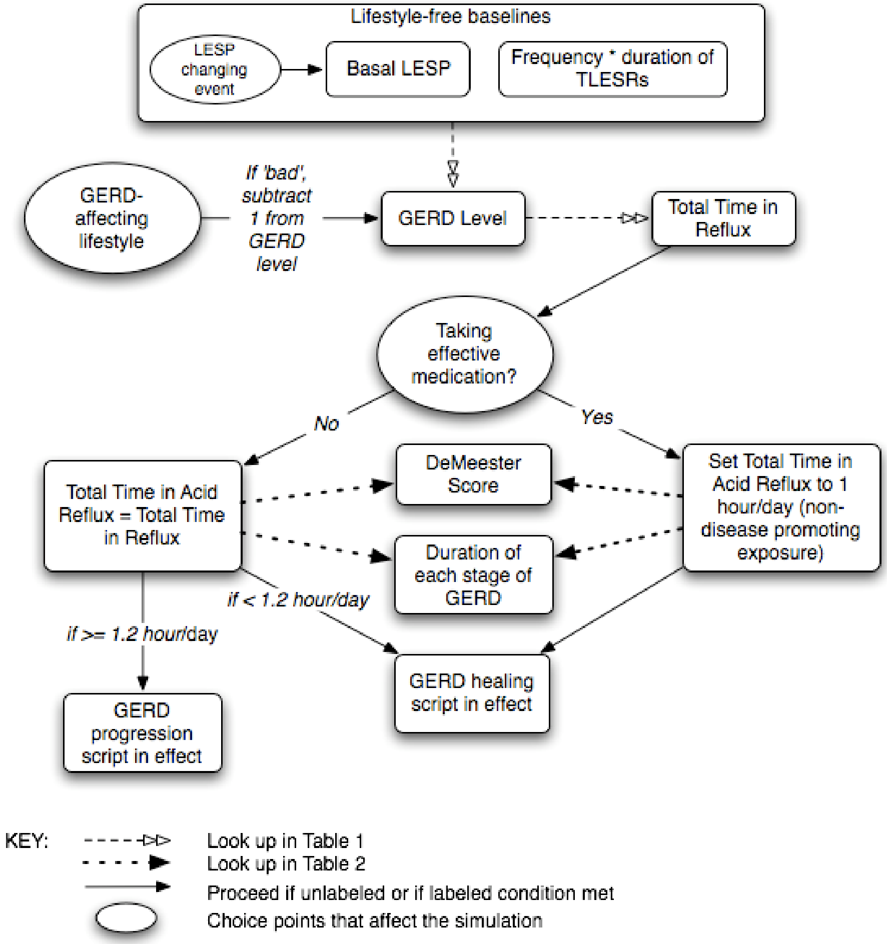

Lifestyle-free baselines
LESP changing event
Basal LESP
Frequency * duration of TLESRs
GERD-affecting lifestyle
If 'bad', subtract 1 from GERD level
GERD Level
Total Time in Reflux
Taking effective medication?
No
Yes
Total Time in Acid Reflux = Total Time in Reflux
DeMeester Score
Duration of each stage of GERD
Set Total Time in Acid Reflux to 1 hour/day (non-disease promoting exposure)
if >= 1.2 hour/day
if < 1.2 hour/day
GERD progression script in effect
GERD healing script in effect
KEY:
Look up in Table 1
Look up in Table 2
Proceed if unlabeled or if labeled condition met
Choice points that affect the simulation

Fig. 1. A representation of how GERD-related property values are set and modified, and how the simulator determines whether GERD is progressing or healing

2 Multi-track Scripts: The Example of GERD

The MVP system currently covers six esophageal diseases: gastroesophageal reflux disease (GERD), laryngopharyngeal extraesophageal reflux disease (LERD), LERD-GERD (a combination of LERD and GERD), scleroderma esophagus, Zenker's diverticulum and achalasia. In this section we describe the knowledge-based model of one of these diseases, GERD, which can be defined as any symptomatic clinical condition that results from the reflux of stomach or duodenal contents into the esophagus. The two sources of GERD are abnormally low pressure of the lower esophageal sphincter (LES) (< 10 mmHg), or an abnormally large number or duration of transient relaxations of the LES (TLESRs), both of which result in increased acid exposure of the esophageal lining. Both basal LES pressure (LESP) and TLESRs can be negatively affected by lifestyle habits such as consuming caffeine, chocolate or fatty foods.

A person can become a GERD patient in four ways: (a) by having an inherent predisposition to low LESP or excessive TLESRs; (b) by engaging in lifestyle habits that negatively impact LESP or TLESRs; (c) as a complication of another disease, like scleroderma esophagus, which decreases LESP; or (d) as a complication of an outside event or intervention, like a Heller myotomy, which results in a hypotensive LES.

The top levels of Figure 1 show the factors contributing to the inception of GERD. The severity of the GERD-producing factors is reflected by the attribute “GERD level”, which was introduced to unify the model, abstracting away from which specific LES-related abnormality gave rise to the disease. The lower the GERD level, the higher the daily esophageal acid exposure and the more fast-progressing the disease. The reason for associating a low GERD level with severe GERD is mnemonic: the GERD levels are the same as the basal LESP for patients who have low-pressure GERD. For example, a patient with a LESP of 1 mmHg will have a GERD level of 1. If a patient has a GERD level of 1 due to TLESRs, that means his daily esophageal acid exposure from the transient relaxations is the same as it would have been if he had had a basal LESP of 1.

Using GERD level as the anchor for modeling provides a simple mechanism for incorporating a patient’s lifestyle habits into the simulation: whenever he is engaging in bad lifestyle habits (assuming he has GERD-related sensitivities to those habits), his GERD level decreases by 1. For patients with a baseline GERD level of 10 – which is not a disease state – this means that engaging in bad habits is sufficient to initiate GERD and discontinuing them is sufficient to promote healing without the need for medication. For patients with a baseline GERD level of less than 10, lifestyle improvements can slow disease progression but not achieve the healing of previous esophageal damage. Once a patient’s GERD level is established, total time in reflux can be determined from Table 1.

Table 1. Property-value correlations for GERD, Part I (an excerpt)

LESP or the equivalent number of TLESRs per day	GERD level	Total Time in Reflux (in hours and percentage of time per day)
10-15	10	1.2 hrs. {5%}
9	9	1.56 hrs. {6.5%}
8	8	1.92 hrs. {8 %}
…	…	…
1	1	5.28 hrs.{22%}
0	0	6.0 hrs. {25%}

However, it is not total time in reflux that actually causes GERD, it is total time in *acid* reflux.[3] The acidity of the reflux can be decreased (i.e., pH can be increased) by taking medication. If effective medication is taken, the patient’s acid exposure is set

[3] New impedance testing permits the detection of GERD due to non-acid reflux, with this rare condition not currently being covered by the MVP system.

to 1 hour per day, which puts the patient in a healing state.[4] If no effective medication is being taken, then total time in acid reflux equals total time in reflux. Total time in acid reflux determines the rate of disease progression (i.e., duration of each conceptually delineated stage of the disease), as well as the patient's DeMeester score[5], as shown in Table 2. If the patient shows incomplete compliance, the simulator can switch him from a healing state to a disease state with great frequency, reflecting every instance of a taken or missed dose.

Table 2. Property-value correlations for GERD, Part II (an excerpt)

GERD level (for orientation)	Total Time in Acid Reflux	Stage Duration	DeMeester Score
10	<= 1.2 hrs. {5%}	na	10
9	1.56 hrs. {6.5%}	180 days	20
8	1.92 hrs. {8 %}	160 days	25
…	…	…	…
1	5.28 hrs. {22%}	40 days	80
0	6.0 hrs. {25%}	30 days	120

GERD can follow any of six clinical manifestations, or tracks:

1. non-erosive GERD, for which inflammation of the esophageal lining is the ending point of the disease
2. GERD with erosive esophagitis leading to erosion(s) but not ulcer(s)
3. GERD with erosive esophagitis leading to ulcer(s)
4. GERD with erosive esophagitis leading to peptic stricture
5. GERD with Barrett's metaplasia
6. GERD progressing past Barrett's metaplasia to adenocarcinoma

Which of these tracks a patient's disease will follow is determined by inherent predispositions. This means that no matter how long a track 1 patient experiences GERD, the disease will never progress past the level of inflammation, whereas if a track 4 patient remains untreated long enough, he will get a peptic stricture. Having a predisposition to later stages of GERD does not, however, necessitate experiencing those complications: ongoing effective treatment can reverse the disease course and ensure esophageal health indefinitely.

All GERD patients are assigned a set of GERD-related predispositions which determine how many and which stages of the disease they will experience. These predispositions can be asserted by patient authors (see below for patient authoring) or can be automatically set by the simulator for cases in which GERD arises spontaneously due to complications of a different disease or intervention.

Disease progression is modeled as changes over time of (a) physiological and symptom-related property values, and (b) the frequency, intensity, etc., of various

[4] It was decided that no pedagogical benefit would be gained by making the total time in acid reflux while on medication a variable: the important point is that the acid exposure is low enough to permit healing.

[5] DeMeester score is a complex reckoning of many factors measured during pH monitoring.

simulated events, like regurgitation. In addition, new objects (such as ulcers or tumors) can be created, modified and destroyed over time.

The simulation is driven by causal chains when they are known and are deemed useful to the goals of the simulation; otherwise, temporally oriented "bridges" reflecting clinical knowledge drive the simulation (e.g., if the disease has reached day 240 of untreated progression, the value of property *x* will be *y*).

2.1 Example: GERD with Erosive Esophagitis Leading to Erosion(s)

We illustrate the progression of GERD using the disease track "GERD with erosive esophagitis leading to erosion(s)", which is conceptually divided into three stages:

1. **preclinical GERD**, during which the value of the property *preclinical irritation percentage* (whose domain is *mucosa of distal esophagus*) increases from 0 to 100. When the *preclinical irritation percentage* reaches 100, the script for the preclinical stage is unasserted, with the simultaneous assertion of the script for
2. **the inflammation stage of GERD**, during which the mucosal layer of the esophageal lining is eroded, going from a depth of 1 mm. to 0 mm. over the duration of the stage. When mucosal depth reaches 0 mm., the script for the inflammation stage is unasserted, with the simultaneous assertion of the script for
3. **the erosion stage of GERD**, at the start of which an erosion object is created whose depth increases from .0001 mm. upon instantiation to .5 mm. by the end of the stage, resulting in a decrease in submucosal depth from 3 mm. to 2.5 mm. When submucosal depth has reached 2.5 mm. the script remains in a holding pattern since this patient does not have a predisposition to ulcer.[6]

Over the course of each stage, property values are interpolated using a linear function (though other functions could, in fact, be used). So halfway through the preclinical stage the patient's "irritation percentage" will be 50, and ¾ of the way through that stage it will be 75.

The length of each stage depends upon the patient's total time in acid reflux (cf. Table 2): e.g., for a patient with a total time in acid reflux of 1.92 hours a day, each stage will last 160 days. Such a patient will either have a baseline GERD level of 8 and not be engaging in bad habits, or have a baseline GERD level of 9 and be engaging in bad habits, with the bad habits "demoting" him to an effective GERD level of 8 (cf. Table 1 and Figure 1).

The experiencing of symptoms varies widely across patients but a fixed inventory of symptoms is associated with each disease, and expected ranges of values for each symptom can be asserted for each stage. For symptoms with abstract values ("On a scale of 1 to 10…"), we use the scale {0,1}, with decimals indicating intermediate values.

[6] Had the patient had a predisposition to ulcer, reaching a submucosal depth of < 2.5 mm. would have unasserted the erosion stage script and asserted the ulcer stage script, leading to the creation of an ulcer object and its increase in size over time.

Table 3 shows the symptom profile table for “GERD with erosive esophagitis leading to erosion(s)”. The ranges shown indicate the possible values for each symptom by the end of the given stage, with default values being shown in parentheses. For example, the patient Beatrice Thompson might experience heartburn severity at the average (default) level of .4 by the end of the inflammation stage, but might experience it at a higher than average level, .8, by the end of the erosion stage. All values between these “stage end” points are interpolated using a linear function. No symptoms are experienced in the preclinical stage, reflecting the definition of a preclinical disease.

Table 3. Symptom profile table for one track of GERD

Stage	Preclinical	Inflammation	Erosion
Heartburn freq. (times per day)	0	3 – 5 (4)	6 – 8 (7)
Heartburn severity	0	.3 – .5 (.4)	.6 – .8 (.7)
Regurgitation (times per week)	0	3 – 5 (4)	6 – 8 (7)
Symptom correlation	0	0 – 1 (random .5 < > .8)	0 – 1 (random .5 < > .8)

2.2 Interactions and Interventions

What we have shown so far is the knowledge needed to simulate GERD, assuming no external interventions. However, external interventions are key to an educationally-oriented simulation, as are realistic patient responses to interventions, be they clinically appropriate (expected) or clinically inappropriate (unexpected).

The MVP system supports verbal interaction with the virtual patient as well as two types of interventions: diagnostic tests and treatments. Questioning the patient and carrying out diagnostic tests are similar in that the system must (a) interpret the question or the request for a test, (b) look up the appropriate physiological or symptom-related property values or stored events in the database populated during the simulation, (c) return a response as an English string. Currently, since we have not yet plugged in natural language processing (NLP) capabilities (see below), steps (a) and (c) are handled by menus and preconstructed strings, respectively. However, once NLP is incorporated, these steps will rely on the language understanding and generation capabilities of the virtual patient, lab technicians and outside specialists.

Typical tests carried out when GERD is suspected are:

- esophagogastroduodenoscopy (EGD), which returns information about the presence of inflammation, erosion, ulcer, etc.
- pH monitoring, which returns time in acid reflux, symptom correlation (the correlation between a patient’s symptoms and the acidity of the distal esophagus), and DeMeester score (cf. footnote 5)
- barium swallow, which can be used to detect tumors.

All positive and pertinent negative results are returned. These tests contribute to a definitive diagnosis of GERD, although a clinical diagnosis can be made based on the efficacy of drugs in reducing symptoms. If tests that are atypical for GERD are launched on a GERD patient, the relevant property value(s) will be sought and, if no abnormal values are found, the return value will be normal. In this way, the MVP system is always prepared for unexpected moves by users.

As concerns treatment, the typical treatment options for GERD are lifestyle modifications in combination with H2 blockers or PPIs (QD or BID).[7] Lifestyle modifications can be completely effective only for patients with a baseline GERD level of 10 and bad habits that they succeed in overcoming. All other manifestations of GERD require medication, and medications can be effective or ineffective for different patients, as recorded in their physiological profiles.

Treatments administered in the MVP environment can improve the patient's condition, be ineffective, or be detrimental. For example, whereas PPI BID typically works for a patient with a GERD level of 5, H2 blockers typically do not work and a Heller myotomy will certainly be detrimental. Since Heller myotomy is not an expected treatment for a GERD patient, there is nothing in his profile to indicate how he will respond to this surgery; instead, the ontologically defined default outcome is used, which is reduction of the basal LESP to 2 mmHg.

If an effective treatment is administered, the GERD progression script for the current stage (e.g., “erosion”) is halted and the “heal GERD” script is launched, improving the physiological and symptom profiles over an appropriate amount of time depending upon the current stage of the disease. For example, healing that starts during the inflammation stage can take up to 4 weeks (depending on how far into the inflammation stage the patient has progressed), while healing that starts during the erosion stage takes up to 8 weeks.

It should be clear that many aspects of GERD are inherent parts of the disease model that cannot be modified by patient authors. For example, a patient with a GERD level of 2 will have a faster-progressing disease than a patient with a GERD level of 9, assuming no effective treatment. The correlation between the amount of acid exposure and the rate of disease progression was considered a fundamental aspect of the disease by the physicians involved in building the model, who saw no practical or pedagogical reason to permit patient authors to override the set correlations between GERD level and stage duration.[8] By contrast, there are practical and pedagogical reasons to permit different patients to display other kinds of differences provided for in the patient authoring process.

2.3 Patient Authoring

Patient authoring involves filling in a one-page on-line questionnaire that permits domain experts or teachers to create a population of patients that display clinically relevant variations on the disease theme. For GERD, the population of patients must include patients with predispositions to each of the six disease tracks as well as different GERD levels, sources of GERD (low LESP vs. TLESRs), food sensitivities,

[7] We have not yet included reflux surgery in the system's repertoire. See [1] for a description of achalasia, whose inventory of treatment options and their possible outcomes is more complex.

[8] For other diseases, such as achalasia, stage correlation can be set explicitly by patient authors. This reflects the fact that the causal chains for such diseases are not as well understood, so there is no physiological variable from which to derive the differences in stage durations among patients that are clinically observed.

symptom profiles, reactions to medications, and compliance to treatment. Once these choices are made, the simulation is fully prepared for interactive use.

Disease scripts can be considered the central axis of the MVP environment, since all interactive and educational functionalities depend upon the ability of the disease scripts to support a robust and realistic simulation. In fact, the emphasis in system development thus far has been on disease scripts, with automatic mentoring (described in [1]) being added just recently, and natural language support not yet integrated, although the latter is the forte of our group, which has spent the past twenty years working in the field of knowledge-based natural language processing.

3 Comparisons with Other Systems and Approaches

The benefits of simulation in medicine have been widely propounded.[9] To give just one example, Satish and Streufert [3] write: "We need to ensure that medical personnel have the factual content knowledge needed to respond to the task at hand, but we also need to make sure that they can respond to complex challenges by processing information optimally. Simulations, if used as part of an appropriate training system, provide an optimal opportunity to acquire both."

Although simulation using live actors has long been a part of medical training, computer-based simulation that provides sufficient authenticity is still in its infancy, particularly as regards simulation of decision-making tasks via cognitive modeling (as contrasted with mannequins trainers for the development of motor skills). As many quality overviews of simulation already exist in the literature (e.g., [4], [5]), we limit our comparisons to two systems that are directly comparable to the MVP along at least one parameter.

CIRCSIM-Tutor [5] is a system dedicated to training medical students about the baroreceptor reflex. Over the course of the system's history, it has shifted from being a dynamic mathematical model that could be interacted with by students but provided no tutoring (MacMan), to being a tutoring system that no longer relies on the mathematical model, instead using a set number of statically stored cases. In terms of overall pedagogical goals – expediting medical education through automatic mentoring using natural language – the CIRCIM and MVP systems are similar. However, whereas the CIRCSIM group moved away from using a live physiological model, the MVP group is committed to developing a robust physiological model that can permit many types of interaction by trainees. For example, whereas one user might prefer to enable the tutor immediately in order to more quickly learn to avoid mistakes, another might prefer to learn by trial and error. Both methods being equally supported by the interactive simulation.

Another striking difference between CIRCSIM and MVP is the approach to NLP. Whereas CIRCISM essentially operates in terms of strings, MVP will carry out the knowledge-rich text processing developed in the theory of Ontological Semantics [6] and implemented in the OntoSem system. Using OntoSem, text strings are automatically translated into interpreted ontological concepts that are combined in logically correct ways to form text-meaning representations. It is over these

[9] Previously we have reported on the potential of this work to further medical pedagogy [2].

text-meaning representations that the language-based reasoners operate. Notably, the same ontological substrate is used both for medical modeling/simulation and for language processing, leading to a highly integrated knowledge base for the MVP system.

A knowledge-based approach to modeling virtual patients to support recertification has been reported in [7,8,9]. The authors describe both the modeling of the patient simulation process and the task of creating knowledge to support such a system. They also consider the situation of disease co-occurrence, which the MVP system also includes (e.g., scleroderma esophagus can have GERD as a side effect, with both continuing their courses simultaneously). One of the many major distinctions between this approach and our project is that Sumner *et al.* create probabilistic models using Bayesian networks and modified Monte Carlo methods, while our approach stresses causal modeling and guided creation (authoring) of virtual patients.

Initial testing of the MVP system was carried out over six hours with 40 second and third year medical students from the University of Maryland School of Medicine. Among the salient observations were that students followed a pattern of evaluation and management that is parallel to the actual process used in clinical care, they managed patients as if their actions had real consequences, and they realized the importance of "observation over time" as a diagnostic maneuver. More extensive evaluation is planned for the near future.

References

1. McShane, M., Jarrell, B., Nirenburg, S., Fantry, G., Beale, S.: Training clinical decision making using cognitively modeled virtual patients (Forthcoming)
2. Jarrell, B., Nirenburg, S., McShane, M., Fantry, G., Beale, S., Mallott, D., Raczek, J.: An interactive, cognitive simulation of gastroesophageal reflux disease. In: Proceedings of medicine meets virtual reality 15 in vivo, in vitro, in silico: Designing the next in medicine, February 6-9, 2007, Long Beach, California (2007)
3. Satish, U., Streufert, S.: Value of a cognitive simulation in medicine: towards optimizing decision making performance of healthcare personnel. Qual. Saf. Health Care 11, 163–167 (2002)
4. Streufert, S., Satish, U., Barach, P.: Improving medical care: The use of simulation technology. Simulation & Gaming 32(2), 164–174 (2001)
5. Evens, M., Michael, J.: One-on-One Tutoring by Humans and Computers. Lawrence Erlbaum and Associates, Publishers, New Jersey and London (2006)
6. Nirenburg, S., Raskin, V.: Ontological Semantics. MIT Press, Cambridge, Mass (2004)
7. Sumner, W., Marek, V.W., Truszczynski, M.: Data transformations for patient simulations. In: Proceedings of the 19th annual symposium for computer applications in medical care, p. 970 (1995)
8. Sumner, W., Marek, V.W., Truszczynski, M.: A formal model of family medicine. Journal of american board of family practice 9, 41–52 (1996)
9. Marek, V.W., Sumner, W., Truszczynski, M.: Creating evolution scenarios for hybrid systems. In: Proceedings of IEEE-SMC Symposium on Control, Optimization and Supervision, at CESA96, Lille, pp. 512–516 (1996)

Co-operative Agents in Analysis and Interpretation of Intracerebral EEG Activity: Application to Epilepsy

Mamadou Ndiaye, Abel Kinie, and Jean-Jacques Montois

Laboratoire Traitement du Signal et de l'Image, INSERM U642 antenne de Saint-Malo,
IUT de Saint Malo, Rue Croix Désilles BP 195
35409 Saint Malo, France
{Mamadou.Ndiaye,Jean-Jacques.Montois,Abel.kinie}@univ-rennes1.fr

Abstract. The paper presents a distributed approach for the interpretation of epileptic signals based on a dynamical vectorial analysis method. The approach associates signal processing methods into a situated, reactive, cooperative and decentralized implementation. The objective is to identify and locate the various interictal and ictal epileptiform events (pathological and/or normal) contained in intracerebral EEG signals (one hundred recording channels in general) recorded in patients suffering from partial temporal lobe epilepsy. This approach associates some signal processing methods (spectral analysis, causality measurements, detection, classification) in a multi-agent system.

Keywords: Epilepsy, signal processing, agents and cooperative systems.

1 Introduction

Epilepsy is a disease characterized by brutal and excessive synchronization of neuronal population. In Stereo-Electroencephalography (SEEG) exploration, we have physiological signals rich of informations on the observed structures. These signals inform various peculiarities of the functioning of a structural entity, an organ or even a system. The human intracerebral EEG can participate in a joint; hierarchical and a reproducible way in cognitive spots or in pathological processes. It is also allowable to admit that some local dynamics taking part in a joint action of structures are different between themselves. Moreover, previous studies were able to show that certain epileptic seizures originate from an area of the brain and propagate in a diffuse way towards other cerebral structures [1]. It is about a multi-dimensional system, multi-variable, split up in a multitude of independent neural entities in mutual influence. The signal processing proposes approaches and answers today to better define the complex concepts of irritative zone and epileptogenic zone [2]. The concepts of topography ("*where is the source of the signal?* ") and of synchrony ("*are these two signals synchronous, thus reflecting a functional connectivity?*") are now clearly set up [3]. However the answers contributed by these methods is based on series of treatments and correlations applied to a small number of signals (selected visually sometimes) among the great number of signals available during the recording, a number which can reach 128. Our work asks the question of the network

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 44–48, 2007.

dynamics in terms of a succession of behaviour produced by inter-structures interactions. The problem is approached by a multi-agents system (MAS) [4], the agents answer partially the resolution of problem by acting locally; making the various computations (signal processing algorithms) according to some control mechanisms (strategies, organization, cooperation and coordination of agents).

The paragraph §2 clarifies the formal frame of our work. It explains the "agentification" of the problem. The third point presents the agents architecture, quantitative extraction information methods used by agents and various phases of the methodology. The experimental results and discussion are exposed in the fourth point.

2 Agentification for the Analysis of Epileptic Signals

We start from the idea that we have to deal with a set of neuronal groups being able to present, on a given temporal window, paroxystic activities of various types and interactions more or less pronounced leading to behaviour of these groups. We try to *approach them locally to be able to explain them globally*. Our work consists in studying a vectorial signal $S(t) = [S_1(t) \dots S_M(t)]$ observed on an interval [0, T]. In other words, it consists : *(i) to characterize each EEG channel $S_i(t)$*, *(ii) to determine the statistical relations between the various channels*, *(iii) to study the organization of group according to the analysis of the signals*, and *(iv) to connect the notion of functional coupling between cerebral structures towards the quantities supplied by signal processing methods supervised by the dedicated multi-agent system (MAS).*

We so associated to each signal $S_i(t)$, a "signal agent" noted A_{S_i} and to each group of "signals agents" of the same cerebral area, a "structure agent" noted S_{P_i} (figure 1).

The collective intelligence, control mechanisms, coordination and signal processing algorithms are spatially distributed in the various constituents of the system. Each structure is associated to an agent, inside whom are implemented local computations processes ("signals agents") and of interpretation ("structures agents").

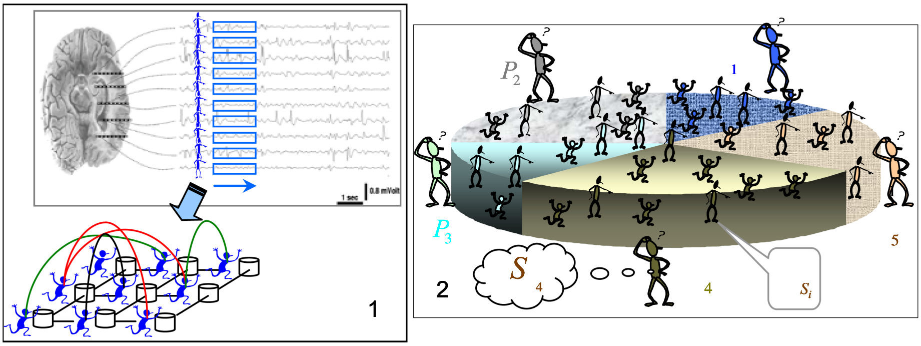


Fig. 1. 1) Vectorial analysis of epileptic signals: each agent has its own choices, knowledge, and algorithms processes. 2) Partitioning: the "signal agent" are supported in groups of agents belonging to the same cerebral area and a "structure agent" is associated to each partition.

3 Multi-Agent System Proposed

The multi-agent system (MAS) is implemented in "Madkit" platform (Multi-agent development kit [6]). Our choice concerned hybrid architecture distributed on two levels; a first level consisted of reactive agents ("signals agents") which try to locate the interesting signals specified by an auto-organized model and the second level established by cognitive agents ("structures agents") which interpret the results realized by the reactive agents. Control's agents complete the system for needs of software treatment ("Scheduler agent", "Server agent" and "Observer agent"). The interactions between signal processing algorithms and MAS are offered in 4 stages which blend the general architecture philosophy.

3.1 Quantitative Extraction Information of Each Agent

This first stage consists in extracting the individual properties of each signal agent. The frequency activities is characterize, in 3 phases on a slippery window : *(i) evaluation of the power spectral density of the signal $S_i(t)$ in defined frequency band, (ii) construction of a characteristic vector for each signal $S_i(t)$ and iii) numeric coding of the vector in a scalar indicator*. Each agent, according its activity, is associated to a scalar indicator and a *characteristic* vector. The signals agents are then grouped by similarity according to 4 classifications methods: *(c_i) Supervised classification using learning classes defined a priori. (c_{ii}) Supervised classification using a Fuzzy approach classification algorithm. (c_{iii}) Pseudo- supervised classification based on the scalar indicator. (c_{iv}) Unsupervised classification, based on euclidian distance between characteristic vectors associated to the agents.*

3.2 Selection of the "Signals Agents"

The selection results from three agent's behaviours grouping-diffusion-decimation. Each signal agent has two local data bases "localGroup" and "localOutGroup" where it registers respectively its remarkable similarity (affinity) and its remarkable differences (rejection). The *grouping behaviour* results from the necessity of grouping together similar signals agents of the same group. Signal agent can be also brought to spread its group by diffusion its local bases according to the criterion "*I accept in my base each agent similar to one of my affinities if he is not already registered in my rejections*" *and* "*I refuse to integrate a group where is registered one of my rejections*". This *diffusion* process takes place at first in a local context (inside a partition). Each group of similar agents so formed is considered as a unique entity carrying the same information from where comes the idea of a sole representative by group. This *decimation behaviour* is activated in each partition (structure agent) to elect the best representatives of the structure.

3.3 Characterization of Connectivity Between Structures

The links between structures are characterized by three measures: similarity, statistical relations and synchronizations between agents. The *similarity* measure is based on a euclidian distance between characteristic vectors. Each leader of group has to evaluate its links with the others leaders. The leaders use on their local data bases to retain their affinities, to reject their rejections and to ignore the elements of which similarities is to be verified by other methods. The s*tatistical relations* measure uses the nonlinear intercorrelation coefficient [5] to characterize the degrees of links between investigated structures. This measure made on local entities defines the links between the bows of the graph produced by the system at the global level. The *synchronization* is based on significant modifications of the intrinsic properties in each signal. The "Observer" agent leans on the temporal synchronization of the alerts received to define the links of synchronizations between agents.

3.4 Characterization of Spatio-temporal Dynamic

The characterization of the spatiotemporal dynamics looks for a global interpretation based on the local modifications of connectivity between structures. Each agent structures pronounces on its links with the other structures all the time and on its possible modifications of connectivity with regard to the previous moments. The combination of all the propositions, by the agent "Observer", produces a graph at the global level. This graph follows the spatiotemporal dynamics of the analyzed seizure. The observation of the modifications between successive graphs (distances between graphs) allows characterizing the organizational dynamics modifications.

4 Results

The data used in this study result from 5 patients suffering from lobe temporal epilepsy and candidates for surgical treatment. The system produces for each patient a layer spatio-temporo-spectral coloured and a graph symbolizing the various connectivity

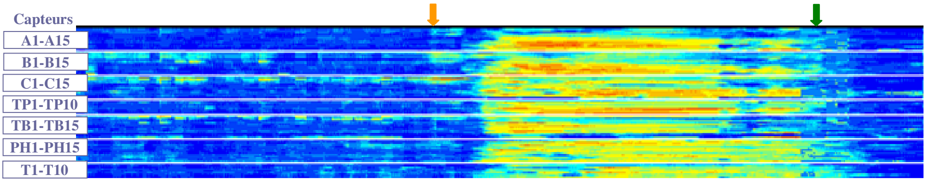


Fig. 2. Spatio-temporo-spectral tablecloths of P1. The "cold" and "warm" colours coding respectively the "low frequency" and the "high frequency" activities. The contacts of electrodes are numbered from 1 to 15 since the internal extremity until the external extremity.

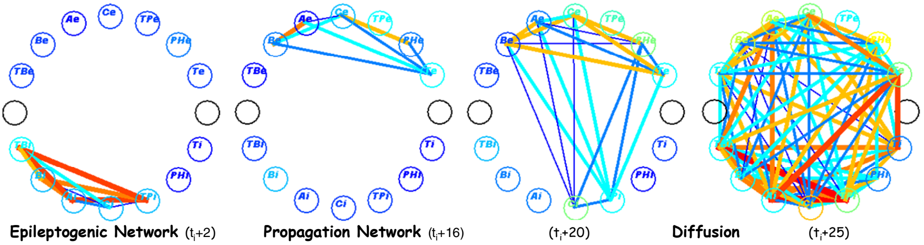


Fig. 3. Graphs of the couplings between structures for P1. The "structures agents" are knots and links between "signals agents" symbolize bows.

of the network formed by the various cerebral structures involved in the epileptic processes. The following paragraphs analyze the results at the first patient.

The analysis of the layer spatio-temporo-spectral (figure 2), coupled with that of the graphs (figure 3) produced by the system lets appear a network organization in the seizure of this patient. This organization includes the hippocampus (Bi and Ci), the internal temporal pole (TPi) and the entorhinal cerebral cortex (TBi). The analysis of P1 puts in evidence the much localized critical activities and a very net implication of a big number of canals SEEG in the distribution of the paroxystic discharge. A systematically premature role of the Hippocampus and the entorhinal cerebral cortex is observed at this patient. A spontaneous synchronization appears at the moment t_i+2 (figure 4) between various internal structures of the temporal lobe (beginnings of the epileptogenic network). This synchronization propagates secondarily in the external structures (temporo-basal cerebral cortex (TBe), gyrus T2 (Be, Ae) and in the external temporal pole)) and becomes widespread in the other investigated structures.

5 Conclusion

We associated in our platform approved signal processing algorithms and cooperative MAS. *A*gents identify the activities contained in each EEG signal, to select the interesting signals and to present the spatiotemporal dynamics of these activities during the seizure. The analysis of this dynamics loosens spatio-temporo-spectral layers and statistical couplings graphs between groups of signals also follow the organization of epileptogenic network. Our approach loosens the progressive involvement of mutual interactions between the cerebral regions. Channels mainly engaged in the release of the process of seizure are clearly made known and channels secondarily implied are also indeed referred.

References

1. Bartolomei, F., et al.: Seizures of temporal lobe epilepsy: identification of subtypes by coherence analysis using SEEG. Clin Neurophysiol 110, 1714–1754 (1999)
2. Bancaud, J.: Stereoelectroencephalography. In: Remond, A. (ed.) Handbook of electroencephalography and clinical neurophysiology, vol. V.10(B), pp. 3–65. Elsevier, Amsterdam (1975)
3. Gotman, J.: L'analyse de l'EEG de Berger à nos jours. Epileptic DisordersNuméro 3 3, 7–10 (2001)
4. Ferber, J.: Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence. Addison-Wesley, London (1999)
5. Bartolomei, F., Chauvel, P., Wendling,: Dynamique des réseaux neuraux dans les épilepsies partielles humaines. Revue Neurologique 161, 767–780 (2005)
6. Gutknecht, O., Ferber, J.: The madkit Agent platform architecture. In: 1st Workshop on Infrastructure for Scalable Multi-Agent Systems (2000)

An Ontology-Driven Agent-Based Clinical Guideline Execution Engine

David Isern, David Sánchez, and Antonio Moreno

ITAKA Research Group - Intelligent Tech. for Advanced Knowledge Acquisition
Dept. of Computer Science and Mathematics. University Rovira i Virgili*
43007 Tarragona, Catalonia (Spain)
{david.isern,david.sanchez,antonio.moreno}@urv.cat

Abstract. One of the hardest tasks in any healthcare application is the management of knowledge. Organisational information as well as medical concepts should be represented in an appropriate way in order to improve interoperability among existing systems, to allow the implementation of knowledge-based intelligent systems, or to provide high level support to healthcare professionals. This paper proposes the inclusion of an especially designed ontology into an agent-based medical platform called HeCaSe2. The ontology has been constructed as an external resource, allowing agents to coordinate complex activities defined in any clinical guideline.

1 Introduction

In order to exploit the great potential that clinical practice guidelines (GLs) offer to improve patient's care delivery quality, tools for adopting them within the clinical routine are required [1]. In this sense, one of the biggest problems is the gap between the codification of these GLs and their use/interpretation in a real organisation. One of the solutions consists on adding content background to the GLs by the use of medical *ontologies*.

The use of ontologies in medicine supposes an important advantage in order to provide a common understandable framework to make explicit the involved medical concepts as well as their relations and properties. Ontologies also provide a high level model of the daily work flow that can be adapted to the particular circumstances of any healthcare organisation. Kumar et. al. [2] studied the implementation of a task ontology named Context-Task Ontology (CTO) to map the knowledge required in the implementation of GLs. They noted that this approach had some drawbacks, such as the difficulty to define and know exactly which relations are needed, as well as the requirement of expert's intervention. The same authors later described the use of ontologies to define clinical guidelines by adding a hierarchy of classes to represent medical procedures and plans

* The work has been partially supported by the K4CARE project (IST-2004-026968). D. Sánchez would like to acknowledge the support of *Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya i del Fons Social Europeu.*

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 49–53, 2007.

[3]. However, this implied a high level of complexity as compared to flow-chart-based representations. Serban et al. [4] proposed the use of an ontology to guide the extraction of medical patterns contained in GLs in order to reconstruct the captured control knowledge. All these works suggest the use of UMLS as a central corpus. Ciccarese et al. [5] introduced an architecture that linked a care flow management system and a guideline management system by sharing all the data and ontologies in a common layer. They proposed to represent medical and organisational information in those ontologies, but they did not use non taxonomic relations in the ontologies. Moreover, Davis and Blanco [6] suggested the use of taxonomies to model the clinical life cycle knowledge. They also described a state-based data flow model to represent all dependencies between enterprise entities.

Previous papers ([7,8]) introduced an agent-based system called HeCaSe2 that proposes an open architecture that models different entities of a healthcare organisation. HeCaSe2 includes the management of clinical guidelines by doctors in the diagnosis or treatment of diseases. All these tasks require an external element to be more flexible and efficient: a representation of the care flow and the terminology used across all entities. In order to address this issue, this paper proposes the inclusion in HeCaSe2 of an application ontology that covers three main areas: *a*) representing all medical terminology used by all partners, *b*) modelling healthcare entities with its relations, and *c*) collecting semantic categories of those medical concepts.

2 HeCaSe2: A Distributed Guideline-Based Health Care System

HeCaSe2 is a dynamic multi-agent system that maps different entities in a healthcare organisation (*i.e.* medical centres, departments, services, doctors, patients) as agents with different roles. This system provides interesting services both to patients (*e.g.* booking a visit with a doctor, or looking up the medical record) and to doctors (e.g. support in the application of a GL to a patient). Guidelines are used to provide a high level supervision of the activities to be carried out to address a specific pathology.

At the top, the patients are represented in the system by *User Agents* (UA). Any UA can talk with the *Broker Agent* (BA). The BA is the bridge between users and medical centres, and it is used to discover information about the system. All UAs can ask this agent in order to find medical centres satisfying certain criteria. The BA covers the medical centres located in a city or in an area. Any user can access the system through a *Medical Centre Agent* (MCA) that centralises and monitors the outsider's accesses. A MCA monitors all of its departments, represented by *Department Agents* (DAs), and a set of general services (represented by *Service Agents* (SAs)). Each department is formed by several doctors (represented by *Doctor Agents* (DRA)) and more specific services (also modelled as SAs). Moreover, in each department there is a *Guideline Agent* (GA) that performs all actions related to guidelines, such as looking for a desired

GL, storing and/or changing a GL made by a doctor, etc. This GA contains only GLs related to the department where it is located (the knowledge is close to the entity that will use it). Each department also contains an *Ontology Agent* (OA) that provides access to the designed medical ontology and complements the information provided by the GA. At the bottom of the architecture there is the *Medical Record Agent* (MRA) which stores all patient medical records in the medical centre.

3 Ontological Representation of Medical Knowledge

Ontologies define the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary [9]. From the available representation languages, the medical ontology has been coded using *OWL DL* ([10]) and constructed following the *101* methodology [11].

The designed ontology is composed by relations established among the agents associated to a healthcare organisation. It has three main groups of concepts: *a*) *agent-based health care concepts*, *b*) *semantic types* of the used concepts (entities and events), and *c*) *medical concepts* related to the managed GLs (see Fig. 1).

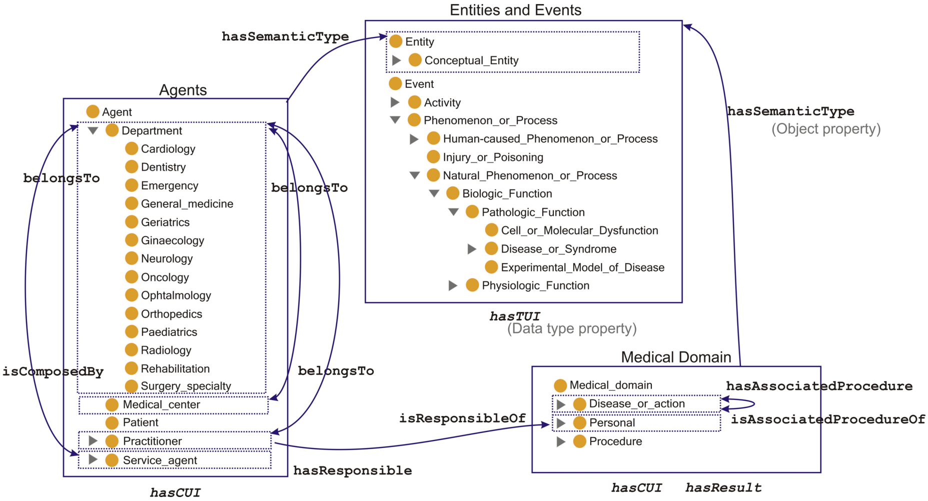


Fig. 1. Subset of the designed *Medical Ontology*

The first group of concepts concerns the multi-agent system. The *Agent* class encloses all main concepts and properties related with the internal organisation. That portion of the medical ontology is composed by *Medical centres*, *Departments*, *Patients*, *Practitioners* and *Services*.

All these elements have internal relations, such as *Oncology* `is-a` *Department* that `belongsTo` *Medical-center* which `isComposedBy` *Service_agents*. More complex relations between doctors and services are also mapped, such as *Nurse* `belongsTo` *Department* because a nurse can be located in any department, or *Family_doctor* `belongsTo` (*General medicine* ∪ *Emergency* ∪ *Paediatrics*) that means that an instance of family doctor could belong to any instance of three departments. Relations between *Agent* subclasses are inspired in the typical structure of healthcare organisations. The inverse relations are also available in order to know which kind of doctors compose a department or which kind of services are located in a department or medical centre.

Although most of the departments are similar in medical centres, it is possible to represent different variations. In those cases, a specialisation of the ontology could be made by creating subclasses. The parent class would keep all common features and the siblings would contain specific features or resources for each one.

The next set of concepts are the semantic types of medical terms according to its context. That portion of the ontology is intended to avoid language ambiguity, and the UMLS Semantic Network was used.

Currently, UMLS defines 135 different semantic types divided in two groups: meanings concerned with healthcare organisations or entities, and meanings related with events or activities in a daily care flow. Those hierarchies are named *Entity* and *Event* respectively, and are organised as a taxonomy with `is-a` relations between concepts, such as *Disease_or_Syndrome* `is-a` *Pathologic_function*.

All this information is used by agents to know exactly which is the function of any required concept and further connections with others. For instance, if a concept is a *Finding*, and a *Finding* `isResponsibilityOf` a *Practitioner*, the agent knows that a patient's finding should be given by a practitioner.

Finally, the last part of the ontology represents the specific vocabulary used in clinical guidelines. This part systematises all specific knowledge required in any guideline execution engine, divided in *Diseases*, *Procedures* and *Personal data*. It is necessary to define a set of relations between each concept and its identifier (*Code Unique Identifier* or CUI), its semantic type, which entity of the system is responsible of its accomplishment, and the produced result (*i.e.* if it as number, a Boolean, an enumerate or a complex object). The established relations are bidirectional because it is interesting to know that the finding *Active_cancer* `isResponsibilityOf` a *Family_Doctor*, and also the inverse relation is important for the family doctor in order to know his responsibilities. Each agent can access the concepts related to its own domain and be aware of the consequences of the application of an action.

4 Conclusions

The inclusion of a medical ontology in the HeCaSe2 multi-agent system has been discussed. As shown in the introduction, the use of ontologies in the medical domain is increasing and offers some advantages such as making domain assumptions explicit, separating domain knowledge from operational knowledge and sharing a consistent understanding of what information means.

In the present work, the designed medical ontology brings the following advantages to the guideline-based execution system: a) to identify the required actors that are able to accomplish an action and to know the source of a data, b) to adapt the execution framework to the particular casuistry of any healthcare organisation without modifying the MAS implementation, and c) to provide an application independent context. Thus, by changing the ontology and its relations, the execution procedure is changed.

Note that the only issue that should be addressed is the manual definition of the appropriate task ontology. This question usually requires the intervention of a domain expert but UMLS provides a large corpus of concepts and relations that can be reused.

References

1. Quaglini, S., Stefanelli, M., Cavallini, A., Micieli, G., Fassino, C., Mossa, C.: Guideline-based careflow systems. Artif Intell Med 20, 5–22 (2000)
2. Kumar, A., Quaglini, S., Stefanelli, M., Ciccarese, P., Caffi, E.: Modular representation of the guideline text: An approach for maintaining and updating the content of medical education. Medical Informatics and the Internet in Medicine 28, 99–115 (2003)
3. Kumar, A., Ciccarese, P., Smith, B., Piazza, M.: Context-Based Task Ontologies for Clinical Guidelines. In: Pisanelli, D.M. (ed.) Ontologies in Medicine. Studies in Health Technology and Informatics, vol. 102, pp. 81–94. IOS Press, Amsterdam (2004)
4. Serban, R., Teije, A.t., Harmelen, F.v., Marcos, M., Polo-Conde, C.: Extraction and use of linguistic patterns for modelling medical guidelines. AI in Medicine 39, 137–149 (2007)
5. Ciccarese, P., Caffi, E., Boiocchi, L., Quaglini, S., Stefanelli, M.: A Guideline Management System. In: Fieschi, M., Coiera, E., Li, Y.-C.J. (eds.) 11th World Congress on Medical Informatics, MedInfo 2004, San Francisco, USA. Studies in Health Technology and Informatics, vol. 107, pp. 28–32. IOS Press, Amsterdam (2004)
6. Davis, J.P., Blanco, R.: Analysis and Architecture of Clinical Workflow Systems using Agent-Oriented Lifecycle Models. In: Silverman, B.G., Jain, A., Ichalkaranje, A., Jain, L.C. (eds.) Intelligent Paradigms for Healthcare Enterprises. Studies in Fuzziness and Soft Computing, vol. 184, pp. 67–119. Springer, Heidelberg (2005)
7. Isern, D., Moreno, A.: Distributed guideline-based health care system. In: 4th International Conference on Intelligent Systems Design and Applications (ISDA-2004), Budapest, Hungary, pp. 145–150. IEEE Press, Orlando (2004)
8. Isern, D., Valls, A., Moreno, A.: Using aggregation operators to personalize agent-based medical services. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 1256–1263. Springer, Heidelberg (2006)
9. Neches, R., Richard, F., Finin, T., Gruber, T., Patil, R., Senator, T., Swartout, W.: Enabling Technology for Knowledge Sharing. AI Magazine 12, 36–56 (1991)
10. McGuinness, D., Harmelen, F.v.: OWL Web Ontology Language (2004)
11. Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. Technical Report SMI-2001-0880, Stanford Medical Informatics (2001)

Part II
Temporal Data Mining

An Intelligent Aide for Interpreting a Patient's Dialysis Data Set

Derek Sleeman[1], Nick Fluck[2], Elias Gyftodimos, Laura Moss, and Gordon Christie

Departments of Computing Science & Medicine,
The University of Aberdeen
Aberdeen AB24 3FX
Scotland UK

Abstract. Many machines used in the modern hospital settings offer real time physiological monitoring. Haemodialysis machines combine a therapeutic treatment system integrated with sophisticated monitoring equipment. A large array of parameters can be collected including cardiovascular measures such as heart rate and blood pressure together with treatment related data including relative blood volume, ultrafiltration rate and small molecule clearance. A small subset of this information is used by clinicians to monitor treatment and plan therapeutic strategies but it is not usually analysed in any detail. The focus of this paper is the analysis of data collected over a number of treatment sessions with a view to predicting patient physiological behaviour whilst on dialysis and correlating this with clinical characteristics of individual patients.

One of the commonest complications experienced by patients on dialysis is symptomatic hypotension. We have taken real time treatment data and outline a program of work which attempts to predict when hypotension is likely to occur, and which patients might be particularly prone to haemodynamic instability. This initial study has investigated: the rate of change of blood pressure versus rate of change of heart rate, rate of fluid removal, and rate of uraemic toxin clearance. We have used a variety of machine learning techniques (including hierarchical clustering, and Bayesian Network analysis algorithms). We have been able to detect from this dataset, 3 distinct groups which appear to be clinically meaningful. Furthermore we have investigated whether it is possible to predict changes in blood pressure in terms of other parameters with some encouraging results that merit further study.

1 Introduction

The human renal system is responsible for a number of different roles, including: Water Balance, Electrolyte Balance (e.g. sodium, potassium), Toxin Removal, Acid/Base Balance, and Blood Pressure Regulation, [Daugirdas et al, 2001]. Patients with

[1] Corresponding author: Derek Sleeman, dsleeman@csd.abdn.ac.uk www.csd.abdn.ac.uk/ research/~sleeman

[2] Renal Medicine Section, Aberdeen Royal Infirmary & School of Medicine, The University of Aberdeen.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 57–66, 2007.

advanced renal failure where overall excretory renal function is below 15% often require renal replacement therapy. The most common form of treatment in the UK is unit based haemodialysis. Most patients receive standard haemodialysis 3 times a week for between 4-5 hours per session. A haemodialysis machine pumps blood from the patient through a high surface area semi-permeable dialysis membrane before returning the blood back to the patient. Fluid is removed through a process of ultrafiltration due to the pressure differential across the membrane and toxins are removed by diffusion across the membrane into a controlled dialysis solution. A range of treatment refinements have been developed that try to improve patient stability and therapeutic efficacy. They include the use of ultrafiltration alone or in combination with dialysis and dialysis treatment profiling where fluid removal or diasylate sodium concentration is dynamically changed during the treatment session. The entire dialysis process requires careful regulation and monitoring and is handled by an on board computer. Internal monitoring equipment records both machine parameters as well as patient physiological variables. Real time data is often used by clinicians to assess current physiological status but long term analyses of the whole treatment data-sets are unusual. We have used a combination of statistical, data mining [Hand et al, 2001], and theory refinement [Craw & Sleeman, 1990] approaches to address a range of clinical issues such as:

- Can major intra-dialytic complications such as hypotension be predicted?
- Are there distinct groups of patients with specific physiological behaviour characteristics?
- What are the major differences between patients who are stable on dialysis and those that are not?
- Is it possible to create patient specific optimum dialysis strategies [Sleeman et al., 2004].

2 Methods

Three dialysis machines (AK200 Ultra S, Gambro) installed at a hospital based satellite unit (Dr Gray's hospital, Elgin, Grampian, Scotland) have been fitted with a data collection interface node wirelessly linked to a server running data collection software (Exalis, Gambro). Data was routinely collected at each dialysis session delivered by the enabled machines. With this configuration data can be collected on a maximum of 12 patients per week of treatment with storage of approximately 144 hours of real time data. Initial pilot investigations have taken 288 hours of complete data sets for analysis. Collected parameters are detailed in Table 2.1.

The initial suite of programs in this series was implemented in the Neurological ICU at Western General Hospital in Edinburgh to monitor patients who have had traumatic head injuries, [Howells, 1994; McQuatt et al, 1999]. They implemented a monitor system which allows clinicians to review real-time data sets, as well as earlier data, for a number of important parameters on a monitor at the patient's bedside. Additionally, they implemented a BROWSER system for use by clinicians and data analysts to view the same data sets in an off-line mode. Further, both of these systems allow the clinician/administrator to define a series of insult levels for each of the

Table 2.1. Parameters collected during dialysis sessions and their frequency of collection

Parameter Name	Frequency Parameter Collected
Heart Rate	30mins
Systolic Pressure	30mins
Diastolic Pressure	30mins
Actual Weight Loss Rate	1min
Blood Volume	1min
Plasma Conductivity (AK)	30mins
Total Blood	1min
Actual Time	1min
NA Concentration	Once per session
KT/V	1min
Ionic Effective Dialysance	1min
Actual Total Weight Loss	1min

"channels"; for instance in the case of the NICU data sets the colours are normal – white, slight raised – yellow, considerably raised – orange, extreme value – red.

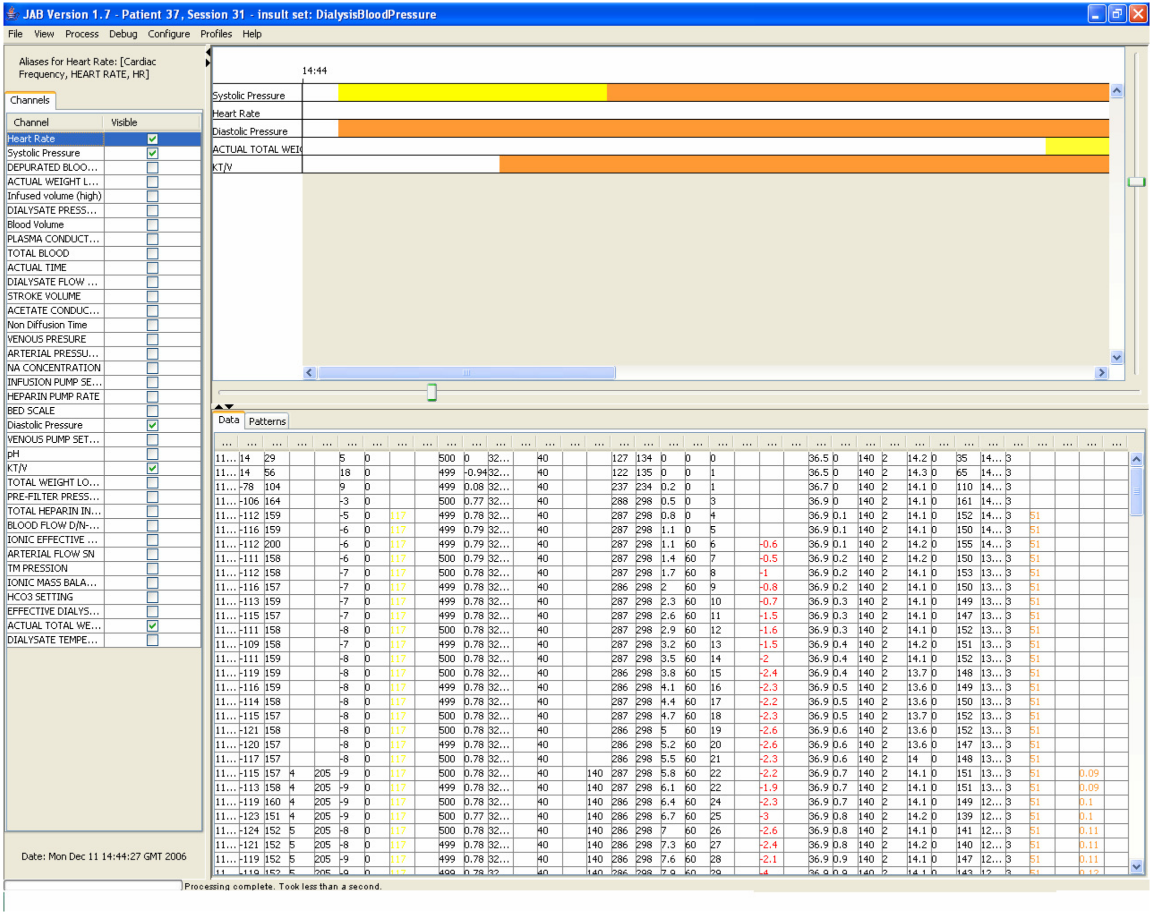
We, at Aberdeen, have extended the system to enable it to deal with not just head injury data sets but with a range of data sets including: dialysis data sets from Pavia (Italy), Aberdeen & Elgin, and an ITU data set. Further types of data sets can be added relatively easily and usually require a further specialized input routine. We took the opportunity to recode the system in Java – hence the new name, JAB (Java version of the Aberdeen Browser). Additionally the user is able to decide which of a range of parameters he/she wishes to have displayed at any time; the system also gives the user the chance to choose between several display formats. Further, the ability to define insult levels has been extended. This implementation thus reads initially details of the parameters collected, followed by the appropriate data set. Figure 2.1 shows the UI for this system displaying part of the Elgin / Aberdeen data set.

3 Results

In the introduction we outlined a number of top-level questions which we hope to address using data sets collected from dialysis patients. Our pilot study aimed to address two specific questions:

a. Is it possible to identify distinct clinical subgroups of patients from the dialysis data alone?
b. Can one predict the blood pressure or change in blood pressure during a dialysis session using the other collected variables?

We selected 72 sessions with complete data from 9 patients (i.e. 8 sessions per patient). This included a range of patients with differing clinical and demographic features. Some were recognised as stable patients whilst others were known to be more challenging to treat effectively.



Aliases for Heart Rate: [Cardiac Frequency, HEART RATE, HR]

Channels

Channel	Visible
Heart Rate	☑
Systolic Pressure	☑
DEPURATED BLOO…	□
ACTUAL WEIGHT L…	□
Infused volume (high)	□
DIALYSATE PRESS…	□
Blood Volume	□
PLASMA CONDUCT…	□
TOTAL BLOOD	□
ACTUAL TIME	□
DIALYSATE FLOW …	□
STROKE VOLUME	□
ACETATE CONDUC…	□
Non Diffusion Time	□
VENOUS PRESURE	□
ARTERIAL PRESSU…	□
NA CONCENTRATION	□
INFUSION PUMP SE…	□
HEPARIN PUMP RATE	□
BED SCALE	□
Diastolic Pressure	☑
VENOUS PUMP SET…	□
pH	□
KT/V	☑
TOTAL WEIGHT LO…	□
PRE-FILTER PRESS…	□
TOTAL HEPARIN IN…	□
BLOOD FLOW D/N-…	□
IONIC EFFECTIVE …	□
ARTERIAL FLOW SN	□
TM PRESSION	□
IONIC MASS BALA…	□
HCO3 SETTING	□
EFFECTIVE DIALYS…	□
ACTUAL TOTAL WE…	☑
DIALYSATE TEMPE…	□

Date: Mon Dec 11 14:44:27 GMT 2006

…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
11…	14	29			5	0			500	0	32…		40			127	134	0	0	0				36.5	0	140	2	14.2	0	35	14…	3				
11…	14	56			18	0			499	-0.94	32…		40			122	135	0	0	1				36.5	0	140	2	14.3	0	65	14…	3				
11…	-78	104			9	0			499	0.08	32…		40			237	234	0.2	0	1				36.7	0	140	2	14.1	0	110	14…	3				
11…	-106	164			-3	0			500	0.77	32…		40			288	298	0.5	0	3				36.9	0	140	2	14.1	0	161	14…	3				
11…	-112	159			-5	0	117		499	0.78	32…		40			287	298	0.8	0	4				36.9	0.1	140	2	14.1	0	152	14…	3	51			
11…	-116	159			-6	0	117		499	0.79	32…		40			287	298	1.1	0	5				36.9	0.1	140	2	14.1	0	150	14…	3	51			
11…	-112	200			-6	0	117		499	0.79	32…		40			287	298	1.1	60	6		-0.6		36.9	0.1	140	2	14.2	0	155	14…	3	51			
11…	-111	158			-6	0	117		500	0.79	32…		40			287	298	1.4	60	7		-0.5		36.9	0.2	140	2	14.2	0	150	13…	3	51			
11…	-112	158			-7	0	117		500	0.78	32…		40			287	298	1.7	60	8		-1		36.9	0.2	140	2	14.1	0	153	13…	3	51			
11…	-116	157			-7	0	117		499	0.78	32…		40			286	298	2	60	9		-0.8		36.9	0.2	140	2	14.1	0	150	13…	3	51			
11…	-113	159			-7	0	117		499	0.78	32…		40			287	298	2.3	60	10		-0.7		36.9	0.3	140	2	14.1	0	149	13…	3	51			
11…	-115	157			-7	0	117		499	0.78	32…		40			287	298	2.6	60	11		-1.5		36.9	0.3	140	2	14.1	0	147	13…	3	51			
11…	-111	158			-8	0	117		500	0.78	32…		40			287	298	2.9	60	12		-1.6		36.9	0.3	140	2	14.1	0	152	13…	3	51			
11…	-109	158			-7	0	117		499	0.78	32…		40			287	298	3.2	60	13		-1.5		36.9	0.4	140	2	14.2	0	151	13…	3	51			
11…	-111	159			-8	0	117		500	0.78	32…		40			287	298	3.5	60	14		-2		36.9	0.4	140	2	14.1	0	152	13…	3	51			
11…	-119	159			-8	0	117		500	0.78	32…		40			286	298	3.8	60	15		-2.4		36.9	0.4	140	2	13.7	0	148	13…	3	51			
11…	-116	159			-8	0	117		499	0.78	32…		40			286	298	4.1	60	16		-2.3		36.9	0.5	140	2	13.6	0	149	13…	3	51			
11…	-114	158			-8	0	117		499	0.78	32…		40			287	298	4.4	60	17		-2.2		36.9	0.5	140	2	13.6	0	150	13…	3	51			
11…	-115	157			-8	0	117		500	0.78	32…		40			287	298	4.7	60	18		-2.3		36.9	0.5	140	2	13.7	0	152	13…	3	51			
11…	-121	158			-8	0	117		500	0.78	32…		40			286	298	5	60	19		-2.6		36.9	0.6	140	2	13.6	0	152	13…	3	51			
11…	-120	157			-8	0	117		499	0.78	32…		40			286	298	5.2	60	20		-2.6		36.9	0.6	140	2	13.6	0	147	13…	3	51			
11…	-117	157			-8	0	117		500	0.78	32…		40			286	298	5.5	60	21		-2.3		36.9	0.6	140	2	14	0	148	13…	3	51			
11…	-115	157	4	205	-9	0	117		500	0.78	32…		40		140	287	298	5.8	60	22		-2.2		36.9	0.7	140	2	14.1	0	151	13…	3	51		0.09	
11…	-113	158	4	205	-9	0	117		499	0.78	32…		40		140	287	298	6.1	60	22		-1.9		36.9	0.7	140	2	14.1	0	151	13…	3	51		0.09	
11…	-119	160	4	205	-9	0	117		500	0.78	32…		40		140	286	298	6.4	60	24		-2.3		36.9	0.7	140	2	14.1	0	149	12…	3	51		0.1	
11…	-123	151	4	205	-9	0	117		500	0.77	32…		40		140	286	298	6.7	60	25		-3		36.9	0.8	140	2	14.2	0	139	12…	3	51		0.1	
11…	-124	152	5	205	-8	0	117		500	0.78	32…		40		140	286	298	7	60	26		-2.6		36.9	0.8	140	2	14.1	0	141	12…	3	51		0.11	
11…	-121	152	5	205	-8	0	117		499	0.78	32…		40		140	286	298	7.3	60	27		-2.4		36.9	0.8	140	2	14.2	0	140	12…	3	51		0.11	
11…	-119	152	5	205	-9	0	117		499	0.78	32…		40		140	286	298	7.6	60	28		-2.1		36.9	0.9	140	2	14.1	0	147	12…	3	51		0.11	
11…	-119	152	5	205	-9	0	117		499	0.78	32…		40		140	286	298	7.9	60	29		-4		36.9	0.9	140	2	14.1	0	143	12…	3	51		0.12	

Processing complete. Took less than a second.

Fig. 2.1. JAB's UI displaying a portion of a dialysis patient's data set

3.1 Determining Distinct Patient Groups in the Data Set: Clustering Analysis and Sub-group Discovery

We have performed an analysis on the group of 9 patients using several clustering algorithms. We used two different methods, namely hierarchical clustering using Ward's method with Euclidean distance [Ward 1963] and k-means clustering (with

Table 3.1. Allocation of patient data records to the three main clusters using hierarchical clustering. Each row contains the patient ID, number of records (percentage in brackets) in each cluster and total number of records for that patient.

Patient	LHS cluster	Centre cluster	RHS cluster	Total
20	48(80%)	0(0%)	12(20%)	60
23	1(2%)	9(19%)	38(79%)	48
25	0(0%)	52(93%)	4(7%)	56
26	0(0%)	59(97%)	2(3%)	61
28	3(5%)	5(9%)	48(86%)	56
30	0(0%)	56(87.5%)	8(12.5%)	64
33	53(95%)	0(0%)	3(5%)	56
35	0(0%)	36(64%)	20(36%)	56
37	0(0%)	0(0%)	56(100%)	56

varying number of clusters), [MacQueen 1967]. Each instance was described by a vector of 10 real-valued attributes: Systolic pressure, diastolic pressure, heart rate, blood volume, and the absolute changes of these four attributes in the previous 30 minute interval, as well as rate of fluid removal (weight loss) and rate of toxin removal (KT/v). So for a patient undergoing a 4-hour dialysis session, this procedure produces 7 such vectors: one at the end of the second 30 minute period when the values are compared with those at the end of the first 30 minute period, one at the end of third 30 minute period when the values are compared with those at the end of the second 30 minute period, etc.

The three clusters of the dendrogram produced by the hierarchical clustering algorithm are shown in Figure 3.1. Labels have the form *pp:ss* where *pp* is patient ID and *ss* is session number. For six patients (ID numbers 20, 28, 30, 33, 35, 37), it can be seen that a high percentage of their instances fall within a single cluster. The algorithm has identified three main sub-clusters: (a) the left-hand side (LHS) cluster in the dendrogram, dominated by records from patients 20 and 33, (b) the centre cluster dominated by patients 25, 26, 30 and 35, and (c) the right-hand side (RHS) cluster dominated by patients 23, 28 and 37. The allocation of patients to these three clusters is shown in more detail in Table 3.1. Comparing these clusters with the demographic data, we see that the LHS cluster contains patients with an average age of 26, compared to 66 in the rest of the dataset. Within the RHS cluster, the two most dominant patients, who are also clustered more compactly, namely patients 28 and 37, are the ones who suffer both from diabetes and cardio-vascular disease. And the third cluster corresponds to the remaining patients.

K-means clustering was performed with two, three, and nine clusters. The advantage of the k-means method is that it gives us a comparative quantitative evaluation of the clusters discovered with hierarchical clustering. The disadvantage is that we have to manually experiment with different numbers of clusters. The clustering was evaluated by matching clusters to the patient IDs, patient age (discretised in two groups), occurrence of diabetes, occurrence of cardio-vascular disease (CVD), and occurrence of both diabetes and CVD. The error rates (incorrectly clustered instances divided by

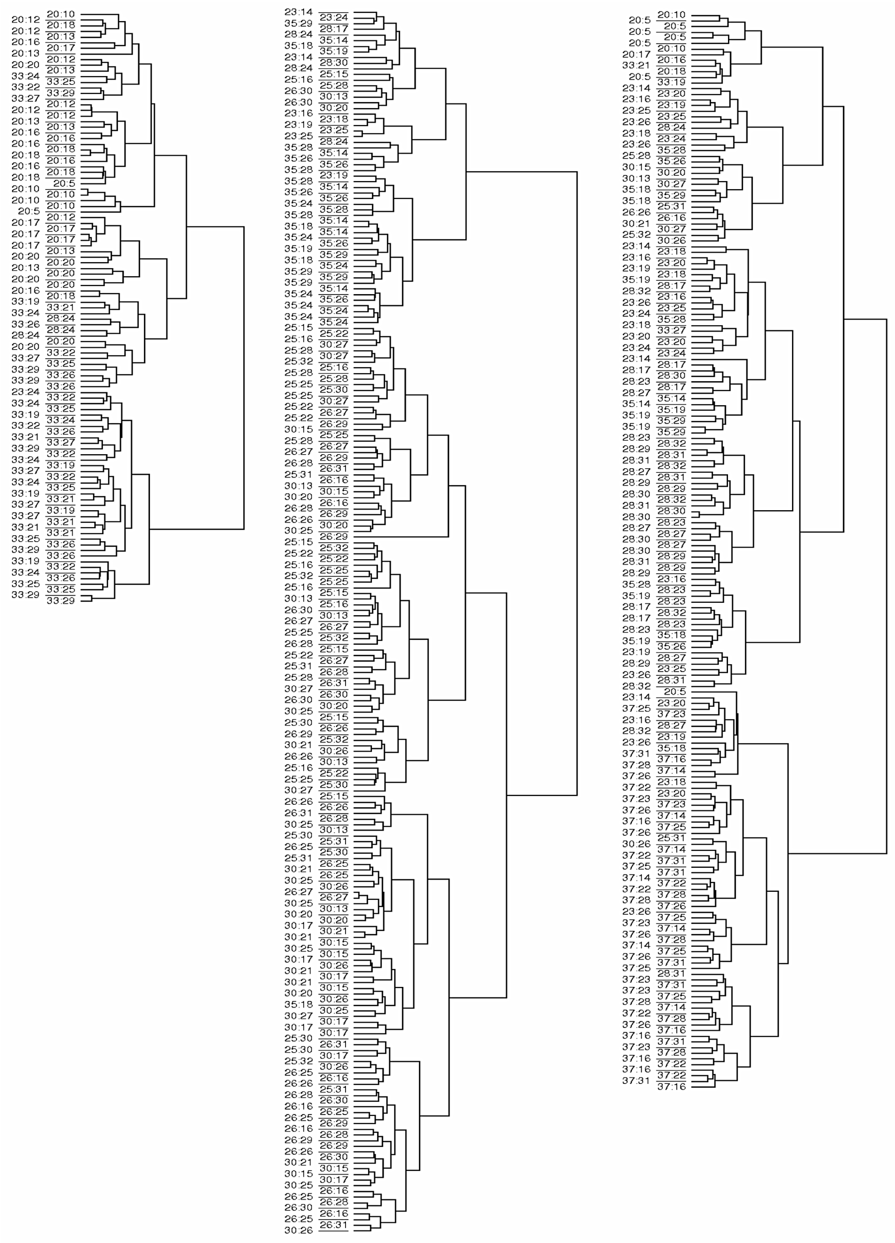


Fig. 3.1. The three main clusters derived using hierarchical clustering in detail. LHS cluster is dominated by patients 20 and 33, and the RHS cluster by patients 28 and 37.

Table 3.2. Error rates for different classes and numbers of clusters

# of clusters	Class	Error
9	Patient ID	27.7%
2	Age	2.9%
2	Cardio-vascular	33.7%
2	Diabetes	41.1%
3	CVD and Diabetes	19.5%

total number of instances) are summarised in Table 3.2. We see that for k=9 the different patients are clustered fairly well (28% error is not very high for a 9-class problem); binary clustering (for k=2) essentially separates the two age groups; finally, while CVD and diabetes independently do not form significant sub-groups, their conjunction (when k = 3) stands out as an important cluster.

Conclusions: Both types of clustering identify essentially the same 3 clusters, which is encouraging. Further, the clinician, using his considerable experience, had selected 3 distinct groups of patients (see the introductory paragraph of this section), and in fact he has confirmed that the 3 clusters identified by the algorithm correspond exactly to those 3 clusters. These 3 clusters are clinically significant as they would be expected to react to dialysis in distinctly different ways; in particular, the third group (diabetes and cardio-vascular disease) are likely to present more challenges to dialyse them successfully. So clinically it is likely that the latter group would be monitored more closely during dialysis. These results are encouraging because it appears to be possible for a machine learning algorithm on the basis of a small amount of demographic data and the information collected during the dialyses sessions, to identify these clinically significant groups. This also suggests that other patients who perhaps can not be clearly identified by their existing medical history, can be detected by the algorithm on the basis of their dialysis data sets, as patients who are likely to be "unstable" / complex patients to dialyse. Further, this suggests that more sophisticated analyses should be applied to the data sets produced by this group of "complex" patients, probing for example for early signs of instability etc.

3.2 Determining If It Is Possible to Predict Changes in the Patient's BP in Terms of Other Parameters

The aim of this analysis is to predict the occurrence of hypotension during a dialysis session. We were particularly interested in investigating for each patient whether the following are correlated:

1. Rate of change of BP and Rate of change of heart rate
2. Rate of change of BP and rate of Fluid removal
3. Rate of change of BP and rate of toxin removal

We consider a hypo-tensive event to be a drop greater than or equal to 10 mmHg in systolic pressure between two successive measurements (at 30 minute intervals). We have considered a dataset obtained from the same 9 patients, for each of which eight complete sessions of approximately four hours were available. Blood pressure (systolic and diastolic) and heart rate were measured at 30 minute intervals during these sessions. Toxin removal and fluid removal rates were available, and these were approximately constant within each individual session. Blood volume was measured every minute; we have converted that to 30 minute averages during pre-processing. The demographic data was also included in the analysis. We have incorporated some of the time-series information by creating the additional features of the changes in heart rate, blood volume, systolic and diastolic pressure since the previous measurement, and appending these to the data vector of each instance. The target attribute in each of these instances was derived from the change in systolic pressure in the next half hour. Each dialysis session yielded about 8 measurements; of these the first and the last were used to derive the changes of values compared to the second and last but one, respectively. We therefore had roughly 6 different instances labelled as hypotensive or non-hypotensive for each dialysis session. In fact since a few of the sessions were slightly longer, we obtained a total of 442 instances, of which 120 were labelled positive (hypotension event occurring) and 322 negative.

Table 3.3. Confusion matrix for hypotension prediction

	Predicted positives	**Predicted negatives**	**Total**
Actual positives	32	88	120
Actual negatives	11	311	322
Total	43	399	442

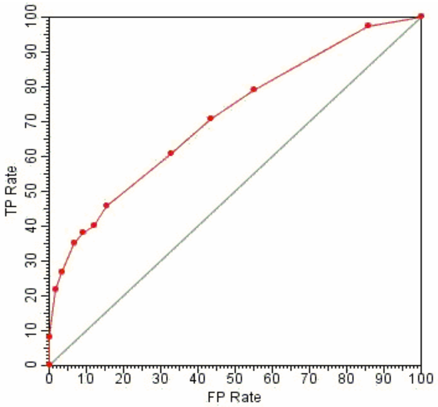


Fig. 3.2. ROC curve for hypotension prediction

Experiments were run using the WEKA data mining software tool [Witten and Frank 2005]. We used a Bayesian network classifier combined with a hill-climbing algorithm for structure learning, and evaluated the classifier using 10-fold cross validation. The ROC curve obtained by the algorithm (shown in Figure 3.2) was analysed to determine the optimal decision threshold for the Bayesian classifier. The confusion matrix for the optimal accuracy point is given in Table 3.3. The maximum accuracy reached was 77.6% (32+311/442), with a true positive rate of 26.7% (32/120) and false positive rate of 3.4% (11/322). This is an encouraging result, since predicting blood pressure trends over a half-hour period is a hard task; the algorithm predicts correctly more than a quarter of the hypo-tensive events, with only 11 false positive cases.

4 Conclusions and Further Work

These pilot studies have indicated that it may be possible to relate a patient's physiological behaviour on dialysis with their underlying pathological status. This offers an intriguing possibility that further analysis may allow more precise characterisation of patients and enable clinicians to tailor therapy more appropriately. Planned further work includes:

- Using the clustering approaches on a much wider range of patients to see if it is still possible for the algorithms to identify a number of clinically significant groups (e.g. those that are "unstable" / complex under dialysis)
- Investigating whether it is possible to improve the blood pressure predictions. As noted in section 3.2 some of the false positives (i.e. avoidance of hypotension) could well be due to nursing interventions (e.g. infusing of saline fluid). So in future we aim for better access to the patients' nursing notes.
- Pursuing some of the other issue list in section 1; addressing this agenda should enable us to progressively customize dialysis sessions to individual patients.
- Collecting opinions from a group of renal experts on a range of dialysis sessions, compare their analyses, and then hold face-to-face session(s) where these differences are discussed / resolved.
- Evaluating these Aides with a number of renal physicians and perhaps nurse practitioners.

Acknowledgements

- Ricardo & Roberto Bellazzi & colleagues (Pavia University & the Northern Italian Renal/ Dialysis Centre) gave us invaluable help in exporting the data collected by the EXALIS system in forms which could be used by our several analysis tools (e.g. the JAB system)
- This work has been supported under the EPSRC's grant number GR/N15764 and the Advanced Knowledge Technologies Interdisciplinary Research Collaboration, which comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University.
- The staff and patients of the Dialysis Unit, Dr Gray's Hospital, Elgin, Grampian.
- Bill Mackie & colleagues (Technical Staff, The Dialysis Unit, Aberdeen Royal Infirmary).

References

1. Craw, S., Sleeman, D.: Automating the Refinement of Knowledge-Based Systems. In: Aiello, L. (ed.) Proceedings of ECCAI-90, pp. 167–172. Pitman, London (1990)
2. Daugirdas, J.T., Blake, P.G., Ing, T.S.: Handbook of Dialysis, 3rd edn. Lippincott, Williams, & Wilkins, Baltimore (2001)
3. Hand, D.J., Mannila, H., Smyth, P.: Principles of Data Mining. MIT Press, Cambridge (2001)
4. Howells, T.P.: Edinburgh Monitor-Browser© Software, [v1.0.0] Computer Program (1994) tim.howells@nc.uas.lul.se
5. McQuatt, A., Andrews, P.J.D., Sleeman, D., Corruble, V., Jones, P.A.: The analyses of Head Injury data using Decision Tree techniques. Artificial Intelligence in Medicine. In: Horn, W., et al. (eds.) Proceedings of AIMDM'99 Conference, Aalborg, Denmark, June 1999, pp. 336–345. Springer, Heidelberg (1999)
6. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, vol. 1, pp. 281–297. University of California Press (1967)
7. Sleeman, D., Luo, Z., Christie, G., Coghill, G.: Analysing Time Series Medical Data-sets. In: Proceedings of Knowledge Based Systems & Services for Health Care, Bonn, May 2004, p. 1–4 (2004)
8. Ward, J.H.: Hierarchical Grouping to optimize an objective function. Journal of American Statistical Association 58(301), 236–244 (1963)
9. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2005)

Temporal Data Mining with Temporal Constraints

M. Campos[1], J. Palma[2], and R. Marín[2]

[1] Informatics and Systems Dept. Computer Science Faculty. University of Murcia
mcampos@dif.um.es
[2] Information and Communications Engineering Dept. Computer Science Faculty.
University of Murcia*

Abstract. Nowadays, methods for discovering temporal knowledge try to extract more complete and representative patterns. The use of qualitative temporal constraints can be helpful in that aim, but its use should also involve methods for reasoning with them (instead of using them just as a high level representation) when a pattern consists of a constraint network instead of an isolated constraint.

In this paper, we put forward a method for mining temporal patterns that makes use of a formal model for representing and reasoning with qualitative temporal constraints. Three steps should be accomplished in the method: 1) the selection of a model that allows a trade off between efficiency and representation; 2) a preprocessing step for adapting the input to the model; 3) a data mining algorithm able to deal with the properties provided by the model for generating a representative output.

In order to implement this method we propose the use of the Fuzzy Temporal Constraint Network (FTCN) formalism and of a temporal abstraction method for preprocessing. Finally, the ideas of the classic methods for data mining inspire an algorithm that can generate FTCNs as output.

Along this paper, we focus our attention on the data mining algorithm.

1 Introduction

The aim of Temporal Data Mining (TDM) algorithms is to extract and enumerate temporal patterns from temporal data. TDM is an intermediate step of the process of knowledge discovery on temporal databases (KDTD), which also includes other steps, such as data integration, cleaning, selection, transformation, preprocessing, and post-processing.

Most TDM techniques are based on conventional data mining techniques, which have been slightly modified in order to be applied to temporal data. However, the rich semantics of temporal information can be exploited to devise TDM algorithms that provide output that is more informative.

* This work was supported by the Spanish Ministry of Education and Science and the European Regional Development Fund of the European Commission (FEDER) under grants TIC2003-09400-C04-01, TIN2006-15460-C04-01 and PET2006-0406.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 67–76, 2007.

Following this idea, a number of methods for discovering expressive temporal patterns have been proposed [13]. Association rules, episodes and sequences are the basic kind of temporal patterns. Temporal association rules have the form $P(v, t_1) \rightarrow Q(w, t_2)$, as in $\{whisky\}_{(today)}\{ginebra\}_{(today)} \rightarrow \{hangover\}_{(tomorrow)}$. A rule establishes a time window in which the consequent and antecedent frequently happen. Sequential patterns are formed by chains of events (time points or time intervals) joined by means of the operator $BEFORE$. Temporal episodes are collections of events that occur relatively close to each other in a given partial order, that is, two events can be sequential or parallel.

The next step is mining temporal relations more complex than the simple chaining of events. However, the more temporal relations are used, the more the complexity of the process is increased. Thus, recently proposed models limit the number of temporal relations used. For example, [15] only uses $CONTAINS$ and $BEFORE$ relations, and [8] establishes a language of patterns in the form $(((a\,rel\,b)\,rel\,c)\,\ldots)$. In order to deal with this kind of temporal relations, a temporal reasoning mechanism must be applied.

In this paper, we address two problems. In the first place, we propose the use of a temporal constraint propagation algorithm as temporal reasoning mechanism. By propagating constraints, we can build expressive, complete and consistent temporal patterns, including temporal relations between both time points and intervals. Computational complexity is bounded to practical limits if small patterns are considered.

The other problem we address is how to represent temporal imprecision in the patterns. Temporal imprecision is inherent to complex domains like medicine. For instance, it is practically impossible for a physician to establish the precise time that should elapse between one manifestation and another for them to be considered as linked within one diagnostic hypothesis. Recently, some works on temporal data mining regarding temporal imprecision (in contrast to imprecision in values), have been published [16,5,14].

Our proposal is based on three elements: representation of temporal information, preprocessing of temporal data and a temporal data mining algorithm. For the first element, we have selected a model that provides powerful temporal reasoning capabilities and establishes a trade off between expressiveness and efficiency. In particular, we propose the use of the Fuzzy Temporal Constraint Networks (FTCN) formalism [9], which will be used for both the input and the output of the mining process. Other models, such as [2], could be eligible. Secondly, the data are preprocessed by means of a temporal abstraction algorithm. Hence, it is possible to work at a higher knowledge level and to reduce the volume of data. The result of this stage is a set of sequences of states. The states present in different sequences can be linked by means of temporal constraints.

Finally, we have developed a temporal data mining algorithm inspired on apriori to discover more informative temporal patterns (FTCNs). The algorithm applies temporal constraint propagation for pruning non-frequent patterns.

This paper focuses on the third element, that is, on the data mining algorithm. The structure of the rest of the document is as follows. In Section 2, we briefly explain the chosen representation model and the preprocessing stage. In Section 3, the algorithm for mining temporal relations is explained. Finally, we describe related works, conclusions and future research.

2 Reasoning and Temporal Abstraction by Means of FTCN

Within an environment rich in temporal data, it is necessary to represent information that can be given in form of points and intervals, and to qualify the data mining method to deal with quantitative or qualitative data in a homogeneous way. Furthermore, in the data mining method we want to deal with data that may not have a concrete mark of time, but that can have a quantitative or qualitative temporal relation with other data. In addition, in certain contexts it is necessary to deal with temporal vague information, e.g., in textual descriptions written by physicians it is usual to read expressions such as “a symptom appears about 1 or 2 days before the admission”.

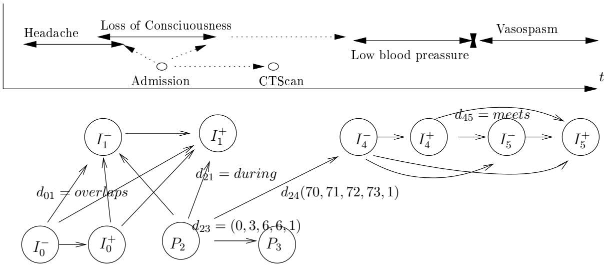
In our model, we have considered the representation of temporal concepts in form of time points or time intervals, and by means of quantitative relations (between points) or qualitative (between points, between points and intervals and between intervals). The relations and the algebra of intervals, point-intervals, and points are widely accepted and used by the community of temporal reasoning. Since the problem of reasoning with the full algebra for temporal relations is NP-complete, we have chosen one of the tractable subalgebras: the convex relations implemented in the Fuzzy Temporal Constraint Network (FTCN).

This model allows us to handle fuzzy temporal information, whose utility in medicine has already been proven by different authors. This model has recently been used to represent information of discharge of patients with a satisfactory result [18]. An FTCN can be represented by a graph in which nodes correspond to temporal variables and arcs to the constraints between them. Each binary constraint between two temporal variables is defined by means of a fuzzy number, that is a convex possibility distribution, which restricts the possible values of the time elapsed between both temporal variables.[1]

The upper part of Figure 1 shows an example of the temporal distribution of an episode of subarachnoid hemorrhage (SAH) in a patient at ICU. Every interval is translated into a point representation, and each qualitative relation is translated into a quantitative one in order to obtain a FTCN. In the lower part of the figure, we show only the explicit temporal constraints between points (I_0^- and I_0^+ denote the beginning and the end of the interval I_0).

Two fundamental operations must be performed with the temporal patterns: determination of the consistency and inference of relations between the

[1] A convex non-normalized trapezoid possibility distribution is given by a 5-tuple=(a, b, c, d, h), where $[a, d]$ defines the support, $[b, c]$ defines the kernel, and $h \in [0, 1]$ is the degree of possibility of the distribution. The precise point of time 4 is represented as $(4, 4, 4, 4, 1)$, and a precise time interval between 5 and 6 is represented as $(5, 5, 6, 6, 1)$. The value of h is 1 when it is omitted.



d_{01}= (Headache (I_0), *overlaps* with loss of consciousness, (I_1));
d_{21}= (Admission, (P_2), *during* loss of consciousness,(I_1));
d_{32}= (CT-Scan, (P_3), *in less than 6 hours after* Admission, (P_2));
d_{52}= (Low blood pressure, (I_4^-), *approximately 72 hours after* Admission, (P_2));
d_{45}= (Low blood pressure, (I_4), *meets* vasospasm, (I_5));

Fig. 1. Possible FTCN example of events of a patient with SAH

temporal variables of the pattern. These operations have a direct representation in the processes of constraint propagation and minimization of the FTCN. These operations have an affordable computational cost by a trade off between representation capacity or expressivity and efficiency. Thus, this model fulfils one of the characteristics we seek for with the purpose of serving as base of a process of data mining: the efficiency in the reasoning process. For instance, by means of the constraint propagation, in the example of Figure 1, we can establish a temporal relation between the vasospasm complication and the loss of consciousness symptom.

In order to fill the gap between the *FTCN* and the high-level temporal language, a temporal reasoner called FuzzyTIME (*Fuzzy Temporal Information Management Engine*) [4] has been used. FuzzyTIME provides procedures for maintaining and querying temporal information (with both points or intervals, and quantitative or qualitative relations) at *FTCN* level. The querying ability, which can be about necessity or possibility, can be used for complex abstractions.

This formalism allows us to define a process of temporal abstraction of data with a double objective. In the first place, the abstraction method has the aim, in a medical context, of interpreting the data of some patient to adapt them to a higher level of expression. In the second place, it provides an abstract explanation, temporarily consistent, of a sequence of events (time points or time intervals). This explanation adopts the form of a sequence of states (intervals) that are obtained by means of an abductive process (the details of this process escape to the scope of this work). As a result, the volume of data is reduced because several points can be subsumed in one interval.

For example, if we consider a series of blood pressure measurements, we could obtain a series of states that indicate the level in a meaningful way to the problem we are dealing with (see the lowest part of Figure 2). For the SAH, the blood pressure is a meaningful variable that can be abstracted with functions such as "if blood pressure is above 130, then is high" for obtaining qualitative states. That is to say, it allows us to transform data into concepts of higher level and facilitates the interpretation of the output patterns.

3 Temporal Data Mining over FTCN

The objective of the algorithm is to discover complex temporal patterns on the input, and represent them as FTCN. The input consists of a set of patients (see Figure 2), where variables are grouped in sequences of temporal points (e.g., diagnostics) or sequences of temporal intervals (e.g., states of blood pressure). Theses sequences are formally represented by FTCNs obtained in the process of abstraction.

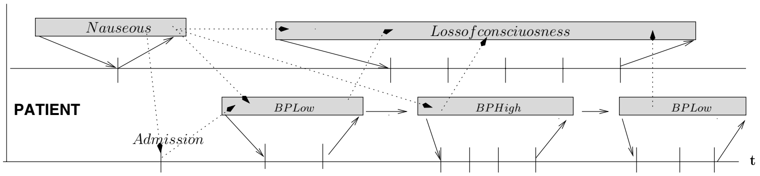


Fig. 2. A partial schematic view of a patient

The data mining algorithm follows a breadth-first approach, inspired on the classic apriori ideas of candidate generation and itemset count [1]. In our case, the search space is the constraints space (instead of the items or entities), so in every step we extend a pattern with a new temporal relation. The generation of candidates becomes a bit more complicated, but it allows us a number of prunings based on the propagation of temporal constraints (together with the ones based on support). Moreover, due to this high cost that entails the generation of all the candidates and frequency count, our strategy consists of generating only frequent patterns, making the count and the extension of the patterns simultaneously.

In order to limit the search space, we use several parameters such as the minimum support (understood as the number of patients where a relation appears), the size of the patterns (given in a maximum number of constraints, including convex ones) and a maximal temporal extension of the pattern (maximum distance between two variables of the pattern).

The skeleton of the algorithm is as follows:

1. Enumerate frequent temporal entities
2. Enumerate frequent basic temporal relations
3. Extend patterns in an iterative way while pruning non-frequent and redundant patterns.

3.1 Frequent Temporal Entities and Relations

As first step in the pattern extraction process, it is necessary to establish which the frequent entities are. An entity is frequent when its *support* is higher than a given threshold, that is to say, the number of patients in which the entity can be observed is above the threshold.

From every pair of frequent entities we can establish a qualitative temporal relation (explicitly represented in the patient or inferred see dotted lines in Figure 2) in the form *TemporalEntity TemporalConstraint TemporalEntity.* Where *TemporalConstraint* stands for Allen's relations when the entities are both intervals; *before*, *equals,* or *after* when both entities are points; and *before, starts, during*, *finishes* or *after* when one entity is a point and the other an interval.

Since our method works on relations, we also use a threshold for the support of these temporal relations as a way of limiting the search.

3.2 Temporal Pattern Extension

Temporal patterns are built incrementally by adding frequent basic relations. A new basic relation can be included in a temporal pattern only if the resulting temporal pattern remains frequent. This extension process starts with the setting of basic temporal relations, enumerated in the previous step, which are considered as temporal patterns of size 1. As mentioned before, these temporal patterns are represented in a FTCN together with a reference to the patients supporting them.

Each time a new frequent basic temporal relation is added to a temporal pattern, the support must be calculated again. This new support can be easily obtained as the cardinality of the intersection of two patients sets, one associated to the temporal pattern and another one associated to the frequent basic temporal relation added. If this new support is below a given threshold, the temporal pattern will not be considered as frequent and can be pruned.

This process has two advantages: 1) it makes an early pruning possible, and 2) it is only necessary to count the patients where both the temporal pattern and the frequent basic relation are present, instead of counting the number of temporal pattern instances in the original data base. However, this technique implies a high memory usage. In order to optimize the use of memory, some authors propose a depth-first approach. The depth-first technique has the limitation of losing the information of previous levels, which would be used to make a more effective pruning and it would require a more intense postprocessing phase in order to avoid repeated temporal patterns.

Avoiding the generation of repeated temporal patterns is one of the main difficulties that arise in TDM process. To this end, the lexicographical order heuristic (LOH) is introduced. LOH imposes an order in the temporal patterns events, allowing us to reduce the number of frequent basic relations that can be considered for temporal pattern extension. For instance, if we have $R1$: $A - before - B$ and $R2$: $B - before - C$, the algorithm would add $R2$ to a pattern containing $R1$, but never the other way round.

Table 1. Weights for constraints from Van Beek's heuristics

constraint	o	oi	d	di	b	bi	s	si	m	mi	f	fi	eq
weights	4	4	4	3	3	3	2	2	2	2	2	2	1

Patient	Event	Constraint
1	Admission (A)	(12,12,12,12)h
1	CT-positive (CP)	(13,13,13,13)h
1	Headache (H)	(1,2,5,6,1)h
1	Loss of consciousness (LC)	(5,6,20,21,1)h
1	Low BP episode (BL)	(5,6,26,27)h
1	Vasospasm (V)	(26,26,28,28)h
2	Nauseous (N)	(1,4,18,20)h
2	Headache (H)	(1,3,12,16)h
2	CT-negative (CN)	(26,26,26,26)h
3	Headache (H)	(1,3,10,12)h
3	Admission (A)	(12,12,12,12)h
3	Low BP episode (BL)	(5,6,16,17)h
3	Loss of consciousness (LC)	(15,16,17,18)h
3	CT-positive (CP)	(17,17,17,17)h
3	Vasospasm (V)	(16,16,20,20)h
4	Admission (A)	(12,12,12,12)h
4	CT-positive (CP)	(14,14,14,14)h
4	Headache (H)	(1,3,8,10)h
4	Loss of consciousness (LC)	(3,5,17,18)h
4	Low BP episode (BL)	(15,16,19,20)h
4	Vasospasm	(20,20,28,28)h

Basic Pattern	Relation
B1	A during LC
B2	A during BL
B5	CP during LC
B6	CP during BL
B18	H overlaps BL
B0	A before CP
B3	A before V
B7	CP before V
B11	LC before V
B20	H before V
B12	LC starts BL
B10	BL meets V

Patterns
B10-B11
B18-B12
B18-B12-B11

Fig. 3. Sample database for 4 patients, frequent basic relations and discovered patterns

Once the temporal pattern is extended, we apply the constraint propagation. On the one hand, this process allows us to detect temporal inconsistent patterns and, thus, prune them. On the other hand, new temporal relations will be inferred. If any of these inferred temporal relations is not a basic temporal relation, the resulting temporal pattern is also pruned. The order in which temporal relations are added to the pattern is based on the weights assigned to each one of the basic constraints proposed by Van Beek (shown in Table 1). This order allows us to use as soon as possible those relations that can provide fewer solutions to the FTCN.

The constraint propagation process has two important advantages. First, by inferring all the possible temporal constraints between all the temporal entities, we can determine when a basic relation, considered for pattern extension, is already present. Therefore, the number of basic frequent relations considered for extension can be reduced. It has to be taken into account that, when an extended temporal pattern is minimized, its temporal constraints become more precise. Thus, if the original temporal pattern is frequent, its minimized version is also frequent and is used in the following steps. The algorithm finishes when there is no possible extension for any pattern.

Table 2. Complete pattern of size 2

	Headache	LowBP	Vasospasm
Loss Consciousness	-	B12(starts)	B11(before)
Low BP	-	-	B10(meets)
Vasospasm	-	-	-

Let us illustrate the previous concept with the example depicted in Figure 3. Two facts can be pointed out. The first one is that there is only a frequent pattern of size 3, B18-B12-B11 and some of the subpatterns may not appear in the list of frequent patterns of size 2. The second one is that there are just a few patterns of size 2 despite the fact that several patterns of size 1 could have been combined. Both facts are motivated by the constraint propagation process; e.g., if we consider the B10-B11 pattern and propagate the constraints, we can see that the B10-B12 pattern can be inferred, thus it is redundant and can be pruned (see Table 2).

If we make the same consideration for the pattern of size three, we can see that there is no need to evaluate more patterns because all possible basic patterns are derived from this one.

4 Conclusions and Future Works

In this paper, we have described a method for TDM that is based on complex temporal reasoning. Three steps have been set out: 1) We propose the use of FTCN to represent the pattern structure, since it provides a formal model able to represent a rich set of temporal relationships; 2) A preprocessing phase is performed by applying a temporal abstraction mechanism for generating a set of sequences of states interlinked by temporal constraints, thus reducing the data volume; 3) We apply an algorithm for TDM that takes advantage of the formal model for temporal reasoning to generate complex patterns.

The proposed method provides several contributions: 1) The input data can include both time points and time intervals; 2) a formal model for imprecision management is applied; 3) the mining algorithm applies temporal constraint propagation to prune non-frequent patterns; 4) the output is a constraint network including explicit and implicit temporal relationships.

The main disadvantage is a higher execution time, but it provides a reasonable trade off between expressive power and efficiency. Time complexity of constraint propagation in FTCN is $O(n^3)$, where n is the number of points in the pattern. This means that, for small patterns, the method is fast enough. In addition, there are efficient versions of FTCN constraint propagation algorithm for specific graph topologies (if the input is a set of sequences without mutual constraint, time complexity is linear).

Currently, we are applying this TDM method to data coming from an ICU. In ICU, temporal imprecision is present in the data; both time point (e.g., diagnostics) and time intervals (e.g., treatments) are needed; dense data (e.g., monitoring data) and sparse data (e.g., laboratory test) coexist. Hence, we can take advantage of the capacities of the proposed method.

A number of related works must be cited. Both [6] and [12] combine temporal abstraction and temporal data mining, although they use temporal abstraction as a means for extracting temporal features that can be used as variables in a learning process. Moskovitch and Shahar [11] emphasize the importance of mining temporal abstractions and their advantages, but they do not provide any implementation.

Bellazzi et al. [3] also combine temporal data mining and temporal abstraction in a supervised search on temporal multi-variate series that are preprocessed to be reduced to states. Their patterns are limited to detect contemporary episodes and those episodes that *precede* a certain given event.

Morris and Khatib [10] describe a general process to apply temporal knowledge to TDM. It is based on a set of abstractions on temporal information whose basic element is the *profile*. A profile is a concise representation of the temporal information on distance or arrangement between a set of intervals. The profiles contain patterns used to determine the consistency of a set of constraints or to detect useful temporal patterns.

Finally, Winarko and Roddick [17] propose an algorithm for mining sequences of intervals generating temporal association rules, in a similar way to [7] but with an approach non based on apriori-like techniques.

Some characteristics differentiate our work from the previous ones. None of them simultaneously includes the main features of our method: the use of a formal temporal reasoning model, the ability to manage temporal imprecision and the possibility of dealing with data in form of intermixed points and intervals.

Some changes in our mining algorithm are in course. The aims are to generate less redundant patterns (following some ideas of the graph mining) and to exploit convex temporal relationships to simplify the management of patterns.

For future work, we plan to introduce some interaction between the qualitative abstraction and the data mining algorithm to determine dynamically the abstraction level suitable for each element. For example, if a patient is receiving a treatment consisting of several drugs of the same kind (e.g. different painkillers) overlapped in the timeline, the temporal abstraction mechanism summarizes all these events in one interval. The mining algorithm could choose a higher level concept to reduce the size of the generated patterns, thus making them more informative and clear. Moreover, subtle temporal nuances can be included in the patterns, by means of linguistic modifiers like “long before” or “shortly before”.

References

1. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S. (eds.) Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C, 26–28, pp. 207–216. ACM Press, New York (1993)
2. Badaloni, S., Falda, M., Giacomin, M.: Integrating quantitative and qualitative fuzzy temporal constraints. AI Communications 17(4), 187–200 (2004)

3. Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R.: Temporal data mining for the quality assessment of hemodialysis services. Artificial intelligence in medicine 34, 25–39 (2005)
4. Campos, M., Cárceles, A., Palma, J., Marín, R.: A general purpose fuzzy temporal information management engine. In: Advances in information and communication technology, in EurAsia-ICT 2002, pp. 93–97 (2002)
5. Guil, F., Bosch, A., Bailón, A., Marín, R.: A fuzzy approach for mining generalized frequent temporal patterns. In: Workshop on Alternative Techniques for Data Mining and Knowledge Discovery. Fourth IEEE International Conference on Data Mining (ICDM 2004), Brighton, UK (2004)
6. Ho, T.B., Nguyen, T.D., Kawasaki, S., Le, S.Q.: Combining Temporal Abstraction and Data Mining Methods in Medical Data Mining. chapter 7, vol. 3, pp. 198–222. Kluwer Academic Press, Dordrecht (2005)
7. Höppner, F., Klawonn, F.: Finding informative rules in interval sequences. Intelligent Data Analysis 6(3), 237–255 (2002)
8. Kam, P.S., Fu, A.W.C.: Discovering temporal patterns for interval-based events. In: Kambayashi, Y., Mohania, M.K., Tjoa, A.M. (eds.) DaWaK 2000. LNCS, vol. 1874, pp. 317–326. Springer, Heidelberg (2000)
9. Marín, R., Mira, J., Patón, R., Barro, S.: A model and a language for the fuzzy representation and handling of time. Fuzzy Sets and Systems 61, 153–165 (1994)
10. Morris, R., Khatib, L.: General temporal knowledge for planning and data mining. Annals of Mathematics and Artificial Intelligence 33(1), 1–19 (2001)
11. Moskovitch, R., Shahar, Y.: Temporal data mining based on temporal abstractions. In: ICDM 2005 Workshop on Temporal Data Mining: Algorithms, Theory and Application. TDM2005, pp. 113–115 (2005)
12. Peek, N., Abu-Hanna, A., Peelen, L.: Acquiring and using temporal knowledge in medicine: an application in chronic pulmonary disease. In: ECAI'02 Workshop on Knowledge Discovery from (Spatio-)Temporal Data, pp. 44–50 (2002)
13. Roddick, J.F., Spiliopoulou, M.: A survey of temporal knowledge discovery paradigms and methods. IEEE Transactions on Knowledge and Data Engineering 14(4), 750–767 (2002)
14. Sudkamp, T.: Discovery of fuzzy temporal associations in multiple data streams. In: Hoffmann, F., Köppen, M., Klawonn, F., Roy, R. (eds.) Advances in Soft Computing, pp. 1–13. Springer, Heidelberg (2005)
15. Villafane, R., Hua, K.A., Tran, D., Maulik, B.: Knowledge discovery from series of interval events. Journal of Intelligent System Information 15(1), 71–89 (2000)
16. Vincenti, G., Hammell, R.J., Trajkovski, G.: Data mining for imprecise temporal associations. In: 6th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks. SNPD/SAWN 2005, pp. 76–81 (2005)
17. Winarko, E., Roddick, J.F.: Discovering richer temporal association rules from interval-based data. In: 7th International Conference on Data Warehousing and Knowledge Discovery, DaWak, pp. 315–325 (2005)
18. Zhou, L., Melton, G.B., Parsons, S., Hripcsak, G.: A temporal constraint structure for extracting temporal information from clinical narrative. Journal of Biomedical Informatics 39(4), 424–439 (2006)

A Nearest Neighbor Approach to Predicting Survival Time with an Application in Chronic Respiratory Disease

Maurice Prijs[1], Linda Peelen[1], Paul Bresser[2], and Niels Peek[1]

[1] Dept. of Medical Informatics
[2] Dept. of Pulmonology
Academic Medical Center – Universiteit van Amsterdam
m.c.prijs@amc.uva.nl

Abstract. The care for patients with chronic and progressive diseases often requires that reliable estimates of their remaining lifetime are made. The predominant method for obtaining such individual prognoses is to analyze historical data using Cox regression, and apply the resulting model to data from new patients. However, the black-box nature of the Cox regression model makes it unattractive for clinical practice. Instead most physicians prefer to relate a new patient to the histories of similar, individual patients that were treated before. This paper presents a prognostic inference method that combines the k-nearest neighbor paradigm with Cox regression. It yields survival predictions for individual patients, based on small sets of similar patients from the past, and can be used to implement a prognostic case-retrieval system. To evaluate the method, it was applied to data from patients with idiopathic interstitial pneumonia, a progressive and lethal lung disease. Experiments pointed out that the method competes well with Cox regression. The best predictive performance was obtained with a neighborhood size of 20.

1 Introduction

In patients with a chronic progressive disease, medical decisions are often influenced by the patient's prognosis. Therefore reliable estimates of remaining lifetime, based on the current condition of the patient, are required. In communication with the patient, the expected prognosis is often expressed in terms of survival probabilities for various time intervals (e.g., 12, 24, and 60 months).

The predominant method for obtaining such survival probabilities is to analyze historical data using Cox proportional hazards (PH) regression analysis and to use the resulting predictive model to construct a survival curve for the individual patient. Although the Cox PH analysis has proven to be a valuable tool in epidemiology and healthcare research, its usefulness in clinical practice is limited by its 'black-box' nature: for the clinical user of a Cox model, it is unclear how the patient's characteristics are used in constructing the survival curve. Instead most physicians prefer to relate a new patient to the histories of similar, individual patients that were treated before. These case histories also

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 77–86, 2007.

provide valuable other information, for instance about mobility status at various time points and subjective experience of the progressing disease.

This paper presents a prognostic inference method which combines Cox regression with the nearest neighbor paradigm, and which can be used to implement a prognostic case-retrieval system. In brief, the method selects nearest neighbors of the new patient from a database of historical observations, using a distance measure based on the variables and their weights resulting from the Cox regression analysis. The survival outcomes of the neighbors are used to construct a nonparametric Kaplan-Meier survival curve for the individual patient. As part of a case-retrieval system, it can be used to collect useful other types of information about similar patients from the past.

This paper is organized as follows. Section 2 provides a brief review of survival analysis, Cox regression models, and introduces our method. Section 3 presents a case study in the field of chronic lung diseases, including a statistical evaluation of predictive performance. The paper is completed with a discussion and conclusions in Sect. 4.

2 Methods

2.1 Survival Analysis

Let $t_1, \ldots, t_n$ be observed survival times for n individuals, and let $\delta_1, \ldots, \delta_n$ be associated censoring indicators, where $\delta_i = 0$ means that individual i was still alive at time t_i and $\delta_i = 1$ means that t_i was the individual's time of death. The statistical basis for both the Cox regression model and our prognostic method is the nonparametric estimation method from Kaplan and Meier [1]. It is a nonparametric estimate of the probability $S(t)$ that a random individual from the given population will have a lifetime exceeding t. When the survival times are ordered such that $t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(j)}$, where $t_{(j)}$ is the jth largest unique survival time, the Kaplan-Meier estimate is defined as:

$$\hat{S}_0(t) = P(T \geq t) = \prod_{j \,|\, t_{(j)} \leq t} \left(1 - \frac{d_j}{r_j}\right) \ , \tag{1}$$

where r_j is the number of individuals at risk (i.e., alive and not censored) just before $t_{(j)}$, and d_j is the number of individuals that died at $t_{(j)}$.

2.2 Cox Regression Models

The method of Kaplan and Meier estimates marginal (i.e., population-averaged) survival probabilities. To arrive at individual survival estimates, we need to use a method that takes information from these individuals into account. Let $\mathbf{X}$ be an $n \times p$ matrix of covariate patterns, where x_{ij} denotes the jth covariate and $\mathbf{x}_i$ the covariate pattern of individual i. Cox PH regression models consist of a nonparametric and a parametric component, that collaborate in the construction of estimated survival curves for individuals, based on their covariate patterns [2]. The non-parametric component is the population-based Kaplan-Meier survival curve $\hat{S}_0$; the parametric part is a linear regression model that adjusts the marginal survival probabilities through an exponential link function. Formally, the survival probability $\hat{S}_i^{\mathrm{Cox}}(t)$ for individual i at time point t is computed as

$$\hat{S}_i^{\mathrm{Cox}}(t) = P(T > t \mid \mathbf{x}_i) = \hat{S}_0(t)^{\exp(\eta(\mathbf{x}_i))} \ , \tag{2}$$

where

$$\eta(\mathbf{x}_i) = \beta_1(x_{1i} - \bar{x}_1) + \cdots + \beta_p(x_{pi} - \bar{x}_p) \ . \tag{3}$$

Here, $\bar{x}_j$ is the average value of the jth covariate in the dataset. Thus, depending on the individual's deviance from average values, the baseline survival probabilities are increased, decreased, or remain unchanged. A fundamental assumption of the Cox PH regression model is that accurate survival functions for individuals (or subgroups of the population) are obtained through a proportional adjustment of the baseline survival function over time.[1]

The regression parameters $\beta_1, \ldots, \beta_p$ are estimated by optimizing a partial likelihood function that is based on Eq. 2. This type of estimation procedure is implemented in all major statistical software packages. Dedicated feature subset selection procedures exist (e.g., [3]) for eliminating irrelevant covariates and preventing the model from overfitting. For further details on the Cox regression model and its various extensions, we refer to [4].

2.3 *k*-Nearest Neighbor Survival Prediction

The *k-Nearest Neighbors* (k-NN) algorithm [5] solves classification and regression problems without explicitly building a model. Instead, when a prediction needs to be made for a particular individual, the algorithm selects a set of k similar instances from the data, and returns their average (regression) or dominant (classification) response. The k-NN algorithm resembles the retrieval step of case-based reasoning methods, a well-known decision support paradigm [6,7].

Classification and regression using the k-NN algorithm have several advantages and disadvantages. On the one hand, the algorithm makes few assumptions about the underlying regularities in the domain, and can therefore be used to approximate virtually every function. On the other hand, however, k-NN methods behave poorly in high-dimensional domains [8]. Furthermore, the algorithm is easily fooled by differences in scales on which the covariates are expressed [9].

The k-NN algorithm is usually applied for classification, smoothing, or binary regression; it has not been used in the context of survival analysis. In this paper we present a k-NN method for making individual survival predictions, which operates as follows. First, a Cox regression analysis with stepwise feature subset selection is conducted on the training dataset. Let η denote the linear predictor

[1] More precisely, the model assumes that such functions can be obtained through a proportional adjustment of the baseline *hazard* function – hence the name 'proportional hazards'.

of the resulting Cox model, as in (3). The value $\eta(\mathbf{x}_i)$ is called the *score* of individual i. We use the score difference

$$d(i,j) = |\eta(\mathbf{x}_i) - \eta(\mathbf{x}_j)| \tag{4}$$

to quantify the distance between individuals i and j.

Second, when making a prediction for an individual i, its k nearest neighbors are selected, and a survival curve is constructed from their survival times, using the method of Kaplan and Meier. Formally, let $\mathbf{x}_{[1]}, , \ldots, \mathbf{x}_{[k]}$, be the k individuals in $\mathbf{X}$ closest to individual i in the dataset $\mathbf{X}$, and let $t^*_{(1)} < \ldots < t^*_{(m)}$, $m \leq k$, be their unique and ordered survival times. Then, the k-NN estimated survival probability for individual i at time point t equals

$$\hat{S}_i^{k\mathrm{NN}}(t) = \prod_{j \,|\, t^*_{(j)} \leq t} \left(1 - \frac{d_j}{r_j}\right) \ , \tag{5}$$

where r_j is the number of neighbors at risk (i.e., alive and not censored) just before $t^*_{(j)}$, and d_j is the number of neighbors that died at $t_{(j)}$.

3 Case Study in Idiopathic Interstitial Pneumonia

Idiopathic interstitial pneumonia (IIP) comprises a group of disorders of unknown etiology which are characterized by a variable pattern of inflammation and/or fibrosis of the pulmonary interstitium, causing shortness of breath on exertion and dry cough. It is a rare disease, with a median survival generally being reported as 2 to 4 years, but there is substantial heterogeneity in survival time among patients [10]. A reliable prediction of survival probabilities for each new patient is of great value for physicians caring for these patients. Decisions on referral for lung transplantation are based on this prognosis, which can have far-reaching consequences for the patient. For this reason IIP is a typical area of disease in which the method we propose in this paper could have added value as compared to the standard Cox model.

3.1 Data

For this evaluation study we used prospectively collected data of patients with IIP, collected between November 1993 and December 2005 at the Department of Pulmonology of the Academic Medical Center in Amsterdam, The Netherlands. Data collection was performed as part of a local protocol for examination and treatment for IIP patients eligible for lung transplantation. The variables used in this study are histopathologic pattern, age, sex, and eight pulmonary function testing (PFT) variables. All of these variables have been identified as predictors of IIP survival in several clinical studies (see e.g., [11]). The dataset we used in this study contained information on 103 patients. Median survival of the patients in the dataset was 47.9 months, and the 5-year survival rate was 35.7%.

3.2 Evaluation of Predictions

The survival probabilities $\hat{S}_i^{\mathrm{Cox}}(t)$ and $\hat{S}_i^{k\mathrm{NN}}(t)$, derived from both the Cox regression model and the k-NN algorithm, provide estimates for survival for each individual i at each observed time point t. The accuracy of these predictions is assessed by comparing the observed individual vital status at each time point $Y_i(t)$ with the predicted survival probabilities $\hat{S}_i(t)$. The means of the squared differences are a measure of prediction error, ranging from 0 to 1, also known as the quadratic or Brier score [12]. The lower the Brier score, the more accurate the prediction. In order to compensate for the loss of information due to censoring, we used an adjusted version of the Brier score $BS_i^c(t)$ as proposed by Graf et al [13], which uses a reweighing of the individual contributions based on their probability of being censored.

With $BS_i^c(t)$ the Brier score for each time point and for each individual is calculated. We used three types of cumulative scores for the comparison of the accuracy of the predictions of both methods. First, BS_i^c describes the cumulative prediction error per patient. Second, $BS^c(t)$ cumulates over individuals for each time point. Third, by cumulating over all individuals and all time points the total cumulative prediction error BS^c is calculated. Another measure of accuracy we used is the area under the receiver-operating characteristic curve (AUC) [14] for specific time points. It represents the probability that a patient who lived at the given time point is assigned a higher survival probability than a patient who then had died. An AUC of 0.5 indicates that the predictions are not better than chance, and is generally found for nondiscriminative models such as the Kaplan-Meier estimate.

3.3 Design

In this section the design of our experiment is outlined. The goal was to measure the performance of both methods in making individual survival predictions. Our strategy to achieve this is summarized in Fig.1. In order to reduce the risk of overfitting, performance was measured in a 10-fold cross-validation setting. Figure 1 represents the operation for a single individual in a single fold.

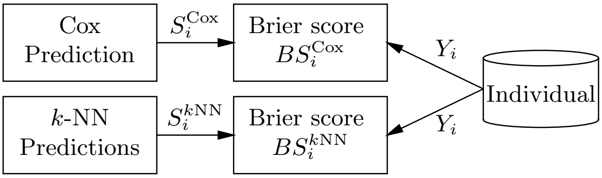


Fig. 1. Measuring the performance of the Cox regression model and the k-NN algorithm. The performance of both methods is measured by means of the Brier score, which compares the estimated probability of survival S_i to the observed vital status Y_i of the individual for each time point.

Calculating Cox- and k-NN Predictions. The Cox predictions and the k-NN predictions were obtained according to the procedures described in Sect. 2. For each of the folds in the cross-validation the following procedure was applied.

Based on the data in the training set a standard univariate Cox regression analysis was used to identify prognostic variables related to survival. Significant variables (p-value<0.05) were entered into a multivariate Cox model. Feature subset selection was performed using a backward stepwise selection method based on exact Akaike's Information Criterion (AIC) [3]. The model with the lowest AIC was considered the final model. From this model, the baseline hazard function and the linear predictor function were extracted.

Subsequently, predictions were made for each patient in the test set. The survival probability function based on Cox regression, $\hat{S}_i^{\text{Cox}}(t)$, was determined by calculating the score $\eta(\mathbf{x}_i)$ for each of the individuals in the test set based on their covariate pattern $\mathbf{x}_i$ and using this in (2). To estimate the survival probability function based on k-NN regression, the score $\eta(\mathbf{x}_i)$ was calculated for each of the individuals in the test set *and* in the training set. Subsequently, $\hat{S}_i^{k\text{NN}}(t)$ for each individual in the test set was calculated by determining the k nearest neighbors from the training set (using the distance function in (4)), and using their survival times to construct a Kaplan-Meier estimate. This procedure was repeated for different values of k, resulting in multiple predictions of $\hat{S}_i^{k\text{NN}}(t)$.

Calculating the Prediction Errors. The error in the predictions obtained by both methods is estimated for each individual by means of the Brier scores $BS_i^{c^{\text{Cox}}}(t)$ and $BS_i^{c^{k\text{NN}}}(t)$. We calculated the mean, standard deviation, and median of all BS_i^cs. To evaluate and display the predictive accuracy of the different methods, error curves were plotted using $BS^{c^{\text{Cox}}}(t)$ and $BS^{c^{k\text{NN}}}(t)$. In interpreting these error curves we focused on the first part of the curves, as the confidence intervals (not shown in the figure for clarity) rapidly increased for higher values of t. To compare the performance of the Cox model and the k-NN algorithm with that of a model that does not include information on the covariate pattern, we used the performance of the Kaplan-Meier estimate (1) as a benchmark value.

3.4 Results

All results are calculated over the 10 cross-validation folds. The mean, standard deviation and median of the cumulative Brier scores are shown in Table 1. The large standard deviations suggest an unbalanced distribution of BS_i^c. Therefore the median is used for comparison. We first compared the performance of the k-NN predictor for different values of k. The lowest median BS_i^c is found for k=20.

In addition, we plotted the prediction error curves over time for the survival predictions based on various neighborhood sizes, as shown in Fig. 2. The prediction error for k=5 is markedly higher over the entire time period than for the other values of k. The differences between 20, 25 and 40 as values of k are less distinct. For approximately the first 50 months, the survival predictions based on k=25 seem to have the lowest prediction error. In the period between 50 and

Table 1. Cumulative Brier scores for the Kaplan-Meier (KM, $\hat{S}_0(t)$), Cox regression ($\hat{S}_i^{\mathrm{Cox}}(t)$), and k-NN method ($\hat{S}_i^{k\mathrm{NN}}(t)$) with different numbers of neighbors k

	KM	Cox	k-NN									
			k=5	k=10	k=15	k=20	k=25	k=30	k=35	k=40	k=50	k=70
Mean	19.9	22.4	30.1	21.3	21.0	20.4	20.4	20.0	19.8	19.1	19.2	19.6
s.d.	21.0	25.1	38.9	24.2	23.0	21.9	21.5	20.9	20.3	19.5	19.7	20.6
Median	18.7	14.8	16.2	13.7	12.6	12.2	13.8	13.6	14.3	15.4	16.4	18.4

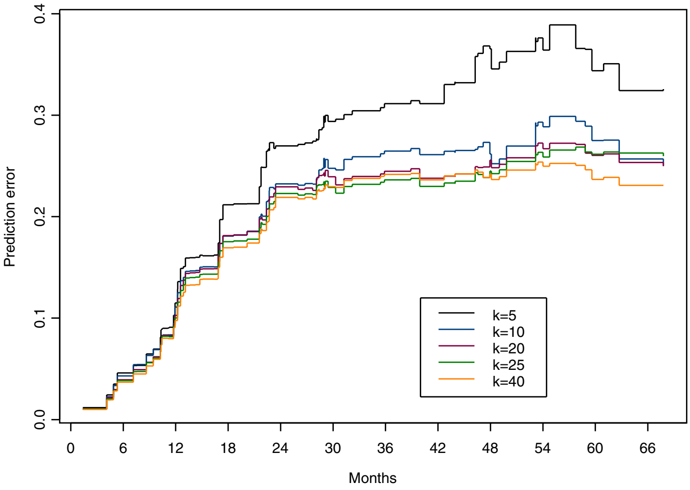


Fig. 2. Estimated prediction error curves for the k-NN algorithm with different numbers of neighbors k for the first 72 months

70 months, this is true for k=40, whereas the prediction error curves of the different values of k after 70 months intersect multiple times due to the increased confidence interval. Based on the median BS_i^c and the error curves, we choose to use 20 as the value of k in the comparison of the different prediction methods.

Figure 3 visualizes the differences between the estimated prediction error over time for the two methods. For purposes of comparison, we have also depicted the performance of the Kaplan-Meier estimate (population average survival). For predictions of survival up to approximately two years, the curves yield a similar pattern, with a slightly better performance of the k-NN algorithm. In the period between approximately 30 and 50 months the k-NN algorithm yields the lowest estimated prediction errors.

We calculated the median of the cumulative Brier scores up to t=12, 24 and 48 months and the AUC at the same time points. As shown in Table 2, the k-NN algorithm yields the lowest median prediction error for each period. The AUC

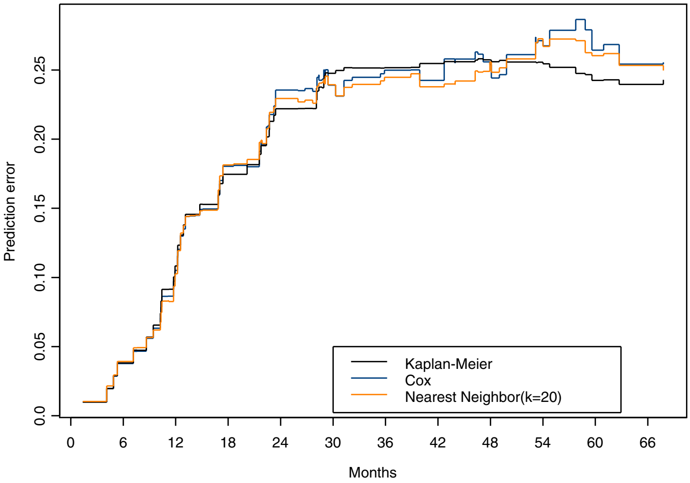


Fig. 3. Estimated prediction error curves for the Cox regression model and the k-NN algorithm with k=20 neighbors for the first 72 months. The error curve of the Kaplan-Meier estimate serves as a benchmark.

Table 2. Median cumulative Brier score per individual (BS^c) and area under the receiver-operating characteristic curve (AUC) for the Kaplan-Meier estimate, the Cox regression model, and the k-NN algorithm for t=12, 24 and 48 months

	12 mo.		24 mo.		60 mo.	
	BS^c	AUC	BS^c	AUC	BS^c	AUC
KM	0.04	0.49	0.77	0.53	11.47	0.46
Cox	0.03	0.61	0.73	0.61	11.73	0.62
20-NN	0.01	0.62	0.63	0.61	8.57	0.59

of the k-NN algorithm is almost equal to that of the Cox regression model at t=12 and 24 months, but lower at t=60 months.

4 Discussion and Conclusions

In this paper we proposed a prognostic inference method that combines the k-NN paradigm with Cox regression. It yields probabilistic survival predictions for individual patients that are based on small sets of similar patients that were seen before by the doctor. The intended application of the method is a prognostic case-retrieval system, but other types of application are conceivable.

From a theoretical point of view, our prediction method is more flexible than the Cox regression model because it does not need the proportional hazards assumption. In some situations covariate patterns are associated with survival curves that are fundamentally different in shape from the population-based survival curve. In the IIP domain, for instance, it is known that differences in histopathologic pattern do not only influence the steepness of the survival curve, but also its shape [15]. However, as all k-NN algorithms, our method relies on statistical estimates from small samples (neighborhoods), and this may cause high variation (and therefore low reliability) of predictions.

In a case study on IIP we compared the predictive performance of this method with the commonly used Cox regression model. On the given dataset, our method has the best predictive performance with k=20. With this number of neighbors, k-NN survival curves were largely similar to those obtained with the Cox model, and the median cumulative Brier score was even slightly lower, indicating superior performance in most cases. However, both methods also performed only slightly better than the population-based Kaplan-Meier survival estimate (in terms of the cumulative Brier score), and had but moderate discriminative abilities (as measured by the AUC). These findings impel to further investigations on survival prediction in this domain.

In the literature, k-NN methods have been used for a variety of prediction tasks, but rarely in the context of survival analysis. Hamilton et al. [16] used a k-NN algorithm to predict survival time of patients with colorectal cancer. Their algorithm returns the median survival time of four closest neighbors, based on a distance measure for three predefined variables. Anand et al. [17] enhanced several basic distance metrics by statistical techniques and used a framework based on Dempster-Shafer's Theory of Evidence to make survival predictions. Although the idea of embedding case retrieval in a multimodal reasoning task in general is not new [18], the idea of combining Cox regression with k-NN regression to predict survival, as proposed in this study, is novel.

There are several limitations to our study. In particular, the design of our evaluation method in the case study is somewhat biased: ten values of k were mutually compared in a cross-validation design, and the best performer (k=20) was compared to the Cox regression model within the same cross-validation loop. This design puts the Cox regression model at a slight handicap. We nevertheless believe that the two methods are competitive on the given dataset. Furthermore, the predictions from both methods could probably be improved by including additional information, e.g., on trends in physiological-respiratory variables [15,19].

The number of k=20 neighbors is quite large for a prognostic case-retrieval system, the intended application of our method. Clinical visits are restricted in time, and discussing 20 historical cases is not feasible for doctor and patient, even if they fancied doing so. Unfortunately, smaller numbers of k (e.g., 5 and 10) produce markedly worse predictions, and should therefore be avoided. We suggest that further research investigates how the information from some 20 cases can be conveniently presented to users of a system. Additional possibilities are to weigh cases according to their distance, and restrict the maximum distance to the query instance, as many search engines on the internet do.

We conclude that our k-NN method can compete in performance with a standard Cox regression model in predicting individual survival for IIP patients, with moderately-sized neighborhoods. It provides the starting point for a prognostic case-retrieval system.

Acknowledgements. We would like to thank Judith Blumenthal for her work on the IIP database.

References

1. Kaplan, E., Meier, P.: Nonparametric estimation from incomplete observations. JASA 53, 457–481 (1958)
2. Cox, D.R.: Regression models and life tables. J Royal Stat Soc Series B 34, 187–220 (1972)
3. Akaike, H.: A new look at the statistical model identification. IEEE Trans Auto Control AC-19, 716–723 (1974)
4. Therneau, T., Grambsch, P.: Modeling Survival Data: Extending the Cox Model. Springer, New York (2000)
5. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans Inform Theory 13, 21–27 (1967)
6. Schmidt, R., Montani, S., Bellazzi, R., Portinale, L., Gierl, L.: Cased-Based Reasoning for medical knowledge-based systems. Int J Med Inform 64, 355–367 (2001)
7. Bichindaritz, I., Marling, C.: Case-based reasoning in the health sciences: Whats next? Artif Intell Med 36, 127–135 (2006)
8. Duda, R., Hart, P.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
9. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. In: Data Mining, Inference, and Prediction, Springer, New York (2001)
10. American Thoracic Society/European Respiratory Society: International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. Am J Respir Crit Care Med 165(2), 277–304 (2002)
11. Perez, A., Rogers, R.M., Dauber, J.H.: The prognosis of idiopathic pulmonary fibrosis. Am J Respir Cell Mol Biol 129(Suppl. 3), 19–26 (2003)
12. Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly Weather Rev 78, 1–3 (1950)
13. Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M.: Assessment and comparison of prognostic classification schemes for survival data. Stat Med 18, 2529–2545 (1999)
14. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1), 29–36 (1982)
15. Latsi, P.I., du Bois, R.M., Nicholson, A.G., et al.: Fibrotic idiopathic interstitial pneumonia: the prognostic value of longitudinal functional trends. Am J Respir Crit Care Med 168(5), 531–537 (2003)
16. Hamilton, P.W., Bartels, P.H., Anderson, N., Thompson, D., Montironi, R., Sloan, J.M.: Case-based prediction of survival in colorectal cancer patients. Anal Quant Cytol Histol 21(4), 283–291 (1999)
17. Anand, S.S., Hamilton, P.W., Hughes, J.G., Bell, D.A.: On prognostic models, artificial intelligence and censored observations. Meth Inf Med 40(1), 18–24 (2001)
18. Aha, D., Daniels, J.J. (eds.): Case Based Reasoning Integrations: Papers from the 1998 Workshop (Technical Report, WS-98-15). AAAI Press, Menlo Park, CA (1998)
19. Collard, H.R., King Jr, T.E., Bartelson, B.B., et al.: Changes in clinical and physiologic variables predict survival in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med 168(5), 538–542 (2003)

Using Temporal Context-Specific Independence Information in the Exploratory Analysis of Disease Processes

Stefan Visscher[1], Peter Lucas[2], Ildikó Flesch[2], and Karin Schurink[1]

[1] Department of Internal Medicine and Infectious Diseases,
University Medical Center Utrecht, The Netherlands
{S.Visscher,K.Schurink}@umcutrecht.nl

[2] Institute for Computing and Information Sciences,
Radboud University Nijmegen, The Netherlands
{peterl,ildiko}@cs.ru.nl

Abstract. Disease processes in patients are temporal in nature and involve uncertainty. It is necessary to gain insight into these processes when aiming at improving the diagnosis, treatment and prognosis of disease in patients. One way to achieve these aims is by explicitly modelling disease processes; several researchers have advocated the use of dynamic Bayesian networks for this purpose because of the versatility and expressiveness of this time-oriented probabilistic formalism. In the research described in this paper, we investigate the role of context-specific independence information in modelling the evolution of disease. The hypothesis tested was that within similar populations of patients differences in the learnt structure of a dynamic Bayesian network may result, depending on whether or not patients have a particular disease. This is an example of temporal context-specific independence information. We have tested and confirmed this hypothesis using a constraint-based Bayesian network structure learning algorithm which supports incorporating background knowledge into the learning process. Clinical data of mechanically-ventilated ICU patients, some of whom developed ventilator-associated pneumonia, were used for that purpose.

1 Introduction

Bayesian networks are known to yield representations that are well suited as a basis for medical decision making [1]. Reasoning with a Bayesian network, which is done by filling in data of a patient into the network and computing posterior probability distributions, often yields considerable insight into the disease process of a patient, as well as concerning the way the disease process can be influenced by the selection of appropriate treatment. However, knowledge of the temporal nature of a disease process may also be relevant in this respect, in which case *dynamic* Bayesian networks are often selected for the construction of models. Such models can be used for temporal reasoning in clinical decision-support systems, as the formalism takes into account the notion of time [2,3,4,5]. Bayesian networks that ignore the notion of time are called *static*.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 87–96, 2007.

So far, Bayesian networks have in particular been popular as models for uncertainty reasoning in clinical decision-support systems; they have been less popular as tools for the analysis of clinical data, despite the availability of a wide range of Bayesian network structure and parameter learning techniques [6]. This is somewhat surprising as the statistical nature of Bayesian networks would render them in principle as useful as data-analytical tools as, say, logistic regression, one of the main statistical tools of multivariate clinical data analysis.

We believe that the reasons why Bayesian networks, both static and dynamic, are used so rarely for data analysis in medicine are threefold: (1) in particular static Bayesian networks are difficult to interpret, as the direction of the arcs is often counterintuitive; (2) whereas dynamic Bayesian networks have the advantage that the direction of some of the arcs is in accordance to the order of time, their structure is usually restricted to being *repetitive* [7], which may not be compatible with the clinical problem at hand; (3) the conditional independences modelled in Bayesian networks only concern random variables and not their individual values; however, in medicine it is often the *context*, i.e., the specific values random variables take, that determine how things relate to each other. Context-specific independences and dependences can be modelled by extensions to Bayesian networks, such as by *multinet* models [8].

In this paper, we demonstrate that by using non-repetitive dynamic Bayesian multi-networks in conjunction with context-specific independence information, an analytical tool results that does indeed yield insight into the evolution of a disease, in comparison to other diseases in a related population of patients. We exploit a constraint-based learning algorithm for that purpose, as these structure learning algorithms allow for the easy incorporation of medical background knowledge in the learning process. The ideas are illustrated by the analysis of temporal data of patients in the Intensive Care Unit (ICU), who either have developed ventilator-associated pneumonia, or VAP for short, and ICU patients without VAP. As only some 10-15% of ICU patients will develop VAP, it was also necessary to exploit background knowledge in the learning process, as sometimes clinically obvious relationships cannot be learnt from the data due to the sparsity of data for a particular type of patient.

The paper is organised as follows. In Section 2, Bayesian networks, dynamic Bayesian networks and context-specific independence are briefly reviewed. Next, in Section 3, the basic theory underlying constraint-based structure learning is reviewed. Finally, in Section 4 we discuss the results achieved. The paper is rounded-off with some conclusions in Section 5.

2 Preliminaries

We briefly review the theory dynamic Bayesian networks, as discussed in more detail in [7]. Furthermore, the medical domain of ventilator-associated pneumonia, is described.

2.1 Dynamic Bayesian Networks

A *Bayesian network* $\mathcal{B} = (G, P)$, BN for short, is a joint probability distribution P of a set of random variables X with an associated acyclic directed graph $G = (V, A)$, where P is assumed to be decomposed into a set of conditional probability distributions in accordance to the structure of G. The random variables X and the vertices V have a 1–1 correspondence; thus we sometimes write X_W, $W \subseteq V$, for the random variables corresponding to the vertices W. Finally, $\mathit{dom}(X)$ denotes the domain of the set of random variables X (a Cartesian product).

Dynamic Bayesian networks (DBNs) are an extension of ordinary Bayesian networks and allow for modelling uncertainty involved in processes regarding the dimension of time. Usually, a DBN is described in terms of a timeslice that has a fixed structure and is repeated several times, i.e., the DBN has a *repetitive* structure [9]. We, however, are convinced that disease processes are more complicated than that in the sense that independences may change over time and, therefore, a repetitive DBN would not suffice in every domain. This motivated some of us to develop a theory of modularisation of DBNs, with both repetitive and non-repetitive DBNs as special cases [7]. Evidence of the practical usefulness of non-repetitive DBNs has also come from work by Tucker et al. [10].

For the formal representation of the uncertain relations between variables over time, we need the following notions. Let T denote the (discrete and finite) time axis. Independence relationships between random variables with the *same* time point t are represented by means of an acyclic directed graph (ADG) $G_t = (V_t, A_t^a)$, called a *timeslice*, with V_t denoting a set of vertices and $A_t^a \subseteq V_t \times V_t$ a set of *atemporal arcs*. Between timeslices, vertices corresponding to random variables may be linked to each other by means of so-called *temporal arcs*. Thus, a DBN consists of two parts: (1) an atemporal part (the timeslices), and (2) a temporal part. First, we consider the atemporal part.

Definition 1. ***(timeslice and atemporal arcs)*** *An ADG $G_t = (V_t, A_t^a)$, with the set of vertices V_t and the set of* atemporal arcs *$A_t^a \subseteq V_t \times V_t$, $t \in T$, is called a* timeslice *at time point t.*

The set of all timeslices G of a DBN is taken as:

$$G = \{G_t \mid t \in T\} = \{(V_t, A_t^a) \mid t \in T\} = (V_T, A_T^a). \tag{1}$$

Let G_t and $G_{t'}$, $t, t' \in T$, be two timeslices. Then, an arc $(u_t, v_{t'})$ with $t < t'$ is called a *temporal arc*. The set of temporal arcs of an ADG is denoted by A^t. Thus, temporal arcs connect timeslices with strict direction from the past to the future.

Definition 2. ***(temporal network)*** *A* temporal network *N is defined as a pair $N = (V_T, A)$, where $G = (V_T, A_T^a)$ and $A = A_T^a \cup A^t$, with A_T^a denoting the set of timeslices.*

Clearly, a temporal network N is also an ADG. A *dynamic Bayesian network* (DBN) is now defined as a pair $\mathcal{DBN} = (N, P)$, where P is the joint probability distribution (JPD) on X_{V_T}.

2.2 Context-Specific Independences

Two sets of random variables X and Y are said to be *conditionally independent* given a third set of random variables Z, denoted by $X \perp\!\!\!\perp_P Y \mid Z$, if it holds that

$$P(X \mid Y, Z) = P(X \mid Z)$$

if $P(Y, Z) > 0$. Such conditional independence statements cannot only be represented in the form of probability distributions P; they can also be read-off from the graphical structure of an associated ADG G using the notion of d-separation. Then, two disjoint sets of vertices A and B in G are said to be *d-separated* given a third disjoint set of vertices C, denoted by $A \perp\!\!\!\perp_G B \mid C$, if each (undirected) path from a vertex in A to a vertex in B is blocked by a vertex in C, taking into account paths with so-called v-structures (i.e., subgraphs of the form $\rightarrow \cdot \leftarrow$).

For Bayesian networks $\mathcal{B} = (G, P)$, it holds that if $A \perp\!\!\!\perp_G B \mid C$ holds, then $X_A \perp\!\!\!\perp_P X_B \mid X_C$ should also be satisfied. It is said that G is an *independence map* of P. Similar, temporal and atemporal, notions of d-separation have been developed for dynamic Bayesian networks, where the *atemporal d-separation* relationship $\perp\!\!\!\perp_G$ is defined for the part of the dynamic Bayesian network where the temporal arcs are ignored, and *temporal d-separation*, denoted by $\perp\!\!\!\perp_{N_{|\Theta}}$, is defined by always taking into account at least one temporal arc when investigating blockage (for details, cf. [7]). Clearly, atemporal d-separation $\perp\!\!\!\perp_G$ can be defined in terms of atemporal d-separation for individual timeslices, i.e., in terms of $\perp\!\!\!\perp_{G_t}$, $t \in T$.

Despite the fact that temporal and atemporal notions of d-separation allow for the study of interesting independence patterns in dynamic Bayesian networks, we believe that many of these patterns are context specific, i.e., independence information may change for particular values of random variables. Formally, a set of variables Y is *conditionally context-specific independent* of a set of variables W given a third set Z in the context φ, written $Y \perp\!\!\!\perp_P W \mid Z; \varphi$, where φ is a nonempty set of random variables U with values u, i.e., $\varphi \equiv U = u$, if $P(Y \mid W, Z, \varphi) = P(Y \mid Z, \varphi)$ and $P(Y \mid W, Z, \varphi') \neq P(Y \mid Z, \varphi')$ for $\varphi' \equiv U = u'$, $u' \neq u$ [11]. For discrete random variables X with finite domain, it is possible to associate an ADG G^{φ} with every context φ. The result is called a *Bayesian multinetwork* $\mathcal{B} = (G, P)$ with $G = \{G^{\varphi} \mid \varphi \equiv X = x, x \in \mathit{dom}(X)\}$. Dynamic Bayesian multinetworks can be be defined along similar lines.

2.3 Ventilator-Associated Pneumonia

Ventilator-associated pneumonia (VAP) occurs in mechanically-ventilated ICU patients. Clinical symptoms, such as fever, indicating that this bacterial infection is present or developing, are usually not very specific. Important symptoms and signs, providing evidence for the development of VAP, include *body temperature*, amount and colour of *sputum*, radiological *signs* on the chest X-ray, duration of *mechanical ventilation*, number of *leukocytes* [12], and abnormal ratio between the arterial oxygen pressure and the fractional inspired oxygen level

(pO_2/FiO_2-ratio). Some of these signs and symptoms, such as fever and number of leukocytes, are due to the fact that VAP is an infectious disease, whereas others, such as increased amount of sputum, abnormal chest X-ray and changed $\mathrm{pO_2/FiO_2}$-ratio are due to the pulmonary location of the infection.

3 Constraint-Based Structure Learning

As only 10-15% of the ICU patients develop VAP, it was unlikely that we would have been able to collect sufficient amount of data for patient with VAP, despite the fact that the amount of data collected for the entire ICU population was large. For situations where data are sparse, it is normally difficult to learn independence relations from the data. However, lack of data can be compensated, in principle, by augmenting the learning process through the exploitation of background knowledge. This is exactly what we have done. Learning algorithms that allow easy incorporation of background knowledge into the learning process are called *constraint based*. These algorithms derive a set of conditional independence statements from the data, taking supplied dependence and independence information as additional constraints, and build a structure with d-separation properties corresponding to the independence information available.

3.1 The NPC Algorithm

One of the best constraint-based Bayesian network structure learning algorithms available is the NPC algorithm. NPC stands for 'Necessary Path Condition'; it is a criterion that has been added to an earlier constraint-based algorithm, PC, by researchers at Siemens in Munich [13]. The algorithm is a variant of the CI algorithm by Verma and Pearl [14], and works as follows:

1. **Automatic phase:**
   (a) An undirected graph $H = (V, E)$, called *skeleton*, is derived through computation of the score $g_{\chi^2,\alpha}(X, Y, S)$, for pairs X, Y of random variables and the set of random variables S (with $X, Y \notin S$). The function $g_{\chi^2,\alpha}$ is based on the χ^2 test with significance level α [13]. Typically, one takes $\alpha \leq 0.05$. If $g_{\chi^2,\alpha}(X, Y, S) > 0$ then the conditional independence hypothesis $X \perp\!\!\!\perp_P Y \mid S$ is rejected.
   (b) Modify subgraphs $X - Y - Z$ of H into $X \rightarrow Y \leftarrow Z$, if $X - Z \notin E$ and $X \not\perp\!\!\!\perp_P Z \mid S$, with $Y \in S$, using the same scoring function $g_{\chi^2,\alpha}$.
   (c) Orientate the remaining lines as to obtain arcs, where the creation of cycles in the resulting directed graph is avoided.
2. **User interaction phase:** To resolve inconsistencies in the conditional (in)dependence statements, the NPC algorithm, unlike the PC algorithm where in case of uncertain dependences directionality is chosen randomly, relies on user interaction where the user gets the opportunity to decide on the addition, removal and orientation of arcs. In addition, α can be arbitrarily chosen, so that lines with calculated p-value (called p below) larger than α are excluded.

In our domain, the direction of an arc has been determined by the use of background knowledge. By doing so, cause and effect can be distinguished. For example, when the NPC algorithm indicates a relation between 'VAP' and 'temperature', the most logical order is that 'VAP' should be parent of 'temperature', as when a patient suffers from VAP, normally, the temperature increases due to fever, and not the other way round. Also, it is possible to supply known relations at the start of the NPC learning process. By doing so, we were able to include constraints that have already been proved to exist and have been described in literature. We used the implementation of the NPC algorithm available in the Hugin tool set [15] for our research. This includes the EM algorithm for parameter learning.

3.2 Data

A dataset D with temporal data of ICU patients, containing 17710 records, was used. Each record represents data of one patient in the ICU during a period of 24 hours. The database contains 2424 admissions to the ICU. For 157 of these patient episodes, VAP was diagnosed by two infectious-disease specialists. From dataset D three subsets were created: D_{vap} containing data of all 157 VAP patients; $D_{\overline{\mathrm{vap}}}$ containing data of patients who were not diagnosed with VAP (so-called controls). Each patient with VAP was matched to 3 control patients, with a similar duration of mechanical ventilation on the days of matching. D_{VAP} contained data of both VAP and control patients, i.e., $D_{\mathrm{VAP}} = D_{\mathrm{vap}} \cup D_{\overline{\mathrm{vap}}}$. Each dataset contained data of 4 consecutive days, each representing a timeslice: t_0 was either the day on which VAP was diagnosed or the day of matching and t_{-3}, t_{-2}, t_{-1} were the three days preceding t_0.

3.3 Procedure of DBN Construction

The construction of the context-specific and combined DBNs was performed as follows:

1. Using the NPC algorithm, under the first author's supervision, the atemporal arcs between vertices in each separate timeslice were determined.
2. In the next run of the NPC algorithm, the temporal relationships of all variables were explored, taking into account the structure of the timeslices.
3. Medical background knowledge was sometimes employed to decide about the direction of arcs, or inclusion or deletion of arcs. However, this was only employed when the algorithm was unable to decide about the inclusion and direction of an arc. The direction of arcs was decided on by looking at the network in terms of cause–effect relationships. Expertise of a medical infectious disease specialist was used for that purpose.

4 Results

Based on the three databases, D_{vap}, $D_{\overline{\mathrm{vap}}}$ and D_{VAP}, three DBNs were constructed using the NPC learning algorithm. As described above, atemporal subgraphs were obtained separately, and then combined to a DBN by learning temporal arcs

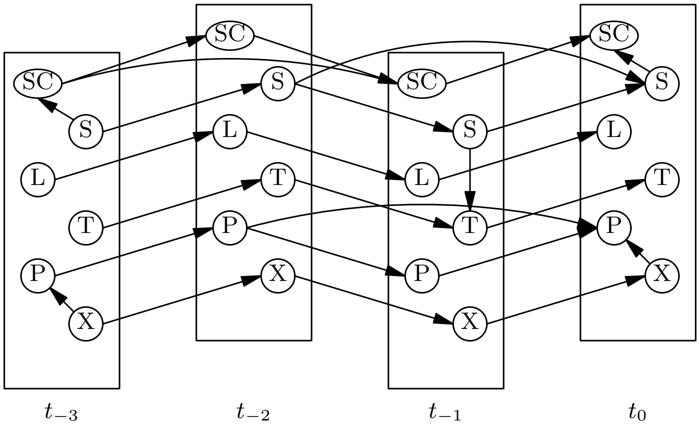


Fig. 1. G_{vap} : Independences obtained for VAP patients. Abbreviations: SC: sputum colour; S: sputum; L: leukocytosis; T: temperature; P: pO_2/FiO_2; X: chest X-ray.

(taking into account the known timeslice structures). Sometimes uncertain arcs were removed (user interaction phase). For example, an arc between the variable *temperature* in timeslice t_0 and variable *leukocytosis* in timeslice t_{-2}, did not seem clinically relevant and was, therefore, excluded.

4.1 VAP Patients (D_{vap})

The timeslices (atemporal subgraphs) for the four different time points show different independences. For example, for t_{-3} an arc between *sputum* and *sputum-colour* ($p = 0.05$) was suggested by the NPC algorithm, whereas for t_{-1} and for t_{-2} that same relation was absent, but for t_0 it was again present. Also, an arc ($p = 0.02$) between *chest X-ray* and *pO_2/FiO_2* (as explained, a measurement of the lungs' functions) was often found, as well as between *temperature* and *sputum-colour* ($p = 0.05$). As the last arc was not considered clinically relevant, it was excluded from the models. All temporal arcs proved to have high significance ($p < 10^{-7}$). Combining the atemporal en temporal parts resulted in a DBN, called G_{vap}, shown in Fig. 1, that includes all signs and symptoms describing the course of the development of VAP.

4.2 Patients Not Diagnosed with VAP ($D_{\overline{\text{vap}}}$)

The timeslices for the non-VAP patients were similar, but not identical, to those for VAP patients; in particular, arcs between *chest X-ray* and *pO_2/FiO_2*, and between *sputum* and *sputum-colour* were found. The only difference was that the strength of the arcs increased towards the time point matching the day of VAP, i.e., t_0, that is, $p(t_{-3}) \approx 10^{-2}$, $p(t_{-2}) \approx 10^{-3}$, $p(t_{-1}) \approx 10^{-4}$ and $p(t_0) \approx 10^{-5}$. Thus, $p(t_{-3}) > p(t_{-2}) > p(t_{-1}) > p(t_0)$. Temporal arcs were suggested between timeslices t_{-2} and t_{-1} for the variables *chest X-ray*, *sputum* and *pO_2/FiO_2*

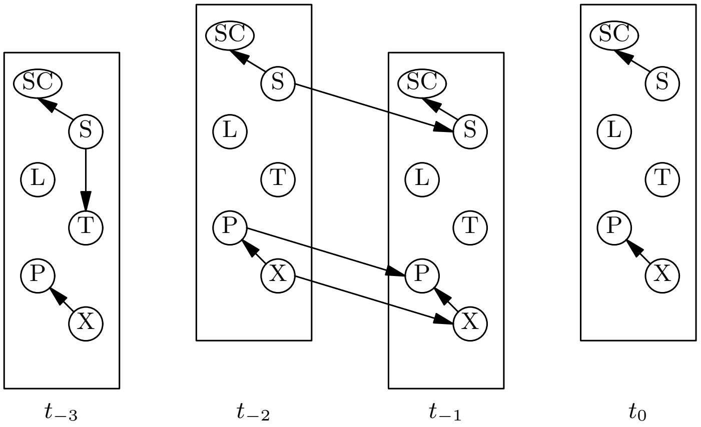


Fig. 2. $G_{\overline{vap}}$: Independences obtained for patients *not* diagnosed with VAP. Abbreviations: SC: sputum colour; S: sputum; L: leukocytosis; T: temperature; P: pO_2/FiO_2; X: chest X-ray.

only. Moreover, these temporal relations proved to be less strong, i.e., $p \approx 0.01$, compared to the temporal relations in the context of VAP. Combining both temporal and atemporal structures resulted in DBN called $G_{\overline{vap}}$, shown in Fig. 2 with again, $\overline{vap}$ representing the matched control patients.

4.3 Patients With and Without VAP ($D_{\mathbf{VAP}}$)

As a model only suitable for VAP patients would not be useful in practice, combining the two datasets mentioned above for building a DBN for VAP and

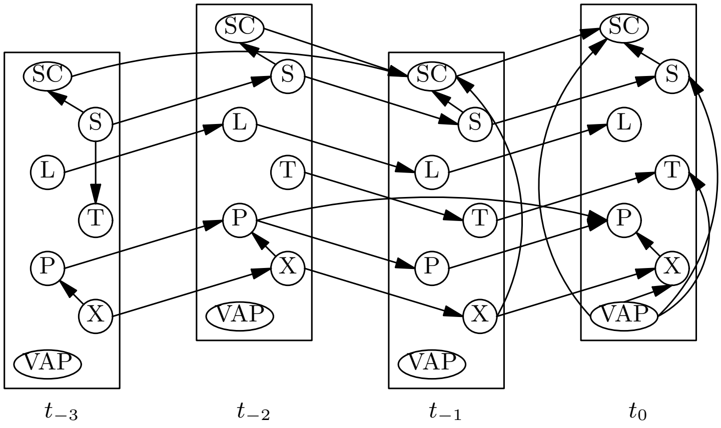


Fig. 3. G_{VAP} : Independence model for variables (excluding context-specific independences). Abbreviations: SC: sputum colour; S: sputum; L: leukocytosis; T: temperature; P: pO_2/FiO_2; X: chest X-ray; VAP: ventilator-associated pneumonia.

non-VAP at the same time yields yet another view on structure learning. Structure learning based on the database including data of VAP as well as non-VAP patients resulted in a combination of the two DBNs G_{vap} and $G_{\overline{vap}}$, from here denoted by G_{VAP}. The temporal arcs were almost identical to those of G_{vap}, though less strong ($p \approx 10^{-4}$). The atemporal arcs had strong correlations and were, not surprisingly, found between the variables *chest X-ray* and pO_2/FiO_2 ($p \approx 10^{-3}$) and between *sputum* and *sputum colour* ($p \approx 10^{-3}$). In all, the temporal arcs again proved to be stronger than the atemporal arcs. The resulting model is shown in Fig. 3. This DBN clearly shows that much of the clarity of the original context-specific DBNs was lost, and that it is no longer possible to gain insight into the development of VAP and non-VAP separately.

5 Conclusions and Discussion

The hypothesis underlying the research described in this paper was that in order to obtain insight into the evolution of disease processes, it is not merely necessary to explicitly model time, but also to consider context-specific independence information. The results we have obtained confirm this hypothesis, and to the best of our knowledge, this is the first paper combining context-specific independence and dynamic Bayesian networks.

The NPC learning algorithm proved to be useful, as it allowed for the incorporation of background knowledge, without which it would have been difficult to obtain clinically meaningful results. This algorithm combines the virtues of offering the capability of automatic learning of independence information from data, whereas uncertainty regarding both the presence of dependences and the directionality of arcs can be resolved by the user. Thus, the algorithm offers a natural role for the incorporation of expert background knowledge in the learning process.

The results obtained for the ICU domain show that signs and symptoms of patients known to develop VAP proved to have strong temporal relationships, whereas the temporal relationships between the signs and symptoms of patients not diagnosed with VAP were very weak. The combined model G_{VAP} included independences from both the VAP and non-VAP models. However, the model was not merely a union of the two independence sets.

When comparing the atemporal parts of the networks, change in the independence information as a function of time was observed. This strengthens our belief that when constructing a dynamic Bayesian model, it is not sufficient to resort to repetitive DBNs if one wishes to obtain models that capture the characteristics of the domain; rather, non-repetitive DBNs should be investigated as well [7]. Further research is needed to evaluate our findings regarding the temporal behaviour of both models for VAP patients and non-VAP patients. It would, for example, be interesting to investigate the predictive value of an increasing body temperature of an ICU patient in relationship to the development of VAP. Moreover, a more detailed comparison of the ADGs of VAP and non-VAP for larger datasets may give more insight into the course of the disease process of VAP.

In conclusion, the combination of a general theory of DBNs, where repetitive and non-repetitive DBNs are both special cases, with the exploitation of context-specific independence information proved to yield a powerful data-analysis tool.

Acknowledgements

This study was supported by a grant from NWO (project number 634.000.026).

References

1. Lucas, P.J.F., van der Gaag, L.C., Abu-Hanna, A.: Bayesian networks in biomedicine and health-care. Artificial Intelligence in Medicine 30(3), 201–214 (2004)
2. Augusto, J.C.: Temporal reasoning for decision support in medicine. Artificial Intelligence in Medicine 33(1), 1–24 (2005)
3. Dagum, P., Galper, A.: Forecasting sleep apnea with dynamic models. In: Proceedings of UAI-1994, pp. 64–71 (1994)
4. Adlassnig, K.-P., Combi, C., Das, A.K., Keravnou, E.T., Pozzi, G.: Temporal representation and reasoning in medicine: Research directions and challenges. Artificial Intelligence in Medicine 38(2), 101–113 (2006)
5. Provan, G., Clarke, J.R.: Dynamic network construction and updating techniques for the diagnosis of acute abdominal pain. PAMI 15(3), 299–307 (1993)
6. Friedman, N., Murphy, K., Russell, S.: Learning the structure of dynamic probabilistic networks. In: Proceedings of the 14th UAI, pp. 139–147 (1998)
7. Flesch, I., Lucas, P.J.F., Visscher, S.: On the modularisation of independence in dynamic Bayesian networks. In: Proceedings of BNAIC-2006, pp. 133–140 (2006)
8. Geiger, D., Heckerman, D.: Knowledge representation and inference in similarity networks and Bayesian multinets. Artificial Intelligence 82, 45–74 (1996)
9. Labatut, V., Pastor, J., Ruff, S., Démonet, J.F., Celsis, P.: Cerebral modeling and dynamic Bayesian networks. Artificial Intelligence in Medicine 39, 119–139 (2004)
10. Tucker, A., Liu, X.: Learning dynamic Bayesian networks from multivariate time series with changing dependencies. In: Proceedings of the 5th International Symposium on Intelligent Data Analysis, Berlin, Germany (2003)
11. Boutilier, C., Friedman, N., Goldszmidt, M., Koller, D.: Context-specific independence in Bayesian networks. In: Proceedings of UAI-1996, pp. 64–72 (1996)
12. Bonten, M.J.M., Kollef, M.H., Hall, J.B.: Risk factors for Ventilator-Associated Pneumonia: from epidemiology to patient management. Clinical Infectious Diseases 38(8) (2004)
13. Steck, H.: Constraint-Based Structural Learning in Bayesian Networks using Finite Data Sets. PhD thesis, Institut für Informatik der Technischen Universität München (1999)
14. Verma, T.S., Pearl, J.: Equivalence synthesis of causal models. In: Proceedings of UAI-1990, pp. 220–227. Morgan Kaufmann, San Francisco (1990)
15. Madsen, A.L., Lang, M., Kjærulff, U.B., Jensen, F.: The hugin tool for learning Bayesian networks. In: Proceedings of 7th ECSQARU, pp. 594–605 (2003), URL: http://www.hugin.com

Discovery and Integration of Organ-Failure Episodes in Mortality Prediction

Tudor Toma[1], Ameen Abu-Hanna[1], and Robert-Jan Bosman[2]

[1] Academic Medical Center, Universiteit van Amsterdam, Department of Medical Informatics, P.O. Box 22700, 1100 DE Amsterdam, The Netherlands
[2] Department of Intensive Care, Onze Lieve Vrouwe Gasthuis, 1e Oosterparkstraat 279, P.O. box 10550, 1090 HM Amsterdam, The Netherlands

Abstract. Current predictive models in the intensive care rely on summaries of data collected at patient admission. It has been shown recently that temporal patterns of the daily Sequential Organ Failure Assessment (SOFA) scores can improve predictions. However, the derangement of the six individual organ systems underlying the calculation of a SOFA score were not taken into account, thus impeding the understanding of their prognostic merits. In this paper we propose a method for model induction that integrates in a novel way the individual organ failure scores with SOFA scores. The integration of these two correlated components is achieved by summarizing the historic SOFA information and at the same time by capturing the evolution of individual organ system failure status. The method also explicitly avoids the collinearity problem among organ failure episodes. We report on the application of our method to a large dataset and demonstrate its added value. The ubiquity of severity scores and sub-scores in medicine renders our approach relevant to a wide range of medical domains.

Keywords: Prognostic models, temporal patterns, Intensive Care, organ failure scores.

1 Introduction

Probabilistic predictions of patient outcomes such as mortality and length of stay in the intensive care unit (ICU) are useful for supporting decisions at the level of individuals and groups [1]. Current models for predicting hospital mortality, after admission to the ICU, use summaries of patient information collected within the first 24 hours of admission. These summaries, which take the form of severity-of-illness-scores such as the APACHE-II [2] and SAPS-II [3], are used as covariates in a logistic regression model (see appendix).

Since a decade ago, some ICUs started collecting Sequential Organ Failure Assessment (SOFA) scores [4] *on each day* of ICU stay. A SOFA score is an integer ranging from 0 to 24 that quantifies the derangement of *all* organs of a patient on a given day, the higher the score the greater the derangement. A SOFA score is calculated as the sum of 6 individual organ system failure (IOSF) scores, each ranging between 0 and 4.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 97–106, 2007.

Although not specifically targeted towards prediction of mortality, the relationship between SOFA scores and mortality has been investigated. In previous work [5] we devised a new method for integrating the SOFA temporal information in the existing logistic regression models. The method, more elaborated on in the next section, is based on the idea of using frequent temporal patterns, called episodes, as covariates in the model. Although the SOFA episodes improved predictions, the use of only SOFA scores has two disadvantages. First, no insight is obtained into the qualitative contribution of the individual organ systems to mortality. Second, it is unclear whether the IOSF scores would further improve the quality of predictions because these scores are *correlated* with the SOFA scores and it is unclear how to combine the two.

In this paper we propose a method for model induction that incorporates IOSF scores alongside the SOFA scores. The method deals with the overlap between the two types of scores by summarizing the historic SOFA information in one summary statistic, and by capturing the evolution of individual organ system failure status in frequent temporal patterns. The summary statistic and the organ failure (OF) episodes are used as covariates in the familiar logistic regression framework. For a given day d, the application of the proposed method results in a model predicting, for patients staying at least d days, the probability of their eventual survival status at discharge (regardless of when this happens). We report on the application of our method to a large real-world dataset and demonstrate the added value in interpreting the models and in their improved predictive performance. In the sequel we will refer to a model using only the SAPS-II (in short SAPS) as the *static model*; to a model using SAPS and SOFA episodes as a *SOFA-model* (as described in [5]); and to a model using SAPS, a summary of SOFA and failure episodes as an organ-failure model (*OF-model*). The resulting OF-models will be subject to comparison with the other models.

The rest of the paper is organized as follows. Section 2 describes the proposed method to induce OF-models and the data types it operates on. Section 3 and Section 4 describe the case study used for demonstrating the method and the obtained results. We discuss our method in Section 5 and put it in perspective by relating it to other work.

2 Data and Methods

Data. We consider two categories of data: the static data, represented by the SAPS score (collected at admission) and temporal data consisting of the daily SOFA score along with its 6 components (IOSF scores) corresponding to the following systems: respiratory (Resp), coagulation (Coag), hepatic (Hepa), cardiovascular (Cardio), central nervous system (CNS), and renal (Ren) systems. Table 1 shows an example of data for a patient who stayed for 4 days in the ICU before dying on the fifth day.

Method. In previous work [5] we showed how to induce SOFA-models. In a nutshell, this is done by the following process. First the SOFA scores, ranging from 0 to 24, are categorized into three qualitative states: Low (L), Medium (M)

Table 1. Example of available temporal data for an ICU patient admitted for 4 days. The SOFA scores indicate a constant health status deterioration.

Day	SOFA	Resp	Coag	Hepa	Cardio	CNS	Ren	Outcome
1	10	4	2	0	0	1	3	
2	12	4	1	2	1	2	2	
3	14	4	2	2	0	4	2	
4	15	4	1	2	1	4	3	
5	–	–	–	–	–	–	–	died

and High (H). For each day d on which hospital mortality is to be predicted the subsample of patients that stayed at least d days is selected. Next, frequent episodes of consecutive SOFA states that are aligned with the day of prediction (later clarified in this paper) are discovered in these patients. The SAPS and a set of binary variables representing the occurrence of the SOFA episodes in patients are then considered as possible covariates (input variables) in a logistic regression model to predict the hospital mortality for day d. For example if the linear predictor LP (see appendix) of the model for day 5 is: $-2 + 0.02SAPS - 1.5LL + 0.7H$ then for a patient with SAPS of 40 having the episode $\{L, L\}$ at days 4 and 5 will be $-2 + 0.02 * 40 - 1.5 = -2.7$ which corresponds to a probability of dying of 0.063 while a patient with SAPS of 40 but having the episode $\{H\}$ on day 5 will have an LP of $-2 + 0.02 * 40 + 0.7 = -0.5$ which corresponds to a probability of death of 0.38. In this paper we adapt and extend our approach described above to induce OF-models. This process is described below followed by a description of the main differences between the new and previous approach.

Categorization. An IOSF score ranging between 0 to 4 is categorized based on clinical definitions into two categories: *failure* (F), for values $\in \{3, 4\}$ and *non–failure* (NF) otherwise. For example, the renal scores during 3 days of 1–4–2 become NF, F, NF. Aside from clinical interpretability, limiting the number of categories allows the emergence of episodes with higher support in the data.

Frequent episode discovery. We rely on the A-priori-like algorithm [6] described in [7] for frequent pattern discovery. This is based on the *downward closure* property which implies that a subsequence of a frequent episode must be frequent too. The discovery procedure is an iterative process. We adapted the algorithm to search for a special type of episodes: their occurrence in a patient's sequence of values is *consecutive* and also *aligned* to the day of prediction d. For example, given the patient's sequence F, F, NF, F, NF starting at admission day, then for d=2 the episode $\{F, F\}$ occurs in the patient data because aligning the episode at day 2 (i.e. positioning the last F in the episode at the second element in the patient's sequence) results in a match with the subsequence F, F in the patient sequence. However, for d=4 the episode is not aligned to the patient's sequence. The decision to use aligned episodes is motivated by the belief that the last days before prediction are more relevant than information at earlier days.

In each iteration, the algorithm extends frequent episodes from the previous iteration with elementary episodes (F and NF for the organ failure data) and assesses the frequency of the resulting episodes. For example given the frequent episode $\{F, F, NF\}$, the extended aligned candidate episodes are $\{F, F, F, NF\}$ and $\{NF, F, F, NF\}$. In general, an episode aligned to day d is said to be frequent when its frequency rate in the subset of patients staying at least d days exceeds a pre-specified threshold (e.g. 5% of cases) referred to as minimum support rate. The discovery process continues until no more frequent episodes are encountered.

Model fitting strategy. Not all the frequent OF episodes are relevant for prediction and their excessive use can lead to overfitting. Our feature selection strategy is based on an information-theoretic measure, the Akaike's Information Criterion (AIC) [8] used in an iterative backward variable elimination selection process. In every iteration, the current model with N variables is used to produce N models, each having one subset of $N - 1$ distinct variables. From the produced models only the one that further reduces, by largest margin, the AIC of the model with N variables is considered for the next iteration. The AIC, defined as $-2logL(\theta) + 2k$, where L(θ) is the maximized likelihood [9] of the model and k the number of parameters, strikes a balance between likelihood and parsimony. Use of an information-theoretic criterion mitigates the problems associated with approaches based on significance testing [10]. Finally, we use background medical knowledge to eliminate model coefficients not compliant to clinical expectations. In particular we: (1) eliminate any episode with "failure" at the day of prediction and a negative β coefficient in the model (e.g. $\beta = -0.7$ for $\{NF, F, F\}$) and (2) eliminate any episode with "non-failure" at the day of prediction and a positive β coefficient in the model (e.g. $\beta = 1.1$ for $\{NF, F, NF\}$). Keep in mind that a negative coefficient reduces the probability of mortality, and a positive one increases it. A similar idea was introduced in [11], under the name "sign OK", defining a variable selection strategy based on the plausibility of the sign of the regression coefficient.

Another thorny issue requiring attention is the phenomenon of *collinearity*, a situation in which at least one of the covariates can be predicted well from the other covariates [10]. This leads to instability of the model and jeopardizes the interpretability of the β coefficients in the logistic model (see appendix) since it is based on the idea of studying a change of a covariate while fixing the others. However, holding down the values fixed of the collinear covariates is unattainable because, by definition, they will be affected. One strong type of collinearity which is ubiquitous in our domain when dealing with aligned episodes, is the occurrence of the *logically entailed* episodes [5]. For example we say that episode $\{NF, F, F\}$ logically entails episode $\{F, F\}$ since the occurrence of the first in a patient implies the occurrence of the second episode. To eliminate logical entailment we included a ranking step in the modeling stage (this procedure is more stringent but simpler than the one suggested in [5]). For each one of the six organ systems, all its discovered frequent OF episodes are ranked, from those with smallest (best) AIC value to the largest, based on a univariate analysis between mortality and the episode. For each organ system we retain only its highest ranked episode.

This eliminates logically entailed episodes and provides simple models. This risks eliminating other possibly useful episodes, but with only 2 categories (F and NF) any two aligned episodes are at least partially correlated. The episodes obtained in this manner are then fed into the AIC-based feature elimination strategy described above.

Evaluation. For each day of prediction d a separate training and testing set are created. An important performance aspect of a probabilistic model is its calibration ability. We applied the commonly used Brier score, $\frac{1}{N}\sum_{i=1}^{N}(P(Y_i = 1 \mid \mathbf{x}_i) - y_i)^2$, where N denotes the number of patients, and y_i denotes the actual outcome for patient i. The vector $\mathbf{x}_i$ represents the covariate values for patient i. The Brier score is a *strictly proper scoring rule* [12] which, unlike measures like the area under the ROC curve, means it is optimal only when the true probability of the event is provided. The performance of each of the OF-models is compared to its corresponding SOFA and static models. To test for statistical significance in performance difference we advocate the use of the non-parametric bootstrap method [13] with 1000 bootstrap samples of differences.

3 Case Study

The ICU patient dataset is available from the OLVG, a teaching hospital in Amsterdam and was collected during July 1998 and October 2006 including all 2785 patients (25% mortality) eligible for analysis [5]. Both SAPS and SOFA scores values were larger in the non-survivors (averages are: SAPS 61±15.3 vs. 39±18.4 for survivors, SOFA: 9.7±3.2 vs. 7.3±2.6). The mean number of failures (IOSF scores values $\in \{3, 4\}$) per patient, give a clear indication of the high association between organ failure and survival outcome (9.8 organ failures in non-survivors versus 4.4 organ failures in survivors).

4 Results

Based on the method described above, four OF-models corresponding to the ICU days 2–5 (day 1 cannot show temporal evolution), were created for predicting the hospital mortality. In episode discovery, a threshold of 5% was used for minimum support rate. Each OF-model includes the SAPS covariates (SAPS, log(SAPS+1)) and, potentially, after variable selection the average SOFA and frequent OF episodes. For comparative purposes the same training set was used to induce the static and SOFA-models for the given day. Table 2 shows the resulting models described by their's linear predictor (LP). logSAPS represent $log(SAPS + 1)$ (used in compliance with the original SAPS model). The organ failure episodes are labeled to identify their type of organ system. For example Resp$\{F, NF\}$ represents a failure followed by a non-failure in the respiratory system. The SOFA-models use the elements $\{L, M, H\}$ to describe frequent SOFA episodes e.g. $\{HM\}$. Table 3 exemplifies the interpretation in terms of odds-ratios (equal to $exp(\beta)$) of the OF-model coefficients for day 2 and 5. For

Table 2. Temporal models (OF and SOFA) and static models for days 2–5 of ICU stay described by their linear predictors (LPs)

Day	OF–model LP	SOFA–model LP	Static model LP
2	-9.3 +0.005SAPS + 1.9logSAPS+0.065meanSOFA -1.85Resp{F,NF} +1.1CNS{F,F}	-5.9 +0.03SAPS + 0.87logSAPS +0.6H -0.7L	-7.7 +0.036SAPS +1.26logSAPS

3	-10.8 -0.01SAPS +2.2logSAPS
	+0.13meanSOFA +0.4Resp{F,F}
	+1.1CNS{F}
		-7.26 +0.01SAPS
		+1.4logSAPS+ 1.1H -0.66L
			-10.35 +0.02SAPS
			+2.2logSAPS

4	-6.7 -0.006SAPS +1.9logSAPS +0.45Resp{F,F,F,F} +1.56CNS{F} -0.62Hepa{NF,NF,NF,NF} -0.8Cardio{NF} -0.8Ren{NF,NF,NF}	-5.5 +0.014SAPS +1.18logSAPS -1.95L -0.83MM -0.65HM	-7.88 +0.027SAPS +1.42logSAPS

5	-6 -0.006SAPS +1.5logSAPS
	+0.5Resp{F,F,F,F,F} -0.9Coag{NF,NF}
	+1.4CNS{F} -0.5Ren{NF,NF,NF,NF}
		-2.5 +0.02SAPS
		+0.12logSAPS -1.04L
		+0.65H
			-5.5 +0.02SAPS
			+0.85logSAPS

Table 3. Model covariates, their coefficients and odds-ratios ($exp(\beta_i)$) in the OF-models for day 2 and 5 of ICU stay

Day	Covariate	β	e^β	Day	Covariate	β	e^β
2	SAPS	0.005	1.005	5	SAPS	-0.006	0.99
	logSAPS	1.9	6.68		logSAPS	1.5	4.48
	meanSOFA	0.065	1.07		Resp{F,F,F,F,F}	0.5	1.64
	Resp{F,NF}	-1.85	0.16		Coag{NF,NF}	-0.9	0.4
	CNS{F,F}	1.1	3		CNS{F}	1.4	4.05
					Ren{NF,NF,NF,NF}	-0.5	0.6

Table 4. Performance evaluation – Brier score

Day	Brier score			OF–model win		SOFA–model win
	OF	SOFA	SAPS	vs. SAPS	vs. SOFA	vs. SAPS
2	0.157	0.158	0.163	Yes	Yes	Yes
3	0.179	0.185	0.197	Yes*	Yes	Yes*
4	0.199	0.209	0.212	Yes	Yes	Yes
5	0.186	0.189	0.207	Yes*	Yes	Yes*

example, for day 2, after adjusting for the other variables, the odds of dying for patients with the episode CNS$\{F, F\}$ is three times the odds for those without it. The OF-models shown in Table 2 where evaluated on an independent test set for day 2 till day 5 of ICU stay and compared to the static and SOFA-models. The Brier scores (the lower the better) are shown in Table 4. An * indicates a statistically significantly better Brier score than the static model.

5 Discussion and Related Work

In this section we discuss the results, our approach in relation to others, delineate further work, and conclude the paper.

Results. Table 4 ascertains that the OF episodes can improve predictions: the OF-models performed better than the SOFA-models on all days. Also, both kinds of temporal models (SOFA and OF) were consistently better (sometimes with statistically significant differences) than the traditional static model (SAPS model). This evidence needs of course corroboration by a more stringent cross-validation design that we plan to do in the future. The results also show the usefulness of the coefficient qualitative and quantitative interpretations of organ derangement (see Table 2). In all days the central nervous and respiratory systems were present in the models. The renal organ system was the next best predictor included in two models. In related work, when the central nervous system was considered for analysis [14,15,16] it was indeed a good discriminator of mortality, otherwise, as in [17,18,19], the cardiovascular system emerged as a strong predictor. When considering the frequent episodes selected we note that those denoting constant organ conditions (failure e.g. $\{F, F\}$ or non-failure $\{NF, NF, NF, NF\}$) were dominant. Similar findings about the "constant patterns" have been reported by [20]. We hypothesize that their dominance is rooted in the high support they enjoy in the data: individual scores are not likely to change often between the only two categories discerned (Non-failure, Failure).

Table 3 can be used to provide quantitative interpretations. For example, in the model for day 5 the central nervous system episode $\{F\}$ is associated with the odds-ratio of about 4: the odds of dying given a failure of the central nervous system failing on day 5 (day of prediction) is about 4 times the odds of dying if the central nervous system did not fail on that day. By the same token, in case of non-failure at the day of prediction an odds-ratio < 1 (corresponding to a negative coefficient in the LP) indicates a beneficial effect on the prognosis. This holds for example for the renal episode $\{NF, NF, NF, NF\}$.

Another finding is that, starting from day 4 of prediction, the models do not include the mean SOFA score anymore. A possible reason is that the importance of the latest days is diluted in the *unweighted* SOFA score mean by the early days which might be less important. Future work consists of using a weighted summary of the SOFA scores where later days enjoy more weight.

Approach, related and future work. The main differences between the OF-model induction approach and the one described in [5] to induce SOFA-models, aside from applying it to other data types, are the following. First, for the induction of OF-models, SOFA scores are aggregated in one summary statistic in order to avoid major overlap with the IOSF scores. Second, each of the 6 organ systems is categorized using clinical knowledge corresponding to organ failure. Third, to avoid strong collinearity between the organ failure episodes we follow a new stringent feature selection strategy based on ranking. Fourth, clinical knowledge is infused in the covariate selection process by using the "sign OK" rule.

The *categorization* we adopted of the IOSF scores in 2 categories can be characterized as vertical (contemporaneous) data abstraction [21] or State Temporal Abstractions [22]. The resulting categories are not necessarily the most useful ones in prediction. Future work will include the use of the outcome (mortality), e.g. by using entropy-based categorization [23], or additional medical knowledge [18] to generate more predictive categories.

Our frequent episodes are discovered in a *data-driven* manner. In using these episodes we also apply various *non-stationarity* assumptions: the episodes are *aligned* to the prediction day, and can be of *different* lengths. These two properties distinguish our work from the work appearing in the intensive care literature in which pre-specified summaries (e.g. maximum value, average value during ICU stay) or qualitative trends of IOSF scores are used [16,18]. This is also in contrast to the methodological work described in [24] which assumes a strict form of stationarity where the mean value of adverse events (highly abnormal values of medical measurements in a patient) is used to summarize the temporal process.

Valuable future work consists of investigating more expressive episodes like those described in [25] where a pattern includes a multivariate set of attributes. This will allow one to capture complex interactions between the IOSF scores in one pattern instead of having various univariate episodes as describes in this paper.

In [20] patterns similar in nature to ours are discovered not based on frequency but on their discriminative capabilities (Area Under the Curve) and forms an interesting future work. The most predictive ones are then used in a Naive Bayes Classifier method. Given the nature of our episodes, the Naive Bayes approach combined with an assessment geared towards discriminatory model capabilities does not provide incentive to predict reliable probabilities. This is because our particular episodes are correlated and because of the overlap between SOFA and IOSF information, clearly violating the conditional independence assumption used in the Naive Bayes approach.

Making use of logistic regression allows a fair comparison between our method and that currently used in the ICU. It also generates coefficients with meaningful interpretation as demonstrated above. These coefficients' interpretability is enhanced by the application of a logical entailment elimination step by means of the AIC criterion avoiding drawbacks related to the p-value based variable selection approaches. In [24] a comparison of various methods showed that in a given setting, different than ours, the logistic regression model was slightly outperformed only by the neural network model in terms of accuracy. Whether neural networks will lead to better calibrated results in our case is unclear, but if one is also interested in the interpretability of the models the logistic regression is a good choice. Future work could investigate how hybrid methods can be employed. For example in [26] a classification tree based on baseline information was used to stratify patients into homogeneous subgroups, then on each of these subgroups a logistic regression model was fit to predict mortality. A similar idea could be applied for temporal data.

In sum, this paper proposed a method for inducing predictive models based on the integration of the information residing in baseline data with the temporal information concerning organ system functioning into the logistic regression framework. The results are encouraging as the temporal organ failure episodes improve predictions' quality and interpretability. The ubiquity of scoring systems in medicine for baseline and repeated measurements suggests the wider applicability of our approach to other domains.

Acknowledgments. This work was supported by the Netherlands Organization for Scientific Research (NWO) under the I-Catcher project number 634.000.020. Thanks for useful feedback are also due to Arno Siebes from the University of Utrecht and to the other members of the Methodology in Medical Informatics group in our department.

References

1. Abu-Hanna, A., Lucas, P.J.F.: Editorial: Prognostic models in medicine - AI and statistical approaches. Methods of Information in Medicine 40, 1–5 (2001)
2. Knaus, W., Draper, E., Wagner, D., Zimmerman, J.: APACHE II: a Severity of Disease Classification System. Crit. Care Med. 13, 818–829 (1985)
3. Le Gall, J., Lemeshow, S., Saulnier, F.: A new Simplified Acute Physiology Score (SAPS-II) based on a European/North American multicenter study. The Journal of the American Medical Association 270, 2957–2963 (1993)
4. Vincent, J.-L., Ferreira, F.L.: Evaluation of organ failure: we are making progress. Intensive Care Med. 26, 1023–1024 (2000)
5. Toma, T., Abu-Hanna, A., Bosman, R.J.: Discovery and Inclusion of SOFA Score Episodes in Mortality Prediction. Journal of Biomedical Informatics (to appear)
6. Agrawal, R., Srikant, S.: Fast algorithms for mining association rules. In: Proc. of the 20th Very Large Data Bases Conf., pp. 487–499 (1994)
7. Mannila, H., Toivonen, H., Verkamo, A.I.: Discovering frequent episodes in sequences. Data Mining Knowledge Discovery 1(3), 259–289 (1997)
8. Burnham, K.P., Anderson, D.R.: Model Selection and Multimodel Inference: A Practical-Theoretic Approach, 2nd edn. Springer, New York (2002)
9. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer Series in Statistics (2001)
10. Harrell Jr., F.E.: Regression Modeling Strategies. 1st edn. Springer Series in Statistics (2001)
11. Steyerberg, E.W., Eijkemans, M.J., Harrell Jr., F.E., Habbema, J.D.: Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat. Med. 19(8), 1059–1079 (2000)
12. Hand, J.D.: Construction and Assessment of Classification Rules. John Wiley & Sons, Chichester (1997)
13. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)
14. Peres Bota, D., Melot, C., Lopes Ferreira, F., Nguyen Ba, V., Vincent, J.L.: The Multiple Organ Dysfunction Score (MODS) versus the Sequential Organ Failure Assessment (SOFA) score in outcome prediction. Intensive Care Med. 28(11), 1619–1624 (2002)
15. Pettila, V., Pettila, M., Sarna, S., Voutilainen, P., Takkunen, O.: Comparison of multiple organ dysfunction scores in the prediction of hospital mortality in the critically ill. Crit. Care Med. 30(8), 1705–1711 (2002)
16. Russell, J.A., et al.: Changing pattern of organ dysfunction in early human sepsis is related to mortality. Crit. Care Med. 28, 3405–3411 (2000)

17. Zygun, D.A., Kortbeek, J.B., Fick, G.H., Laupland, K.B., Doig, C.J.: Non-neurologic organ dysfunction in severe traumatic brain injury. Crit. Care Med. 33(3), 654–660 (2005)
18. Levy, M.M., Macias, W.L., Vincent, J.L., Russell, J.A., Silva, E., Trzaskoma, B., Williams, M.D.: Early changes in organ function predict eventual survival in severe sepsis. Crit. Care Med. 33(10), 2194–2201 (2005)
19. Nfor, T.K., Walsh, T.S., Prescott, R.J.: The impact of organ failures and their relationship with outcome in intensive care: analysis of a prospective multicenter database of adult admissions. Anaesthesia 61, 731–738 (2006)
20. Kayaalp, M., Cooper, G., Clermont, G.: Predicting with variables constructed from temporal sequences. In: Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics. pp. 220–225 (2001)
21. Shahar, Y.: A framework for knowledge–based temporal abstraction. Artificial Intelligence 90, 79–133 (1997)
22. Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R.: Temporal data mining for the quality assessment of hemodialysis services. Artificial Intelligence in Medicine 34, 25–39 (2005)
23. Kohavi, R., Sahami, M.: Error-Based and entropy-based discretization of continuous features. In: Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, pp. 114–119 (1996)
24. Silva, A., Cortez, P., Santos, M.F., Gomes, L., Neves, J.: Mortality assessment in intensive care units via adverse events using artificial neural networks. Artificial Intelligence in Medicine 36, 223–234 (2006)
25. Srikant, R., Agrawal, R.: Mining Sequential Patterns: Generalizations And Performance Improvements. In: Proc. 5th Int. Conf. Extending Database Technology, vol. 1057, pp. 3–17 (1996)
26. Abu-Hanna, A., Keizer, N.F.: Integrating Classification Trees with Local Logistic Regression in Intensive Care Prognosis. Artificial Intelligence in Medicine 29(1–2), 5–23 (2003)
27. Hosmer, D.W., Lemeshow, S.: Applied logistic regression. John Wiley & Sons, Inc., New York (1989)

Appendix: Logistic Regression

A logistic regression model [27] specifies the conditional probability of a binary outcome variable Y, given the values of the covariate vector $\mathbf{x} = (x_1, ..., x_m)$: $p(Y = 1 \mid \mathbf{x}) = \frac{e^{LP(\mathbf{x})}}{1+e^{LP(\mathbf{x})}}$. For m covariates the natural logarithm of the odds (*logit*) is equal to the *linear predictor* $LP(\mathbf{x})$: $log(\frac{p(Y=1 \mid \mathbf{x})}{1-p(Y=1 \mid \mathbf{x})}) = LP(\mathbf{x}) = \beta_0 + \sum_{i=1}^{m} \beta_i \cdot x_i$ where β_i, $i = 1, ..., m$, denote the coefficients of the m covariates. A coefficient (β_i) can be interpreted in terms of an *odds-ratio*. Suppose the linear predictor is $\beta_0 + \beta_1 \cdot SAPS + \beta_2 \cdot Ep$ where $Ep = 1$ for patients having some specific episode and 0 for patients not having the episode. The odds of dying for those having the episode, $odds(Ep = 1)$ is $P(Y = 1|Ep = 1)/P(Y = 0|Ep = 1)$ and for those not having the episode, $odds(Ep = 0)$, is $P(Y = 1|Ep = 0)/P(Y = 0|Ep = 0)$. The quantity e^{β_2} is equal to the odds-ratio $odds(Ep = 1)/odds(Ep = 0)$. A positive coefficient corresponds to an odds-ratio > 1 and indicates that, when adjusting for all other variables (here SAPS), the odds of the event is higher for those with the episode compared to those without it.

Part III

Machine Learning and Knowledge Discovery

Contrast Set Mining for Distinguishing Between Similar Diseases

Petra Kralj[1], Nada Lavrač[1,2], Dragan Gamberger[3], and Antonija Krstačić[4,⋆]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia
[3] Rudjer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia
[4] University Hospital of Traumatology, Draškovićeva 19, 10000 Zagreb, Croatia

Abstract. The task addressed and the method proposed in this paper aim at improved understanding of differences between similar diseases. In particular we address the problem of distinguishing between thrombolic brain stroke and embolic brain stroke as an application of our approach of contrast set mining through subgroup discovery. We describe methodological lessons learned in the analysis of brain ischaemia data and a practical implementation of the approach within an open source data mining toolbox.

1 Introduction

Data analysis in medical applications is characterized by the ambitious goal of extracting potentially new relationships from data, and providing insightful representations of detected relationships. Methods for symbolic data analysis are preferred since highly accurate but non-interpretable classifiers are frequently considered useless for medical practice.

A special data mining task dedicated to finding differences between contrasting groups is contrast set mining [1]. The goal of our research is to find discriminative differences between two groups of ischaematic brain stroke patients: patients with thrombolic stroke and those with embolic stroke. The problem is introduced in Section 2.

Contrast set mining can be performed by a specialized algorithm STUCCO [1], through decision tree induction and rule learning [13], and—as shown in our recent work—through subgroup discovery [7]. Section 3 presents the results of decision tree induction on our contrast set mining task and discuss advantages and disadvantages of this approach. In Section 4 we show an approach to contrast set mining through subgroup discovery by providing a mathematically correct translation from contrast set mining to subgroup discovery [7] and an implementation of the approach in the Orange [2] open source data mining toolbox. Next

[⋆] This work was supported by Slovenian Ministry of Higher Education, Science and Technology project “Knowledge Technologies”, Croatian Ministry of Science, Education and Sport project “Machine Learning Algorithms and Applications”, and EU FP6 project “Heartfaid: A knowledge based platform of services for supporting medical-clinical management of the heart failure within the elderly population”.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 109–118, 2007.

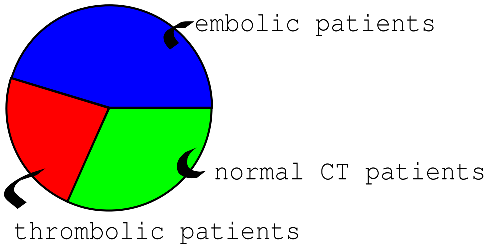


Fig. 1. Distribution of diagnosis of patients in our dataset

we show that the direct "round robin" contrast set mining approach to solve our descriptive induction task leads to rather disappointing results. We discuss the reasons for this undesired performance. This lesson learned resulted in a different, more appropriate "one-versus-all" transformation of contrast set mining to subgroup discovery, justified by the improved results of our experiments, confirmed by the medical expert.

2 The Brain Ischaemia Data Analysis Problem

A stroke occurs when blood supply to a part of the brain is interrupted, resulting in tissue death and loss of brain function. Thrombi or emboli due to atherosclerosis commonly cause ischemic arterial obstruction. Atheromas, which underlie most thrombi, may affect any major cerebral artery. Atherothrombotic infarction occurs with atherosclerotic involving selected sites in the extracranial and major intracranial arteries. Cerebral emboly may lodge temporarily or permanently anywhere in the cerebral arterial tree. They usually come from atheromas (ulcerated atheroscleritic plaques) in extracranial vessels or from thrombi in a damaged heart (from mural thrombi in atrial fibrillation). Atherosclerotic or hypertensive stenosis can also cause a stroke. Embolic strokes, thrombolic strokes and stokes caused by stenosis of blood vessels are categorized as ischaemic strokes. 80% of all strokes are ischaemic while the remaining 20% are caused by bleeding.

We analyze the brain ischaemia database, which consists of records of patients who were treated at the Intensive Care Unit of the Department of Neurology, University Hospital Center "Zagreb", Zagreb, Croatia, in year 2003. In total, 300 patients are included in the database (Figure 1):

- 209 patients with the computed tomography (CT) confirmed diagnosis of brain stroke: 125 with embolic stroke, 80 with thrombolic stroke, and 4 undefined.
- 91 patients who entered the same hospital department with adequate neurological symptoms and disorders, but were diagnosed (based on the outcomes of neurological tests and CT) as patients with transition ischaemic brain attack (TIA, 33 patients), reversible ischaemic neurological deficit (RIND, 12 patients), and severe headache or cervical spine syndrome (46 patients).

Patients are described with 26 descriptors representing anamnestic, physical examination, laboratory test and ECG data, and their diagnosis. Anamnestic data: aspirin therapy *(asp: yes, no)*, anticoagulant therapy *(acoag: yes, no)*, antihypertensive therapy *(ahyp: yes, no)*, antiarrhytmic therapy *(aarrh: yes, no)*, antihyperlipoproteinaemic therapy – statin *(stat: yes, no)*, hypoglycemic therapy *(hypo: none, yesO – oral, yesI – insulin)*, sex *(m or f)*, age *(in years)*, present smoking *(smok: yes, no)*, stress *(str: yes, no)*, alcohol consumption *(alcoh: yes, no)* and family anamnesis *(fhis: yes, no)*. Physical examination data are: body mass index *(bmi: ref. value 18.5–25)*, systolic blood pressure *(sys: normal value* < 139 *mmHg)*, diastolic blood pressure *(dya: normal value* < 89 *mmHg)* and fundus ocular *(fo: discrete value 0-4)*. Laboratory test data: uric acid *(ua: ref. value for men* < 412 $\mu mol\ L^{-1}$*, for women* < 380 $\mu mol\ L^{-1}$*)*, fibrinogen *(fibr: ref. value 2.0–3.7 g* L^{-1} *)*, glucose *(gluc: ref. value 3.6–5.8 mmol* L^{-1}*)*, total cholesterol *(chol: ref. value 3.6–5.0 mmol* L^{-1}*)*, triglyceride *(trig: ref. value 0.9–1.7 mmol* $L^{-}1$*)*, platelets *(plat: ref. value 150000–400000)* and prothrombin time *(pt: ref. value without th. 0.7–1.2, with anticoagulant th. 0.25–0.4)*. ECG data: heart rate *(ecgfr: ref. value 60–100 beats/min)*, atrial fibrillation *(af: yes, no)* and left ventricular hypertrophy *(ecghlv: yes, no)*.

In this paper, the goal of data analysis is to discover regularities that discriminate between thrombolic and embolic stroke patients. Despite the fact that the immediate treatment for all kinds of ischeamic strokes is the same, the distinction between thrombolic and embolic stroke patients is important in later phases of patient recovery and to better determine the risk factors of the specific diseases.

It must be noted that this dataset does not consist of healthy individuals but of patients with serious neurological symptoms and disorders. In this sense, the available database is particularly appropriate for studying the specific characteristics and subtle differences that distinguish patients with different neurological disorders. The detected relationships can be accepted as generally true characteristics for these patients. However, the computed evaluation measures only reflect characteristics specific to the available data, not necessarily holding for the general population or other medical institutions [12].

3 Searching for Contrast Sets by Decision Tree Induction

Decision trees [11] are a classical machine learning technique. By selecting the attribute that best distinguishes between the classes and putting it as a root node, they partition the examples into subsets of examples where the same method is recursively applied. We have induced a decision tree in Figure 2 for a problem of distinguishing between two types of patients with brain stroke (marked “emb” and “thr”) and patients with normal (marked “norm”) brain CT test results.[1]

[1] In the experiments we used rigorous pruning parameters to induce small and comprehensible decision trees, using a decision tree learner implemented in the Orange data mining toolbox [2]. Due to noisy data and harsh pruning the decision tree has low classification accuracy (58% accuracy estimated by 10 fold crossvalidation).

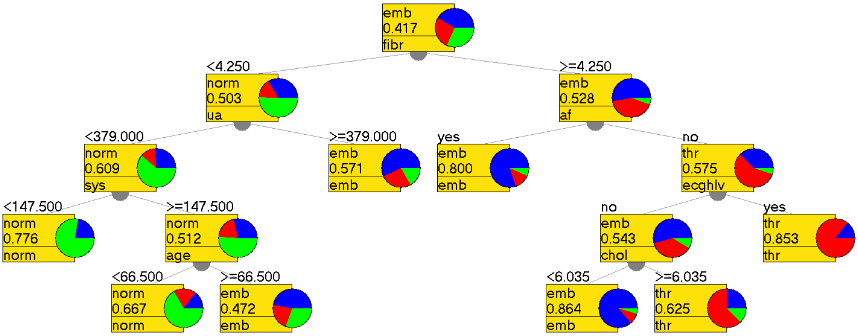


Fig. 2. A decision tree distinguishing between patients with embolic brain stroke, thrombolic brain stroke and patients with normal brain CT test results

The interpretation of the decision tree by the medical expert is that fibrinogen ("fibr") is the "most informative" attribute distinguishing between patients with and without brain ischaemia, and that atrial fibrillation ("af") is the attribute that best distinguishes between groups of embolic and thrombolic patients. While the induced decision tree well represents the medical knowledge applied in patient diagnosis, the intention of this experiment was not to produce a classifier, but to generate a descriptive model and to investigate the advantages and disadvantages of decision tree induction for contrast set mining.

In the contrast set mining setting, the main advantage of decision trees is the simplicity of their interpretation. On the other hand, there are many disadvantages. A decision tree partitions the space of covered examples, disallowing the overlapping of the discovered patterns. All the contrasting patterns (rules formed of decision tree paths) include the same root attribute (*fibrinogen*), which is disadvantageous compared to contrast set rule representations. Moreover, due to attribute repetitions and thus a limited set of attributes appearing in decision tree paths, the variety of contrasting patterns is too limited.

4 Contrast Set Mining Through Subgroup Discovery

A data mining task devoted to finding differences between groups is *contrast set mining* (CSM). It was defined by Bay and Pazzani [1] as finding "conjunctions of attributes and values that differ meaningfully across groups". If was later shown that contrast set mining is a special case of a more general rule discovery task [13]. Finding all the patterns that discriminate one group of individuals from all other contrasting groups is not appropriate for human interpretation. Therefore, as is the case in other descriptive induction tasks, the goal of contrast set mining is to find only the descriptions that are "unexpected" and "most interesting" to the end-user [1].

On the other hand, a *subgroup discovery* (SD) task is defined as follows: Given a population of individuals and a property of those individuals that we are

Table 1. Table of synonyms from different communities

Contrast Set Mining (CSM)	Subgroup Discovery (SD)	Rule Learning (RL)
contrast set	subgroup description	rule condition
group	class (property of interest)	class
attribute value pair	feature	condition
examples in groups $G_1, \ldots G_n$	examples of $Class$ and $\overline{Class}$	examples of $C_1 \ldots C_n$
examples for which contrast set is true	subgroup	covered examples

interested in, find population subgroups that are statistically "most interesting", i.e., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the given property of interest [14].

Putting these two tasks in a broader rule learning context, note that there are two main ways of inducing rules in multiclass learning problems: learners either induce the rules that characterize one class compared to the rest of the data (the standard *one versus all* setting, used in most classification rule learners), or alternatively, they search for rules that discriminate between all pairs of classes (known as the *round robin* approach used in classification rule learning, proposed by [3]). Subgroup discovery is typically performed in a one vs. all rule induction setting, while contrast set mining implements a round robin approach (of course, with different heuristics and goals compared to classification rule learning).

Section 4.1 shows that, using a round robin setting, a CSM task can be directly translated into a SD task. The experiments in brain ischaemia data analysis were performed using a novel implementation of our subgroup discovery algorithms in the Orange data mining toolbox, characterized by excellent data and model visualization facilities (see Section 4.2).

The direct transformation of a CSM task into a SD task in the round robin setting showed some problems when used for contrast set mining for distinguishing between thrombotic and embolic patient groups (see Section 5). This lead to a modified task transformation, following the more "natural" one-versus-all subgroup discovery setting (see Section 6).

4.1 Round Robin Transformation: Unifying CSM and SD

Even though the definitions of subgroup discovery and contrast set mining seem different, the tasks are compatible [7]. From a dataset of class labeled instances (the class label being the property of interest) by means of subgroup discovery [4] we can find contrast sets in a form of short interpretable rules. Note, however, that in subgroup discovery we have only one property of interest (class) for which we are building subgroup descriptions, while in contrast set mining each contrasting group can be seen as a property of interest.

Moreover, using the dictionary of Table 1, it is now easy to show that a two-group contrast set mining task $CSM(G_1, G_2)$ can be directly translated into the following two subgroup discovery tasks: $SD(Class = G_1$ vs. $\overline{Class} = G_2)$ and $SD(Class = G_2$ vs. $\overline{Class} = G_1)$. And since this translation is possible for a two-group contrast set mining task, it is—by induction—also possible for a general contrast set mining task. This induction step is as follows:

```
$CSM(G_1, \ldots, G_n)$
    for i=1 to n do
        for j=1, j$\neq$ i to n do
            $SD(Class = G_i\ vs.\ \overline{Class} = G_j)$
```

4.2 Implementations of Subgroup Discovery Algorithms and Subgroup Visualization in Orange

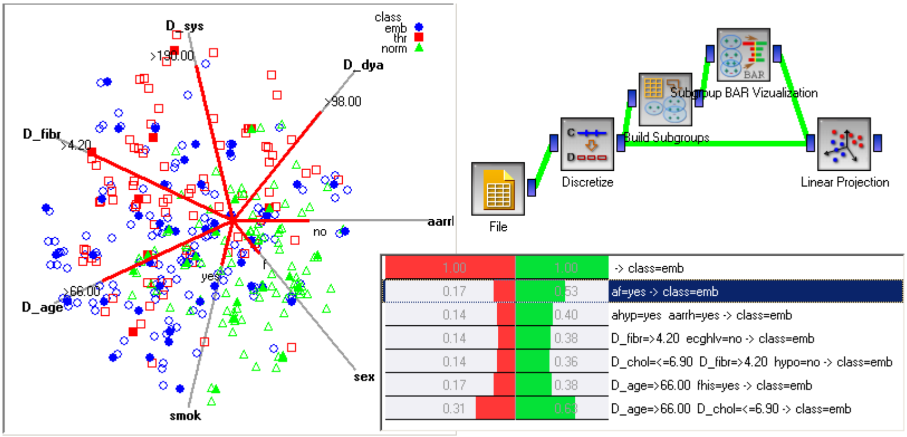
There are several algorithms that are adaptations of rule learners to perform the subgroup discovery task: SD [4], CN2-SD [10] and Apriori-SD [6]. We have reimplemented these algorithms [9] in Orange [2] with some minor adaptations compared to the descriptions in the original papers. The difference arises from the internal representation of the data in Orange, based on attributes and not on features (attribute values). Data need to be discretized in the preprocessing phase, as the implementations construct attribute-value pairs from discretized data on the fly while constructing rules. Despite this data representation limitation, the algorithm reimplementation in Orange is worthy, as it offers various data and model visualization tools and has excellent facilities for building new visualizations.

We here briefly describe just the APRIORI-SD algorithm [6], an adaptation of the algorithm for mining classification rules with association rule learning techniques APRIORI-C [5], which was used in our experiments. The main modifications of APRIORI-C, making it appropriate for subgroup discovery, involve the implementation of an example weighting scheme in rule post-processing, a modified rule quality function incorporating example weights and a probabilistic classification scheme.

Orange goes beyond static visualization, by allowing the interaction of the user and combination of different visualization techniques. In Figure 3 an example of a visual program in the Orange visual programming tool Orange Canvas is shown.[2] The first widget from the left (*File*) loads the dataset (in this example we load the Brain Ischemia dataset with three classes). The following widget (*Discretize*) takes care of data discretization in the preprocessing phase. It is followed by the new widget *Build Subgroups* which is in charge of building subgroups. In this widget the user chooses the algorithm for subgroup discovery and sets the algorithm parameters.

We have implemented a new subgroup visualization technique called the *visualization by bar charts* [8], described in the next paragraph. The widget *Subgroup Bar Visualization* provides the visualization of the subgroups. It can be

[2] This visual program is just one example of what can be done by using the Subgroup discovery tool implemented in Orange. Subgroup evaluation and different method for visualizing the contents of subgroups are also available.



1.00	1.00	-> class=emb
0.17	0.53	af=yes -> class=emb
0.14	0.40	ahyp=yes aarrh=yes -> class=emb
0.14	0.38	D_fibr=>4.20 ecghlv=no -> class=emb
0.14	0.36	D_chol=<=6.90 D_fibr=>4.20 hypo=no -> class=emb
0.17	0.38	D_age=>66.00 fhis=yes -> class=emb
0.31	0.63	D_age=>66.00 D_chol=<=6.90 -> class=emb

Fig. 3. An example of a visual program in the interactive interface for subgroup discovery implemented in Orange

connected to several other widgets for data visualization. In our case we connected it to existing *Linear Projection* visualization (see the left-hand side of Figure 3) which visualizes the entries of the entire dataset as empty shapes and the entries belonging to the group selected in the *Subgroup Bar Visualization* widget as full shapes. By moving the mouse over a certain shape in the *Linear Projection* widget the detailed description of the entry is displayed.

In the bar chart visualization (shown below the Orange Canvas in Figure 3) the first line's purpose is to visualize the distribution of the entire example set. The area on the right represents the positive examples and the area on the left represents the negative examples. Each following line represents one subgroup. The positive and the negative examples of each subgroup are drawn below the positive and the negative examples of the entire example set. Subgroups are sorted by the relative share of positive examples. Examples of this visualization are shown in Figures 4 and 5.

This visualization method allows simple comparison between subgroups and is therefore useful. It is very intuitive and attractive to end-users. All the displayed data is correct and not misleading. It is very simple and does not display the contents of data, but it can be connected to other data visualizations in Orange (Figure 3) in order to allow in depth investigations.

5 Experimental Evaluation of the Round Robin CSM

The goal of our experiments was to find characteristic differences between patients with thrombolic and embolic ischeamic stroke. We approached this task by applying the round robin transformation from contrast set mining to subgroup discovery, described in Section 4.1. We ran this experiment and asked the experts for interpretation.

The resulting rules mainly include the feature $AF = no$ for thrombolic patients and $AF = yes$ for embolic patients, which are very typical for the corresponding diseases. However, the rules turned out to be non-intuitive to the medical expert. For example, the rule

$$af = yes \ \& \ sys < 185 \ \& \ fo = 1 \rightarrow embolic$$

covering many embolic and just one thrombolic patient (TP =33, FP = 1) was interpreted as "people with suspected thromb in the heart ($af = yes$) and visible consequences of hypertension in the eyes ($FO = 1$)". The feature $sys < 185$ says: patients with not extremely high systolic blood pressure, though high blood pressure is characteristic for both the diseases and the boundary 185 is very high, since everything above 139 is considered high in medical practice.[3]

We investigated further the reasons why the rules were difficult to interpret for domain experts. The reason comes from the task itself: Medical physicians are not used to distinguish between two types of disease given the condition that a patient has a disease, but are rather used to find characteristics for a specific type of a disease compared to the entire population. Another motivation is to avoid rules as

$$fhis = yes \ \& \ smok = yes \ \& \ asp = no \ \& \ dya < 112.5 \rightarrow embolic$$

This rule has good covering properties (TP=28, FP=4), but practically describes healthy people with family history of brain stroke. It is undoubtedly true that this pattern is present in the dataset, but it is not the reason why these patients have a certain type of disease. The algorithm just could not know that the combination of these features is not characteristic for group differentiation simply because it did not have normal people as a reference.

6 Experimental Evaluation of the One-Versus-All CSM

As the medical expert was not satisfied with the results of the comparison of thrombolic and embolic patients, we investigated the reasons and learned a lesson in medical contrast set mining. To overcome the problems related to the original definition of contrast set mining we need to modify the task: instead of using the round robin approach where we compare classes pairwise, we use a one vs. all approach which is standard in subgroup discovery. In this way we give the algorithm also the information about healthy patients.

Our dataset is composed of three groups of patients, as described in Section 2 and shown on Figure 1. An approach we claim is applicable in many similar domains where the differences between two varieties of one disease are as follows: To find characteristics of the embolic patients we perform subgroup discovery on the embolic group compared to the rest of the patients (thrombolic and those with a normal CT). Similarly, when searching for characteristics of thrombolic patients, we compare them to the rest of the patients (embolic and those with a normal CT).

[3] In our dataset there are 56 patients with $sys > 185$.

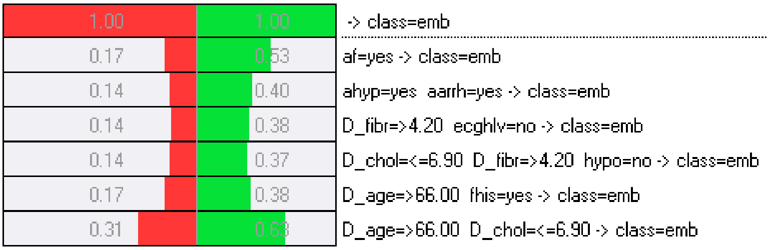


Fig. 4. Characteristic descriptions of embolic patients displayed in the bar chart subgroup visualization: on the right side the positive cases, in our case embolic patients, and on the left hand side the others - thombolic and normal CT

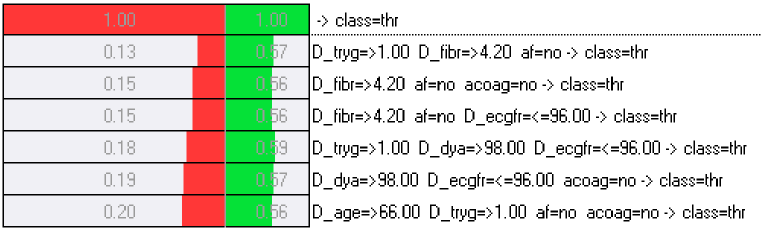


Fig. 5. Characteristic descriptions of thrombolic patients

In this setting, we ran the experiment with Orange implementation of Apriori-SD. We used the following parameter values: minimal support = 15%, minimal confidence = 30%, the parameter for tuning the covering properties k = 5. The results are displayed in Figures 4 and 5.

Strokes caused by embolism are most commonly caused by heart disorders. The first rule displayed on Figure 4 has only one condition confirming atrial fibrillation ($af = yes$) as an indicator for embolic brain stroke. The combination of features from the second rule also shows that patients with antihypertensive therapy ($ahyp = yes$) and antiarrhytmic therapy ($aarrh = yes$), therefore patients with heart disorders, are prone to embolic stroke.

Thrombolic stroke is most common with older people, and often there is underlying atherosclerosis or diabetes. In the rules displayed in Figure 5 the features presenting diabetes do not appear. The rules rather describe patients without heart or other disorders but with elevated diastolic blood pressure and fibrinogen. High cholesterol, age and fibrinogen values appear characteristic for all ischeamatic strokes.

7 Conclusions

This paper has shown that contrast set mining and subgroup discovery are very similar data mining tasks, and has presented approaches to solving a contrast set mining task by decision tree learning and by transforming the contrast set mining problem to a subgroup discovery problem. As shown in [7], the subgroup discovery approach to contrast set mining has several advantages. Its application in brain ischemia data analysis has shown that sometimes the right task to address is one-vs-all contrast set mining rather then the classical round robin formulation of contrast set mining.

Acknowledgment. This work was partially supervised by Blaž Zupan from Faculty of Computer and Information Science, Ljubljana, Slovenia.

References

1. Bay, S.D., Pazzani, M.J.: Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery 5(3), 213–246 (2001)
2. Demšar, J., Zupan, B., Leban, G.: Orange: From experimental machine learning to interactive data mining, white paper. Faculty of Computer and Information Science, University of Ljubljana (2004) www.ailab.si/orange
3. Fürnkranz, J.: Round robin rule learning. In: ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 146–153 (2001)
4. Gramberger, D., Lavrač, N.: Expert-guided subgroup discovery: Methodology and application. Journal of Artificial Intelligence Research 17, 501–527 (2002)
5. Jovanovski, V., Lavrač, N.: Classification rule learning with APRIORI-C. In: Proceedings of the 10th Portuguese Conference on Artificial Intelligence, pp. 44–51 (2001)
6. Kavšek, B., Lavrač, N.: APRIORI-SD: Adapting association rule learning to subgroup discovery. Applied Artificial Intelligence, 543–583 (2006)
7. Kralj, P., Lavrač, N., Gramberger, D., Krstačić, A.: Contrast set mining through subgroup discovery applied to brain ischaemia data. In: PAKDD '07: Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Heidelberg (2007)
8. Kralj, P., Lavrač, N., Zupan, B.: Subgroup visualization. In: IS '05: Proceedings of the 8th International Multiconference Information Society, pp. 228–231 (2005)
9. Kralj, P., Lavrač, N., Zupan, B., Gramberger, D.: Experimental comparison of three subgroup discovery algorithms: Analysing brain ischemia data. In: IS '05: Proceedings of the 8th International Multiconference Information Society, pp. 220–223 (2005)
10. Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research 5, 153–188 (2004)
11. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufman Publishers Inc., San Francisco (1993)
12. Victor, M., Ropper, A.H.: Cerebrovascular disease. Adams and Victor's Principles of Neurology, 821–924 (2001)
13. Webb, G.I., Butler, S., Newlands, D.: On detecting differences between groups. In: KDD '03: Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 256–265 (2003)
14. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Proceedings of the 1st European Conference on Principles of Data Mining and Knowledge Discovery, pp. 78–87. Springer, Heidelberg (1997)

Multi-resolution Image Parametrization in Stepwise Diagnostics of Coronary Artery Disease

Matjaž Kukar[1], Luka Šajn[1,⋆], Ciril Grošelj[2], and Jera Grošelj[2]

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

[2] University Medical Centre Ljubljana, Nuclear Medicine Department,
Zaloška 7, SI-1001 Ljubljana, Slovenia
ciril.groselj@kclj.si

Abstract. Coronary artery disease is one of the world's most important causes of early mortality, so any improvements of diagnostic procedures are highly appreciated. In the clinical setting, coronary artery disease diagnostics is typically performed in a sequential manner. The four diagnostic levels consist of evaluation of (1) signs and symptoms of the disease and ECG (electrocardiogram) at rest, (2) ECG testing during a controlled exercise, (3) myocardial perfusion scintigraphy, and (4) finally coronary angiography (which is considered as the "gold standard" reference method). In our study we focus on improving diagnostic performance of the third diagnostic level (myocardial perfusion scintigraphy). This diagnostic level consists of series of medical images that are easily obtained and the imaging procedure represents only a minor threat to patients' health. In clinical practice, these images are manually described (parameterized) and subsequently evaluated by expert physicians. In our paper we present an innovative alternative to manual image evaluation – an automatic image parametrization on multiple resolutions, based on texture description with specialized association rules, and image evaluation with machine learning methods. Our results show that multi-resolution image parameterizations equals the physicians in terms of quality of image parameters. However, by using both manual and automatic image description parameters at the same time, diagnostic performance can be significantly improved with respect to the results of clinical practice.

Keywords: machine learning, coronary artery disease, medical diagnosis, image parametrization, association rules, stepwise diagnostic process.

1 Introduction

Coronary artery disease (CAD) is one the world's main cause of early mortality, and there is an ongoing research for improving diagnostic procedures. The usual clinical process of coronary artery disease diagnostics consists of four diagnostic steps (levels): (1) evaluation of signs and symptoms of the disease and ECG (electrocardiogram) at rest; (2) ECG testing during the controlled exercise; (3) stress myocardial scintigraphy; and (4) coronary angiography.

[⋆] Parts of work presented in this paper are taken from the second author's doctoral dissertation.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 119–129, 2007.

In this process, the fourth diagnostic level (coronary angiography) is considered as the "gold standard" reference method. As this diagnostic procedure is invasive, comparatively expensive, and potentially dangerous for the patients, there is a tendency to improve diagnostic performance and reliability of earlier diagnostic levels, especially of myocardial scintigraphy [9, 10]. Approaches used for this purpose include applications of neural networks [1], expert systems [7], subgroup mining [6], statistical techniques and rule-based approaches [11]. In our study we focus on different aspects of improving the diagnostic performance of myocardial scintigraphy.

Results of myocardial scintigraphy consist of series of medical images that are taken both during rest and a controlled exercise. These images are readily available in PC or Mac format by respective SPECT cameras and such and imaging procedure does not threaten patients' mostly frail 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 evaluated the images.

In our paper we present an innovative alternative to manual image evaluation – automatic multi-resolution image parametrization, based on texture description with specialized association rules, and image evaluation with machine learning methods. Our results show that multi-resolution image parametrization equals or even betters the physicians in terms of the quality of image parameters. Additionally, by using both manual and automatic image description parameters at the same time, diagnostic performance can be significantly improved with respect to the results of clinical practice.

2 Methods and Materials

2.1 Stepwise Diagnostic Process

Every medical diagnosis inherently contains some uncertainty and is therefore not completely reliable. Sometimes it is crucial to know the magnitude of diagnosis' reliability in order to minimize risks for patient's health or even life.

In a stepwise diagnostic process diagnostic tests are ordered by some pre-determined criteria, such as increasing cost, diagnostic accuracy, and invasiveness (in this order). Key elements of stepwise testing are the estimate of the prior (pre-test) probability of a disease, and the sensitivity and specificity of different diagnostic levels. With this information, 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 (Fig. 1). The process results in a series of tests where

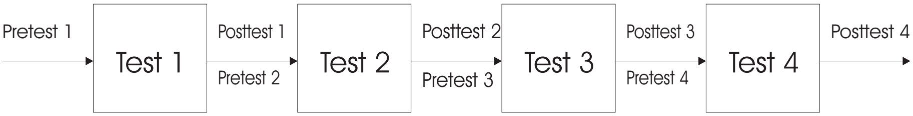


Fig. 1. Increasing diagnostic test levels in stepwise diagnostic process

each test is performed independently. Its results may be interpreted with or without any knowledge of the other test results. In diagnostic problems, the performance of a diagnostic test is described with diagnostic (classification) accuracy (*Acc*), sensitivity (*Se*) and specificity (*Sp*). Test results from earlier levels are used to obtain the final probability of disease. Stepwise diagnostic tests are performed until the post-test probability of disease's presence or absence exceeds some pre-defined threshold value [3].

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 test result the probability $P(d|+) = P(disease|positive\ test\ result)$ is calculated:

$$P(d|+) = \frac{P \cdot Se}{P \cdot Se + (1-P) \cdot (1-Sp)} \tag{1}$$

For negative test result the probability $P(d|-) = P(disease|negative\ test\ result)$ is calculated:

$$P(d|-) = \frac{P \cdot (1-Se)}{P \cdot (1-Se) + (1-P_1) \cdot Sp} \tag{2}$$

The post-test probability after a diagnostic test represents the pre-test probability for the subsequent test. This approach may not only incorporate several test results but also the data from the patient's history [3].

2.2 Image Parametrization

Images in digital form are normally described with matrices which are spatially complex and yet do not offer features that could uniformly distinguish between their predefined classes. Determining image features that can satisfactorily discriminate observed image classes is a hard task for which different algorithms exist. They transform the image from the matrix form into a set of numeric or discrete features (parameters) that convey useful information for discrimination between classes.

The ArTeX Algorithm. The use of association rules used for texture description were first described in [16]. We follow a slightly different approach introduced in [2], where different texture representation and different algorithm for association rules are used.

Fig. 2 illustrates the association rule $(1, 1) \wedge (2, 10) \implies (1, 15) \wedge (3, 5)$, which can be read as follows: if a pixel of intensity 1 is found on distance 1 and a pixel of intensity

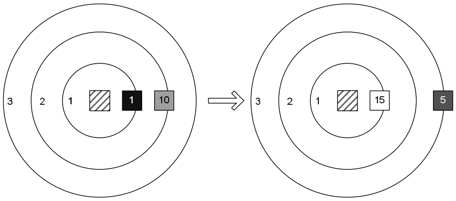


Fig. 2. An illustration of association rule $(1, 1) \wedge (2, 10) \implies (1, 15) \wedge (3, 5)$

10 is found on distance 2, then there is also a pixel of intensity 15 on distance 1 and a pixel of intensity 5 on distance 3.

Using association rules on textures allows to extract a set of features (attributes) for a particular domain of textures. Here is a general description of the ArTeX algorithm:

- *Select a (small) subset of images F for feature extraction.* The subset F can be considerably small. Use at least one example of each typical image in the domain. That is at least one sample per class, or more if the class consists of subclasses.
- *Pre-processing of images in F.* Pre-processing involves the transformation of images to grey scale if necessary, the quantization of grey levels and the selection of proper neighborhood size R. The initial number of grey levels per pixel is usually 256. The quantization process downscales it to say 16 levels per pixel. Typical neighborhood sizes are 3, 4, 5.
- *Generate association rules from images in F.* Because of the representation of texture, it is possible to use any algorithm for association rules extraction. We use the well-known algorithms *Apriori* and *GenRules*.
- *Use generated association rules to extract a set of features.* There are two features associated with each association rule: support and confidence. Use these two attributes of all association rules to construct a feature set. The number of extracted features is twice the number of association rules, which could be quite a lot.

Earlier experiments [18] show excellent results when using ArTeX-type texture descriptions in conjunction with machine learning algorithms. One of the reasons for the success of ArTeX is that it describes images in a multi-resolution and rotation-invariant manner. This parametrization is also invariant to image brightness which is in our case necessary due to varying radiopharmaceutical agent absorption. These features make ArTeX a promising tool for analyzing myocardial scintigrams.

Multi-resolution Parametrization. Algorithms for image parametrization are suitable either for images (imaging some content of different classes) or textures (representing some repeating patterns). Image illumination, scale and affine transformations often obstruct the parametrization. Algorithms use different pixels' properties and relations between them since images in digital representation are described with pixels. Due to the time and space complexity only a predefined size of pixel neighborhood is observed, which makes detectable relations between pixels quite dependent on image resolution. Not only different image scales require appropriate resolutions, but also when there are more shapes of different size present in a picture more resolutions are desired. By using only a single resolution, we may miss the big picture, and proverbially not see the forest because of the trees.

Another issue is the pattern scale. Not every combination of scale and neighborhood size can guarantee that the pattern would be detected. This yields a solution where more resolutions are simultaneously observed in one image and obtained features for each resolution are combined together in one feature vector.

If we want to use more resolutions it is necessary to determine which ones to use. Many existing applications use fixed resolutions irrespectively of the image content and usually three or four are used [5, 14]. Multi-resolution algorithms usually perform better when using only a few resolutions; more resolutions typically yield worse results.

We have developed an algorithm [17] for determining the resolutions for which more informative features can be obtained. The idea for the algorithm is derived from the well known SIFT algorithm [13]. In this way also resolutions for the hearth scintigraphy are determined.

When detecting the appropriate resolutions the image is consequently resized from 100% down to some predefined lowest threshold at some fixed step. At each resize peaks are counted. Peaks are represented by pixels which differ from their neighborhood either as highest or lowest intensity. This algorithm can be implemented also with DOG (Difference-Of-Gaussian) [13] method which improves the time complexity with lower number of actual resizes required to search the entire resolution space.

Detected peak counts are recorded over all resolutions as a histogram. From the histogram the best resolutions are detected at the highest counts. The number of resolutions we want to use in our parametrization is predefined. When there are more equal counts we chose as diverse resolutions as possible [17]. When optimal resolutions are determined, an image parametrization algorithm (Artex in our case, but could be anything) is used to describe images.

2.3 Medical Data

In our study we used a dataset of 288 patients with suspected or known CAD. All patients had performed proper clinical and laboratory examinations, exercise ECG, stress myocardial perfusion scintigraphy (complete image sets were available for analysis), and coronary angiography. Features for the ECG an scintigraphy data were extracted manually by the clinicians. 10 patients were excluded for data pre-processing and calibration required by multi-resolution ArTeX, 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 thousands patients who were examined at the Nuclear Medicine Department between 2001 and 2004. We selected only the patients with complete diagnostic procedures (all four levels), and for whom the imaging data was readily accessible. Some characteristics of the dataset are shown in Tab. 1.

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 in experiments.

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	7
Class distribution	129 (46.40%) 149 (53.60%)		CAD negative CAD positive

It must be noted that our patients represent a highly specific population, since many of them had already had performed cardiac surgery or dilatation of coronary vessels. This clearly reflects the situation in Central Europe with its aging population. It is therefore not surprising that both the population and the predictive performance are considerably different than that of our previous study, where data were collected between years 1991 and 1994 [8]. These differences are a consequence of rapidly progressing interventional cardiology. and are therefore not applicable to the general population, but only to comparable population in developed countries. Similarly, general findings about CAD only partially apply to our population.

Scintigraphic Images. In each patient series of images were taken with the General Electric Millennium SPECT gamma camera, both at rest and after a controlled exercise, thus producing the total of 64 grayscale images in resolution of 64×64 8-bit pixels. Because of patients' movements and partial obscuring the heart by other internal organs, these images are not suitable for further use without heavy pre-processing. For this purpose, a General Electric workstation running eNTEGRA software was used (more specifically we used the Emory Cardiac Toolbox [4]). One of ECToolbox's outputs, a series of 9 polar map (bull's eye) images was used for each patients. Polar maps were chosen because previous work in this field [12] showed that they have useful diagnostic value. The 9 polar map images consist of the following images [4]:

- three raw images (the stress and the rest image, as well as the reversibility image, calculated as a difference between normalized rest and stress images
- three blackout (defect extent) images (which are the stress and the rest image, compared with the respective database of normal images, and suitably processed). Again the reversibility image, calculated as a difference between normalized rest and stress blackout images.
- three standard deviation images that show relative perfusion variance when compared to the respective database of normal images.

An example of polar map images for a typical patient with well-defined CAD is shown in Fig. 3. Unfortunately, in most cases (and especially in our specific population) the differences between images taken during exercise and at rest are not so clear-cut as shown in Fig. 3. Interpretation and evaluation of scintigraphic images therefore requires considerable knowledge and training of expert physicians. Although specialized tools such as the ECToolbox software can aid in this process, they still require a lot of training and medical knowledge for evaluation of results. The aim of our study is to use automatic image parametrization in conjunction with machine learning methods in order to provide additional diagnostic tools.

3 Results

As already mentioned in Sec. 2.3, out of the 288 patients, 10 were excluded for data preprocessing and calibration required by the multi-resolution ArTeX parametrization procedure. These patients were not used in further experiments. The remaining 278 patients with 9 images each were parameterized for three resolutions in advance. The

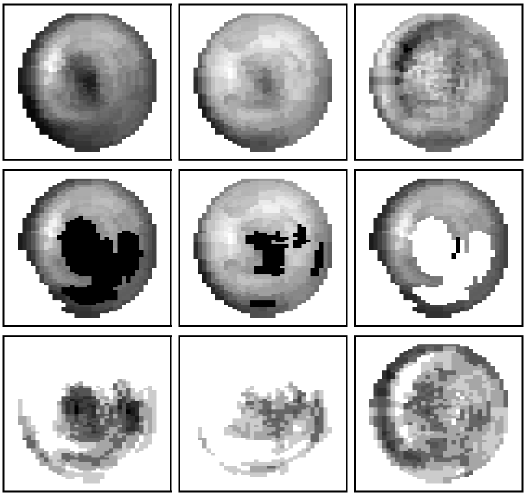


Fig. 3. Typical polar maps taken after exercise (first column), at rest (second column), and their difference (third column). The first row consists of raw images, the second of blackout images, and the third of standard deviations. Black regions indicates less perfused cardiac tissue (a potential defect). Images shown in this figure correspond to the patient with a very clear manifestation of CAD.

proposed three resolutions[1] were 0.95×, 0.80×, and 0.30× of the original resolution, producing together 2944 additional attributes (features). Since this number is too large for most practical purposes, we filtered[2] it to 200 best features as estimated with the ReliefF algorithm[15]. We also did some experiments with other image parametrization approaches such as wavelet and DFT transform, Gabor filters, and combined them with SIFT-like resolution selection; they, however, mostly performed considerably worse than ArTeX. We omit further analysis of these results due to lack of space.

We applied three popular machine learning algorithms: naive Bayesian classifier, support vector machine, and C4.5 (J4.8) decision tree. We performed experiments with Weka [19] machine learning toolkits.

When necessary, continuous attributes were discretized in advance. Testing was performed in the 10-fold cross-validation setting. Aggregated results of the coronary angiography (CAD negative/CAD positive) were used as the class variable.

Experimental results are compared with diagnostic accuracy, specificity and sensitivity of expert physicians after evaluation of scintigraphic images (Tab. 2). The results of clinical practice were validated by careful blind evaluation of images by an independent expert physician.

For machine learning experiments we considered several different settings: evaluation of attributes extracted by physicians; evaluation of all attributes extracted by multi-resolution ArTeX; evaluation of all attributes extracted by multi-resolution ArTeX in

[1] A resolution of 0.30× means $0.30 \cdot 64 \times 0.30 \cdot 64$ pixels instead of 64×64 pixels.

[2] Even better results could be expected if the wrapper approach were used instead of filtering. We chose not to follow this lead for now because of its time consumption.

Table 2. Diagnostic results of the physicians compared with results of machine learning classifiers obtained from the original attributes, as extracted by physicians. Results that are significantly $(p < 0.05)$ different from clinical results are emphasized.

	All basic attributes		
	Accurracy	Specificity	Sensitivity
Physicians	64.00	71.10	55.80
Naive Bayes	**68.34**	69.80	**67.10**
SVM	65.10	62.80	**67.10**
J4.8	**57.19**	53.50	**60.40**

conjunction with attributes, extracted by physicians (Tab. 3); as well as the above variants reduced to 200 best attributes with ReliefF (Tab. 4). Significance of differences to clinical results was evaluated by using the McNemar's test.

From Tab. 2 we can see that machine learning algorithms are approximately on level with expert physicians when evaluating the original data, as collected by physicians. Naive Bayesian classifier even achieves significantly higher diagnostic accuracy and slightly lower sensitivity than physicians, while the J4.8 decision tree achieves significantly lower diagnostic accuracy. However, for physicians, improvements of specificity are more important than improvements of sensitivity or overall accuracy, since increased specificity decreases the number of unnecessarily performed higher-level diagnostic tests, and consequently shorter waiting times for truly ill patients.

Table 3. Experimental results of machine learning classifiers on parameterized images obtained by using all available attributes. Results that are significantly $(p < 0.05)$ better than clinical results are emphasized.

	All image and basic attributes			All image attributes		
	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity
Physicians	64.00	71.10	55.80	64.00	71.10	55.80
Naive Bayes	**70.50**	**72.10**	**69.10**	**70.14**	**72.10**	**68.50**
SVM	**69.40**	69.80	**69.10**	61.15	58.10	**63.80**
J4.8	65.10	60.50	**69.10**	59.71	63.80	55.00

In Tab. 3 we can see that some machine learning algorithms have difficulties when handling a huge number (2944) of attributes, with only 278 learning examples. This can lead to overfitting the learning data and thus lower their diagnostic performance. Only naive Bayesian classifier is significantly better than physicians when using all 2944 attributes. However, using these 2944 attributes together with the original attributes invariably improves the physicians' results, in two of three cases even significantly.

Tab. 4 depicts the situation where machine learning algorithms considerably benefit from attribute filtering. In all cases the results are significantly better than the results of physicians. Especially nice results are that of naive Bayesian classifier, which improves diagnostic accuracy, sensitivity and specificity.

Table 4. Experimental results of machine learning classifiers on parameterized images obtained by selecting only the best 200 attributes. Results that are significantly better ($p < 0.05$) than clinical results are emphasized.

	200 best image and basic attributes			200 best image attributes		
	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity
Physicians	64.00	71.10	55.80	64.00	71.10	55.80
Naive Bayes	**74.10**	**79.80**	**69.10**	**72.30**	**79.80**	**65.80**
SVM	**69.42**	65.90	**72.50**	**70.14**	**72.90**	**67.80**
J4.8	**67.62**	63.60	**71.10**	**68.34**	63.60	**72.50**

We also experimented with machine learning classifiers in stepwise process, as shown in Fig. 1 and described in Sec. 2.1. By the stepwise diagnostic process, after the third diagnostic level we get the following percentages reliable diagnoses (post-test probability ≥ 0.90), which are practically the same as the results of expert physicians:

- 30.94% reliable true positive diagnoses and 10.50% erroneously reliable false positive diagnoses
- 19.88% reliable true negative diagnoses and 7.83% erroneously reliable false negative diagnoses

Our preliminary experiments show, that by using additional attributes from parameterized images, we can increase the number of reliable positive and negative diagnoses by almost 10% while keeping the number of incorrect diagnoses lower than the physicians in clinical practice.

4 Discussion

Although our study is still in early stages, the results are promising. We have shown that multi-resolution ArTeX parametrization in conjunction with machine learning techniques can be successfully used as an intelligent tool in image evaluation, as well as as a part of the stepwise diagnostic process. Automatic image parametrization and machine learning methods can help less experienced physicians evaluate medical images and thus improve their combined performance (in terms of accuracy, sensitivity and specificity).

From the practical use of described approaches two-fold improvements of the diagnostic procedure can be expected. Due to higher specificity of tests (by almost 9%), fewer patients without the disease would have to be examined with coronary angiography which is invasive and therefore dangerous method. Together with higher sensitivity this would also save money and shorten the waiting times of the truly ill patients.

The most significant result of our study may well be the improvement in the predictive power of the stepwise diagnostic process. The almost 10% improvement of positive and negative patients who would not need to be examined with costly further tests, represents 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.

However, it should be emphasized that the results of sour 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 of the University Clinical Centre in Ljubljana.

Acknowledgements

This work was supported by the Slovenian Ministry of Higher Education, Science, and Technology.

References

[1] Allison, J.S., Heo, J., Iskandrian, A.E.: Artificial neural network modeling of stress single-photon emission computed tomographic imaging for detecting extensive coronary artery disease. J Nucl Cardiol 95(2), 178–181 (2005)

[2] Bevk, M., Kononenko, I.: Towards symbolic mining of images with association rules: Preliminary results on textures. Intelligent Data Analysis (2006)

[3] 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)

[4] General Electric. ECToolbox Protocol Operator's Guide (2001)

[5] Ferreira, C.B.R., Borges, D.L.: Automated mammogram classification using a multi-resolution pattern recognition approach. In: SIBGRAPI01, vol. 00, pp. 76 (2001)

[6] 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)

[7] 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)

[8] Grošelj, C., Kukar, M., Fettich, J., Kononenko, I.: Impact of machine learning to the diagnostic certainty of the patient's group with low coronary artery disease probability. In: Proc. Computer-Aided Data Analysis in Medicine, Bled, Slovenia, pp. 68–74 (1997)

[9] Kukar, M.: Transductive reliability estimation for medical diagnosis. Artif. intell. med. 81–106 (2003)

[10] 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)

[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] Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)

[15] Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning 53, 23–69 (2003)

[16] Rushing, J.A., Ranganath, H.S., Hinke, T.H., Graves, S.J.: Using association rules as texture features. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(8), 845–858 (2001)

[17] Šajn, L.: Multiresolution parameterization for texture classification and its application in analysis of scintigrafic images. PhD thesis, Faculty of Computer and Information Science, University of Ljubljana, in Slovene (2007)

[18] Šajn, L., Kukar, M., Kononenko, I., Milčinski, M.: Computerized segmentation of whole-body bone scintigrams and its use in automated diagnostics. Computer Methods and Programs in Biomedicine 80(1), 47–55, 10 (2005)

[19] Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

Classifying Alarms in Intensive Care - Analogy to Hypothesis Testing

Wiebke Sieben and Ursula Gather

Department of Statistics, University of Dortmund, 44227 Dortmund, Germany
sieben@statistik.uni-dortmund.de,
gather@statistik.uni-dortmund.de

Abstract. Monitoring devices in intensive care units observe a patient's health status and trigger an alarm in critical situations. The alarm rules in commercially available monitoring systems are usually based on simple thresholds set by the clinical staff. Though there are some more advanced alarm rules integrated in modern monitoring devices, even for those, the false alarm rate is very high. Decision trees have proven suitable for alarm classification and false alarm reduction. Random forests which are ensembles of trees can improve the accuracy compared to single trees in many situations. In intensive care, the probability of misclassifying a situation in which an alarm is needed has to be controlled. Subject to this constraint the probability of misclassifying a situation in which no alarm should be given has to be minimized - an analogy to a hypothesis test for testing "situation is alarm relevant" vs. "situation is non alarm relevant" based on an ensemble of trees. This yields a classification rule for any given significance level, which is the probability of misclassifying alarm relevant situations. We apply this procedure to annotated physiological data recorded at an intensive care unit and generate rules for false alarm reduction.

1 Introduction

Monitoring devices in intensive care units observe a patient's health status and trigger an alarm in critical situations. The alarm rules in commercially available monitoring systems are mostly simple thresholds set by the clinical staff. There are some more advanced alarm rules integrated in modern monitoring devices but even for those, the false alarm rate still is very high (Lawless 1994, Tsien, Fackler [1997], Chambrin [2001]). This reduces the attention the clinical staff pays to alarms. As a reaction, alarm limits are set very wide so that it is unlikely that a physiological measurement will cross the limit and trigger an alarm. Occasionally the alarm function is even deactivated completely for some physiological parameters. By this, the alarm system literally loses its meaning.

Intelligent alarm rules are needed to reduce the high false positive rate of commercially available monitoring systems. Artificial Intelligence methods have already been successfully applied in the context of intensive care (Imhoff et al. [2006]). For example, Gather et al. [2000] used support vector machines in

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 130–138, 2007.

combination with time series analysis for medicational suggestions. Tsien [2000] and Zhang [2003] both applied decision trees and neural networks for event detection in intensive care monitoring. One major problem in event detection in intensive care is the urgently required high classification accuracy. In case of identifying false alarms, the rate of misclassifying true alarms needs to be close to zero.

Reducing the number of false positive alarms can be achieved by validating the alarms of conventional monitoring devices. In terms of the underlying classification problem, situations indicated by an alarm are the objects that will be classified as “non alarm relevant” or “alarm relevant” by an ensemble of trees. In Sect. 2 the basic concept of trees and ensembles of trees (forests) is presented. Based on a forest, a procedure is proposed in Sect. 3 in analogy to a statistical test. This procedure implies a classification rule for any given significance level. We apply this procedure to annotated physiological data recorded at an intensive care unit and generate rules for false alarm reduction (Sect. 4).

2 Trees and Random Forests

2.1 Notation

We consider a population $\mathit{\Pi}$ of all situations indicated by an alarm that is a union of the disjoint populations $\mathit{\Pi}_0$ of all alarm relevant situations and $\mathit{\Pi}_1$ of all non alarm relevant situations. Every situation is characterized by a measurement x of a p-dimensional random vector X of physiological parameters. Classification means assigning an object to one of these populations according to the observed x. This corresponds to a partition of the sample space $\mathcal{X}$ into regions B_0 and B_1. An object with realization x in B_i is assigned to population $\mathit{\Pi}_i$, $(i = 0, 1)$.

Assuming that both population membership and the vector of physiological measurements are random the problem can be formalized as follows:

Population membership. Let the random variable G be the population membership of an object, G taking values $\{0, 1\}$, with priors π_i $(i = 0, 1)$.

Physiological parameters of an object from population $\mathit{\Pi}_i$. The random vector X containing physiological parameters of an object from population $\mathit{\Pi}_i$ is a measurable mapping into the sample space $\mathcal{X}$, with density f_i.

We seek a decision rule $\delta : \mathcal{X} \to \{0, 1\}$, that assigns every observation x to one of the populations:

$$\delta(x) = \begin{cases} 0 & : \quad x \in B_0 \\ 1 & : \quad x \in B_1 \end{cases}, \quad B_0 \,\dot{\cup}\, B_1 = \mathcal{X}.$$

When building decision trees, these regions are found by *Recursive Partitioning.*

2.2 Decision Trees

Recursive Partitioning results in a *tree* which is a directed, acyclic, finite graph. Every node is a subset of $\mathcal{X}$ and $\mathcal{X}$ itself is a node, called the root node, having no incoming edges. Every other node in the tree has exactly one incoming edge and no or at least two outgoing edges. Nodes without outgoing edges are called leaf nodes. A *split* divides a node into two or more child nodes which form a partition of the parent node.

Several algorithms are available to perform Recursive Partitioning. They split the sample space into disjoint subsets (child nodes) according to some *splitting rule*. All possible splits on all variables are compared regarding the gain of purity in the child nodes with respect to the purity of the parent node. The splitting rule chooses that variable on which splitting improves the purity the most. The resulting subsets are then split in the same way. Splitting is stopped when a suitable *stop splitting rule* suggests that further splitting is of no great worth. The three most popular algorithms are Chi Square Automatic Interaction Detection (CHAID) (Kass [1980]), Classification and Regression Trees (CART) (Breiman et al. [1984]) and C4.5 (Qinlan [1993]). The CHAID algorithm was designed for categorial data. Continuous variables are categorized. The splitting rule involves a χ^2-test and splitting is stopped if the p-values are too large. This algorithm allows multiway splitting. CART and C4.5 can handle categorial and continuous variables. The CART algorithm produces binary trees, while C4.5 creates a subset for every category when splitting on categorial variables. There are only minor differences between CART and C4.5 in the case of continuous and binary variables as observed in intensive care. From these two algorithms, we choose the CART algorithm because of Breiman's more general approach to purity of nodes (see below) compared to Quinlan's restriction to measuring the impurity in terms of the entropy.

In a two population classification problem an impurity function is a function ϕ defined on all pairs (p_0, p_1) satisfying $p_0, p_1 \geq 0$ and $p_0 + p_1 = 1$ with the properties

ϕ has its only maximum at (0.5, 0.5),
ϕ has its minimum only at (0, 1) and (1, 0),
ϕ is symmetric.

Define the proportions $p_{j|t}$, $j = 0, 1$, to be the proportions of objects in node t belonging to population j so that $p_{0|t} + p_{1|t} = 1$. The function $i(t) = \phi(p_{0|t}, p_{1|t})$ measures the impurity of node t. A split s divides a node t containing n objects into a left child node t_l containing n_l objects and a right child node t_r containing n_r objects. Then, the decrease in impurity of split s is defined by

$$\triangle i(s,t) = i(t) - \frac{n_l}{n} i(t_l) - \frac{n_r}{n} i(t_r).$$

Examples for impurity measuring functions $i(t)$ are

- Gini index: $i(t) = 2p_{0|t}p_{1|t}$
- Entropy: $i(t) = -p_{0|t}log(p_{0|t}) - p_{1|t}log(p_{1|t})$.

The splitting rule selects the split $s^\star$ that maximizes the decrease in impurity. Stop splitting rules can involve a minimal required nodes size after splitting or a minimal decrease in impurity.

The union of all leaf nodes with a majority of objects from population Π_0 defines region B_0 and the union of all other leaf nodes defines region B_1.

2.3 Random Forests

A Random Forest (Breiman [2001]) is an ensemble of trees. Every tree is grown on an independently and randomly chosen subset of the learning sample. When searching for the best split not all but a fixed number of randomly selected variables are considered. An object is classified by every single tree which is understood as a vote for its population membership. The object is democratically assigned to that population that gets the majority of votes.

Let $\delta_j(x)$ denote the j-th of N trees in the forest

$$\delta_j(x) = \begin{cases} 0 & : \quad x \in B_0^j \\ 1 & : \quad x \in B_1^j \end{cases}, \quad j = 1, ..., N,$$

then the forest $\delta_{forest}(x)$ is

$$\delta_{forest}(x) = \begin{cases} 0 & : \quad \text{if } \sum \delta_j(x) < N/2 \\ 1 & : \quad \text{if } \sum \delta_j(x) \geq N/2 \end{cases}.$$

Generally, the voting need not be strictly democratic. Especially in intensive care monitoring, consequences of misclassification are not equally severe in both populations. We must control the probability of misclassification and achieve nearly perfect classification of alarm relevant situations. Having this in mind, one can try to improve the classification rate in the population of non alarm relevant situations. Essentially, this is a statistical hypothesis test.

3 Analogy to Statistical Hypothesis Tests

We in a sense test the null hypothesis that a clinical situation is alarm relevant against the alternative that the situation is non alarm relevant:

$$H_0: \text{ object belongs to } \Pi_0 \text{ (alarm relevant)}$$

vs.

$$H_1: \text{ object belongs to } \Pi_1 \text{ (non alarm relevant).}$$

For a given significance level α, which is the probability of wrongly rejecting H_0, find the critical value q for

$$\delta_{forest}(x) = \begin{cases} 0 & : \quad \text{if } \sum \delta_j(x) < q \\ 1 & : \quad \text{if } \sum \delta_j(x) \geq q \end{cases},$$

so that $P(\delta_{forest} = 1|G = 0) \leq \alpha$. The critical value q is the $(1-\alpha)$-quantile of the distribution of our test statistic $\sum \delta_j(x)$ under H_0 as

$$\begin{aligned} & P(\delta_{forest} = 1|G = 0) \leq \alpha \\ \Leftrightarrow \quad & P(\textstyle\sum \delta_j(x) > q|G = 0) \leq \alpha \\ \Leftrightarrow \; & P(\textstyle\sum \delta_j(x) \leq q|G = 0) \geq 1 - \alpha. \end{aligned}$$

4 Generating Alarm Rules

From a clinical study conducted at the University Hospital Regensburg following measurements are recorded at a sample rate of one per second:

- respiration rate, oxygen saturation
- arrhythmia indicator
- heart rate, pulse, premature ventricular contraction
- arterial systolic, diastolic and mean blood pressure
- temperature.

Additionally, the threshold settings and alarm information from the monitoring system are provided. Every alarm given by the monitoring system at the bedside is annotated by a clinician as alarm relevant or non alarm relevant. It is also noted whether an alarm was induced by manipulation. Only alarms without manipulation are included in the data set as manipulation implies the presence of clinical staff. The presence of clinical staff affects the judgement of the annotating clinician and shifts the measurements into an abnormal state that needs not to be automatically detected because it is induced by the present person. Data preprocessing covers missing value imputation and the generation of indicator variables for missing values. The data set is randomly divided into three sets: learning set, estimation set and test set.

The forest is grown on the learning set. It consists of 1000 trees, each built on 200 objects selected randomly without replacement. Splits are restricted to 5 randomly chosen variables and the minimum node size is set at 5 objects. The unknown distribution of the test statistic is estimated by the empirical distribution function. Therefore, every object from the estimation set that is annotated as alarm relevant is dropped down the trees in the forest and the votes for non alarm relevant are counted. The (1-α)-quantile of these counts is the critical value for the test. Then, the objects from the complete test sample are dropped down every tree in the forest and the votes are compared to the critical value. We assign an object to the population of alarm relevant situations if less than q trees vote for non alarm relevant.

The results clearly depend on the partition into learning, estimation and test set. To see if this procedure works in general, the data set is divided 1050 times randomly into these sets and the proposed procedure is applied to each of these partitions. The obtained sensitivities and false alarm reductions on the test sets (illustrated in Fig. 1) allow us to analyze its performance. We first consider a significance level of 5 percent. On the test sets, the mean and median sensitivities of 94.7 and 95.0 respectively achieve the target of 95 percent sensitivity almost exactly. Half of the forests have a sensitivity between 93.4 and 96.3 percent.

Only few forests achieve a sensitivity below 90 percent (Fig. 2). The mean and median false alarm reductions are 45.6 and 45.8 percent respectively. Half of the forests are able to reduce the number of false alarms by 41.2 to 49.9 percent.

As higher sensitivities usually go along with lower false alarm reductions we seek a forest with a good combination of both. For the given significance level of 5 percent, one of the best performing forests, representing several equally good

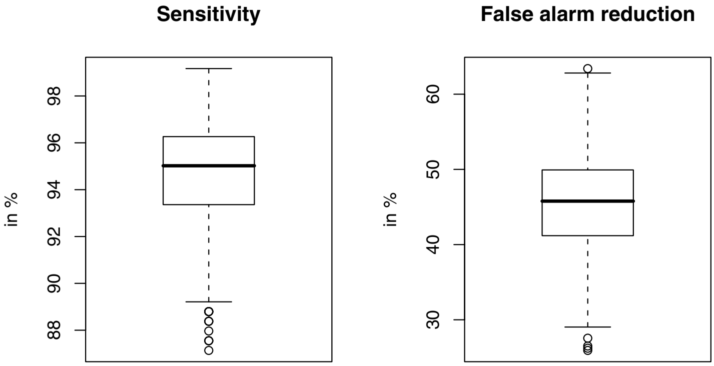


Fig. 1. Sensitivities and false alarm reductions at a significance level of 5 %

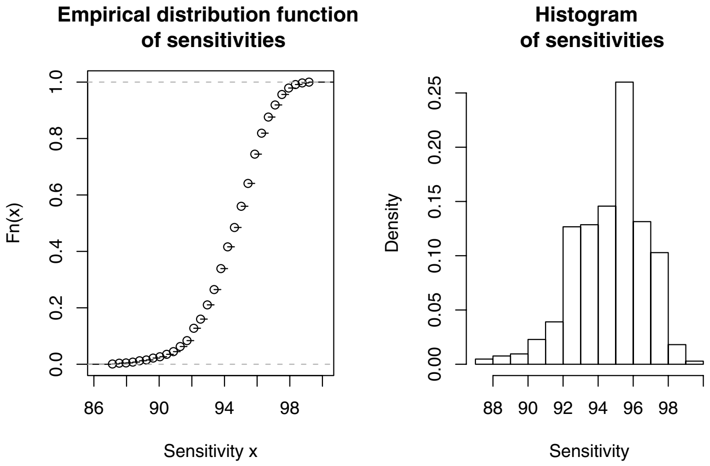


Fig. 2. Empirical distribution function and histogram of sensitivities at a significance level of 5 %

candidates, classifies 228 of 241 alarm relevant situations in the test set correctly which is a sensitivity of 94.6 percent. It classifies 381 of 675 non alarm relevant situations correctly and so reduces the false alarms by 56.4 percent (Tab. 1). The critical value for this forest is 826, which means that at least 826 of 1000 trees need to vote for non alarm relevant so that the forest classifies an object as non alarm relevant.

A sensitivity of 95 percent might not be adequate in patient monitoring. A significance level of 2 percent, however, yields a considerably lower false alarm reduction. Again, 1050 forests are constructed with a target sensitivity of 98 percent. The median and mean achieved sensitivities on the test sets are 97.9 and 97.6 respectively. On the average, false alarms are reduced by 30.1 percent.

Table 1. Confusion matrix of one of the best performing forests, significance level 5 %

	classified as	
situation	alarm relevant	non alarm relevant
alarm relevant	228	13
non alarm relevant	249	381

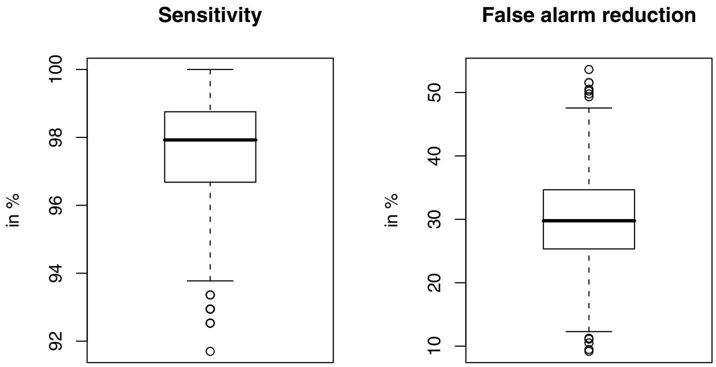


Fig. 3. Sensitivities and false alarm reductions at a significance level of 2 %

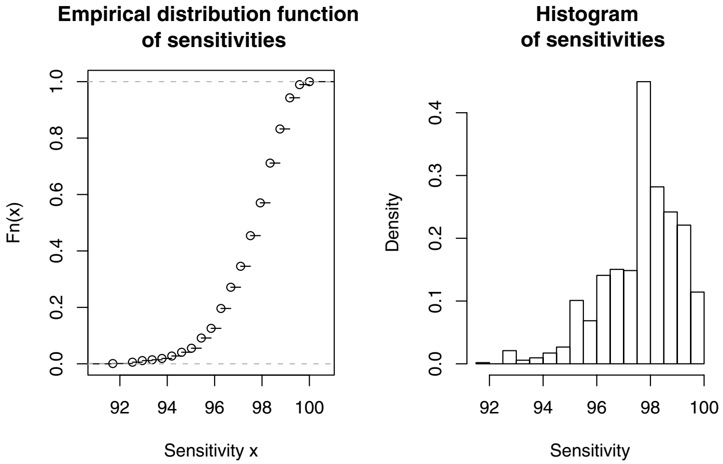


Fig. 4. Empirical distribution function and histogram of sensitivities at a significance level of 2 %

The range of achieved sensitivities is narrow while still reasonable reductions of false alarms are possible (see Fig. 3). The majority of forests have a sensitivity between 96 and 100 percent (Fig. 4). One of the best performing forests classifies

Table 2. Confusion matrix of one of the best performing forests, significance level 2 %

	classified as	
situation	alarm relevant	non alarm relevant
alarm relevant	237	4
non alarm relevant	451	224

237 of 241 alarm relevant situations correctly (sensitivity 98.3 percent) and at the same time detects 224 of 675 false alarms (Tab. 2).

Using this random forest with a critical value of 917 requiring that 917 of 1000 trees vote for non alarm relevant to classify an object as such reduces the false alarms by 33.2 percent.

5 Discussion

In intensive care, frequently occurring false alarms from the bedside monitoring system distract and annoy the clinical staff. To reduce this high false alarm rate, a procedure is proposed that validates the alarms generated by a commercially available monitoring system. Misclassifying an alarm relevant situation might have severe consequences and so the probability of misclassifying these situations needs to be controlled. Subject to this constraint, the proportion of correctly classified non alarm relevant situations has to be maximal. A standard approach to improve the classification accuracy in one class is to use a loss function. The costs for misclassifying alarm relevant situations should then be chosen to be quite high compared to the costs for misclassifying non alarm relevant situations. As it is generally not known which costs exactly are to be chosen to achieve a prespecified classification accuracy, different loss functions must be tried to find a suitable one. Controlling the probability of misclassifying alarm relevant situations is only indirectly possible. The proposed procedure allows us to control this probability for any given significance level.

The procedure consists of a random forest classification that is understood as a statistical hypothesis test. The test statistic is the sum of votes for non alarm relevant from the trees in the forest. For any given significance level, the critical value is estimated by a quantile of the empirical distribution function of this test statistic. Objects are classified by the random forest as non alarm relevant if the number of trees voting for non alarm relevant is at least the critical value.

This procedure is applied to data from an intensive care unit. In a clinical study, alarms were collected along with data on physiological parameters and monitor settings. Additionally, annotations from a clinician are available who labelled the alarm situations as alarm relevant or non alarm relevant. The data set is randomly divided into learning, estimation and test sets 1050 times and forests are constructed. The results from these 1050 forests show that it is possible to control the probability of misclassifying alarm relevant situations with this procedure. The target sensitivities of 95 and 98 percent are achieved on the average. At the same time, false alarms are reduced by 45 and 30 percent on the average.

Acknowledgement

The support of the German Science Foundation (Deutsche Forschungsgemeinschaft, Collaborative Research Center 475, Reduction of Complexity in Multivariate Data Structures) is gratefully acknowledged. Moreover, we are grateful to unknown referees for their valuable comments. We wish to thank Dr. Christian Wrede and Dr. Sylvia Siebig, Department of Internal Medicine I, University Hospital Regensburg, for their intense commitment to conducting the clinical study that forms the basis for this research.

References

[2001] Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001)

[1984] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman & Hall, New York (1984)

[2001] Chambrin, M.-C.: Alarms in the intensive care unit: How can the number of false alarms be reduced? Critical Care 4, 184–188 (2001)

[2000] Gather, U., Morik, K., Imhoff, M., Brockhausen, P., Joachims, T.: Knowledge discovery and knowledge validation in intensive care. Artificial Intelligence in Medicine 19, 225–249 (2000)

[2006] Imhoff, M., Kuhls, S.: Alarm Algorithms in Critical Care Monitoring. Anesthesia & Analgesia 102, 1525–1537 (2006)

[1980] Kass, G.V.: An exploratory technique for investigating large quantities of categorial data. Journal of the Royal Statistical Society: Series C (Applied Statistics) 2, 119–127 (1980)

[1994] Lawless, S.T.: Crying wolf: False alarms in a pediatric intensive care unit. Critical Care Medicine 22 (6), 981–985 (1994)

[1993] Quinlan, J.R.: C4.5, Programms in Machine Learning. Morgan Kaufmann Series in Machine Learning, San Mateo, California (1993)

[1997] Tsien, C.L., Fackler, C.: Poor prognosis for existing monitors in the intensive care unit. Critical Care Medicine 25 (4), 614–619 (1997)

[2000] Tsien, C.L.: TrendFinder: Automated detection of alarmable trends. MIT Ph.D. dissertation, Massachusetts Institute of Technology (2000)

[2003] Zhang, Y.: Real-time analysis of physiological data and development of alarm algorithms for patient monitoring in the intensive care unit, MIT EECS Master of Engineering Thesis, Massachusetts Institute of Technology (2003)

Hierarchical Latent Class Models and Statistical Foundation for Traditional Chinese Medicine

Nevin L. Zhang[1], Shihong Yuan[2], Tao Chen[1], and Yi Wang[1]

[1] Hong Kong University of Science and Technology, Hong Kong, China
{lzhang,csct,wangyi}@cs.ust.hk
[2] Beijing University of Traditional Chinese Medicine, Beijing, China
yuanshih@yahoo.com.cn

Abstract. The theories of traditional Chinese medicine (TCM) originated from experiences doctors had with patients in ancient times. We ask the question whether aspects of TCM theories can be reconstructed through modern day data analysis. We have recently analyzed a TCM data set using a machine learning method and found that the resulting statistical model matches the relevant TCM theory well. This is an exciting discovery because it shows that, contrary to common perception, there are scientific truths in TCM theories. It also suggests the possibility of laying a statistical foundation for TCM through data analysis and thereby turning it into a modern science.

1 Introduction

In TCM Diagnosis, patient information is collected through an overall observation of symptoms and signs rather than micro-level laboratory tests. The conclusion of TCM diagnosis is called *syndrome* and the process of reaching a diagnostic conclusion from symptoms is called *syndrome differentiation*. There are several syndrome differentiation systems, each focusing on a different perspective of the human body and with its own theory. The theories describe relationships between syndrome factors and symptoms, as illustrated by this excerpt:

> KIDNEY YANG [1] (Yang *et al.* 1998) is the basis of all YANG in the body. When KIDNEY YANG is in deficiency, it cannot warm the body and the patient feels cold, resulting in intolerance to cold, cold limbs, and cold lumbus and back. Deficiency of KIDNEY YANG also leads to SPLEEN disorders, resulting in loose stools and indigested grain in the stool.

Here syndrome factors such as KIDNEY YANG FAILING TO WARM THE BODY and SPLEEN DISORDERS DUE TO KIDNEY YANG DEFICIENCY are not directly observed. They are similar in nature to concepts such as 'intelligence' and are indirectly measured through their manifestations. Hence we call them *latent variables*. In contrast, symptom variables such as 'cold limbs' and 'loose stools' are directly observed

[1] Words in small capital letters are reserved for TCM terms.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 139–143, 2007.

and we call them *manifest variables*. TCM theories involve a large number of latent and manifest variables. Abstractly speaking, they describe relationships among latent variables, and between latent variables and manifest variables. Hence they can be viewed as *latent structure models* specified in natural language.

TCM is an important avenue for disease prevention and treatment for ethnic Chinese and is gaining popularity among others. However, it suffers a serious credibility problem especially in the west. One reason is the lack of rigorous randomized trials in support for the efficacy of TCM herb treatments (Normile 2003). Another equally important reason, on which this paper focuses, is the lack of scientific validations for TCM theories. Researchers in China have been searching for such validations in the form of laboratory tests for more than half a century, but there has been little success. We propose and investigate a statistical approach. In the next three paragraphs, we explain the premise and the main idea of the approach.

We human beings often invoke latent variables to explain regularities that we observe. Here is an experience that many might share. I (the first author) was looking at some apartment buildings nearby one night. I noticed that, for a period of time, the lighting from several apartments was changing in brightness and color at the same times and in perfect synchrony. This caught my attention and my brain immediately concluded that there must be a common cause that was responsible for changes. My brain did so without knowing what the common cause was. So, a latent variable was introduced to explain the regularity that I observed. What I tried to do next was to find the identity of the latent variable.

We conjecture that, in a similar vein, latent syndrome variables in TCM were introduced to explain observed regularities about the occurrence of symptoms. Take the concept KIDNEY YANG FAILING TO WARM THE BODY as an example. We believe that in ancient times it was first observed that symptoms such as intolerance to cold, cold limbs, and cold lumbus and back often occur together in patients, and then, to explain the phenomenon, the latent variable KIDNEY YANG FAILING TO WARM THE BODY was created.

When explaining the phenomenon of synchronous change in lighting, I resorted to my knowledge about the world and concluded that the common cause must be that residents in those apartments were watching the same TV channel. Similarly, when explaining patterns about the occurrence of symptoms, ancient Chinese resorted to their understanding of the world and the human body. This explains why concepts from ancient Chinese philosophy such as YIN and YANG are prevalent in TCM theories. Words such as KIDNEY and SPLEEN also appear in TCM theories because there was primitive anatomy in ancient times. However, the functions that TCM associates with KIDNEY and SPLEEN are understandably different from the functions of kidney and spleen in modern western medicine.

Thus, the premise of our work is that TCM theories originated from regularities ancient Chinese doctors observed in their experiences with patients. The main idea of our approach, called *the latent structure approach*, is to collect patient symptom data systematically, analyze the data based on statistical principles, and thereby obtain mathematical latent structure models. If the mathematical latent structure models match the relevant aspects of TCM theories, then we would have validated those aspects of TCM theories statistically. A case study has been conducted to test the idea. In the following, we describe the case study and report the findings.

2 Data and Data Analysis

The data set used in the case study involves 35 symptom variables, which are considered important when deciding whether a patient suffers from the so-called KIDNEY DEFICIENCY syndrome, and if so, which subtype. Each variable has four possible values: none, light, medium, and severe. The data were collected from senior citizen communities, where the KIDNEY DEFICIENCY syndrome frequently occurs. There are totally 2,600 records. Each record consists of values for the 35 symptom variables, but there is no information about syndrome types.

We refer the relevant TCM theory that explains the occurrence of the 35 symptoms as the *TCM* KIDNEY *theory*. As mentioned earlier, this is a latent structure model specified in natural language. The objective of the case study is to induce a mathematical latent structure model from the data based on statistical principles and compare it with the TCM KIDNEY theory to see whether and how well they match.

The statistical models used in the case study are called hierarchical latent class (HLC) models (Zhang 2004), which were developed specifically for latent structure discovery. An HLC model is a rooted tree where each node represents a random variable. The leaf nodes represent manifest variables, while the internal nodes represent latent variables. Quantitative information includes a marginal probability distribution for the root variable and, for each of the other variables, a conditional probability distribution for the variable given its parent. The quality of an HLC model with respect to a data set is determined by the Bayesian information criterion (BIC) (Schwarz 1978). According to this widely used model selection principle, a good model should fit the data well, that is, explain the regularities well, and should be as simple as possible. To find a model with a high BIC score, one can search in the space of all possible HLC models. The current state-of-the-art is an algorithm known as HSHC (Zhang and Kocka 2004).

The KIDNEY data were analyzed using the HSHC algorithm. The best model that we obtained is denoted by M. Its BIC score is -73,860 and its structure is shown in Fig. 1. In the model, Y_0 to Y_{34} are the manifest variables that appear in the data, while X_0 to X_{13} are the latent variables introduced in the process data analysis.

3 Latent Variables

We now set out to compare the structure of model M with the TCM KIDNEY theory. According to the semantics of HLC models, the left most part of model M states that there is a latent variable X_1 that is (1) directly related to the symptoms intolerance to cold (Y_2), cold lumbus and back (Y_3), and cold limbs

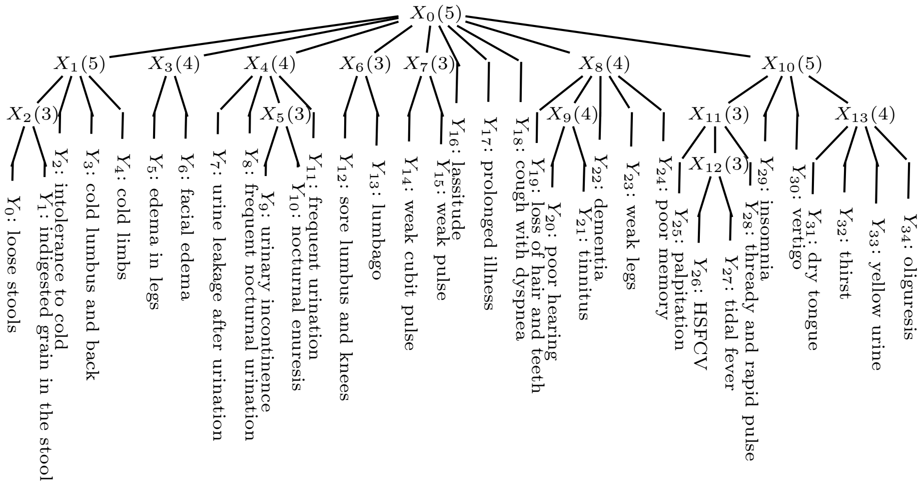


Fig. 1. The structure of the best model M found for KIDNEY data. The abbreviation HSFCV stands for Hot Sensation in Five Centers with Vexation, where the five centers refer to the centers of two palms, the centers of two feet, and the heart. The integer next to a latent variable is the number of possible states of the variable.

(Y_4); and (2) through another latent variable X_2 indirectly related to loose stools (Y_0) and indigested grain in the stool (Y_1). On the other hand, the TCM KIDNEY theory asserts that when KIDNEY YANG is in deficiency, it cannot warm the body and the patient feels cold, resulting in manifestations such as cold lumbus and back, intolerance to cold, and cold limbs. Deficiency of KIDNEY YANG also leads to SPLEEN disorders, resulting in symptoms such as loose stools and indigested grain in the stool. Here, we have a good match between model M and the TCM KIDNEY theory. The latent variable X_1 can be interpreted as KIDNEY YANG FAILING TO WARM THE BODY, while X_2 can be interpreted as SPLEEN DISORDERS DUE TO KIDNEY YANG DEFICIENCY (KYD).

According to the TCM KIDNEY theory, clinical manifestations of the KIDNEY ESSENCE INSUFFICIENCY syndrome includes premature baldness, tinnitus, deafness, poor memory, trance, declination of intelligence, fatigue, weakness, and so on. Those match the symptom variables in model M that are located under X_8 fairly well and hence X_8 can be interpreted as KIDNEY ESSENCE INSUFFICIENCY. The clinical manifestations of the KIDNEY YIN DEFICIENCY syndrome includes dry throat, tidal fever or hectic fever, fidgeting, hot sensation in the five centers, insomnia, yellow urine, rapid and thready pulse, and so on. Those match the symptom variables under X_{10} fairly well and hence X_{10} can be interpreted as KIDNEY YIN DEFICIENCY. Similarly, X_3 can be interpreted EDEMA DUE TO KYD and X_4 can be interpreted as KIDNEY FAILING TO CONTROL UB, where UB stands for the urinary bladder.

It is very interesting that some of the latent variables in model M correspond to syndrome factors such as KIDNEY YANG FAILING TO WARM THE BODY, SPLEEN

DISORDERS DUE TO KYD, EDEMA DUE TO KYD, KIDNEY FAILING TO CONTROL UB, KIDNEY ESSENCE DEFICIENCY, and KIDNEY YIN DEFICIENCY, as each of them is associated with only a subset of the symptom variables in the TCM KIDNEY theory. As the latent variables were introduced by data analysis based on a statistical principle, the case study had provided statistical validation for the introduction of those syndrome factors to the TCM KIDNEY theory and for what are asserted about their relationships with symptom variables.

4 Latent Classes

By analyzing the KIDNEY data using HLC models, we have not only obtained a latent structure, but also clustered the data in multiple ways. For example, the latent variable X_1 has 5 states. This means that the data has in one way been grouped into 5 clusters, with one cluster corresponding to each state of X_1. We have examined the meaning of those *latent classes* and found that they, like the latent variables, provide statistical validations for aspects of the TCM KIDNEY theory. The reader is referred to a longer version of the paper for the details.

5 Conclusion

The TCM KIDNEY theory was formed in ancient times, while model M was obtained through modern day data analysis. It is very interesting that they match each other well. This shows that, contrary to popular perception, there are scientific truths in TCM theories. It also suggests the possibility of laying a statistical foundation for TCM through data analysis.

Acknowledgement

Research on this work was supported by Hong Kong Grants Council Grant #622105 and The National Basic Research Program (aka the 973 Program) under project No.2003CB517106.

References

1. Normile, D.: The new face of traditional Chinese Medicine. Science 299, 188–190 (2003)
2. Yang, W., Meng, F., Jiang, Y.: Diagnostics of Traditional Chinese Medicine. Academy Press, Beijing (1998)
3. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics 6(2), 461–464 (1978)
4. Zhang, N.L.: Hierarchical latent class models for cluster analysis. Journal of Machine Learning Research 5(6), 697–723 (2004)
5. Zhang, N.L., Kocka, T.: Efficient Learning of Hierarchical Latent Class Models. In: Proc. of the 16th IEEE International Conference on Tools with Artificial Intelligence, Boca Raton, Florida (2004)

Interpreting Gene Expression Data by Searching for Enriched Gene Sets

Igor Trajkovski and Nada Lavrač

Department of Knowledge Technologies, Jožef Stefan Institute,
Jamova 39, Ljubljana, Slovenia
{igor.trajkovski,nada.lavrac}@ijs.si

Abstract. This paper presents a novel method integrating gene-gene interaction information and Gene Ontology (GO) for the construction of new gene sets that are potentially enriched. Enrichment of a gene set is determined by Gene Set Enrichment Analysis. The experimental results show that the introduced method improves over existing methods, i.e. that it is capable to find new descriptions of the biology governing the experiments, not detectable by the traditional methods of evaluating the enrichment of predefined gene sets, defined by a single GO term.

1 Introduction

In a typical experiment, mRNA expression profiles are generated for thousands of genes from a collection of samples belonging to one of two classes - for example, tumor vs. normal tissue. The genes can be ordered in a ranked list L, according to the difference of expression between the classes. The challenge is to extract the meaning from this list. A common approach involves focusing on a handful of genes at the top of L to extract the underlying biology responsible for the phenotypic differences. This approach has two major limitations:

- After correcting for multiple hypotheses testing, no individual gene may meet the threshold for statistical significance, because the relevant biological differences are small relative to the noise, or the opposite situation, one may be left with a long list of statistically significant genes without any common biological function.
- Single-gene analysis may miss important effects on pathways. An increase of 20% in all genes encoding members of a biological process may dramatically alter the execution of that process than a 10-fold increase in a single gene.

To overcome these analytical challenges, a recently developed method, called Gene Set Enrichment Analysis (GSEA) [1], can evaluate microarray data at the level of gene sets. Biologically defined sets, for example genes that have the same biological function, are good examples of such gene sets. The most popular choice for gene sets are genes annotated with some term from GO [2].

The goal of GSEA is to determine whether members of a gene set S tend to occur toward the top of the list L, in which case the gene set is correlated with the phenotypic class distinction.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 144–148, 2007.

In this work we propose a method for generating new gene sets that have relevant biological interpretations, by combining the existing gene sets, and by inclusion of gene-gene interaction information available from the public gene annotation databases. The experimental results show that our method can find descriptions of interesting enriched gene sets, that traditional methods are unable to discover. We applied the proposed method to three gene expression data sets and we find that significant number of discovered gene sets have description which highlights the underlying biology that is responsible for distinguishing one class from the other classes.

2 Generation of New Gene Sets

In the last years several methods that test for enrichment of GO terms have been proposed. A comparative study of these methods was presented by [3]. [4] presented two novel algorithms that improve GO term scoring using the underlying GO graph topology.

None of the papers includes the gene interaction information, and none of them presents a method for the construction of novel gene sets, but rather they just calculate the enrichment of an a-priory given list of gene sets.

First, let us mention some properties of the gene annotations with GO terms:

- one gene can be annotated with several GO terms,
- a GO term may have thousands of genes annotated to it,
- if a gene is annotated with a GO term A then it is annotated with all ancestors of A.

From this information, we can conclude that each GO term defines one gene set, that one gene can be member of several gene sets, and that some gene sets are subsets of other gene sets.

Second, let $Func$, $Proc$ and $Comp$ denote the sets of gene sets that are defined by the GO terms that are a subterm of the term “molecular function”, “biological process” and “cellular component”, respectively.

Our method relies on two ideas, that are used in the construction of new gene sets:

- **Inclusion of gene interaction information.** There are cases when some abrupted processes are not detectable by the enrichment score, one reason can be that the genes had a slight increase/decrease in their expression, but had a much larger effect on the interacting genes. Therefore we think that it is reasonable to construct a gene set whose members interact with another gene set. Formally: if $G_1 \in Func$ (or $Proc, Comp$, respectively), then $G_2 = \{g_2 | g_2$ is a gene, and g_2 interacts with $g_1 \in G_1$ $\}$ was added to $Func$ (or $Proc, Comp$).
- **Intersection of gene sets.** There are cases where two or three given gene sets are not significantly enriched, but their intersection is significantly enriched. Formally: if $G_1 \in Func$, $G_2 \in Proc$ and $G_3 \in Comp$, then $G_4 = G_1 \bigcap G_2 \bigcap G_3$ is a new defined gene set.

It can happen that a gene set defined by the molecular function F is not enriched, because a lot of genes in different parts of the cell execute it and one can not expect that all of them will be over/under expressed, but if genes with that function in some specific part of the cell C_{part} are abnormally active, then it can be elegantly captured by the following gene set:

$$\text{function(F)} \bigcap \text{component}(C_{part}).$$

The newly defined gene sets are interpreted very intuitively. For example, the gene set defined as intersection of a "functional" term A and "process" term B:

$$\text{Func(A), Proc(B)} \equiv \text{function(A)} \bigcap \text{process(B)}$$

is interpreted as: *Genes that are part of the process B and have function A.*

The number of the newly defined gene sets is huge. In December 2006, $|Func| = 7513$, $|Proc| = 12549$ and $|Comp| = 1846$. Then the number of newly generated gene sets is:

$$|Func| \times |Proc| \times |Comp| \approx 1.4 \times 10^{12}$$

For each of these sets we need to compute its enrichment score, ES, that takes linear time in the number of genes ($\approx 2 \times 10^4$), we get $\approx 3 \times 10^{16}$ floating operations. If we want to statistically validate founded enriched gene sets, usually with 1000 permutation tests, we get $\approx 10^{20}$ operations, that is well above the average performance of today PC's. Therefore we need to efficiently search the space of newly generated gene sets for possible enriched gene sets.

The first idea for improvement is that we are not interested in generating all possible gene sets, but only those that are potentially enriched, and have some minimum number of genes at the top of the list, for example 5 in the first 100. That is a weak constraint concerning the biological interpretation of the results, because we are not really interested in the gene sets that do not have this number of genes at the top of the list, but it is a hard constraint concerning the pruning of the search space of all gene sets. By having this constraint we can use the GO topology to efficiently generate all gene sets that satisfy it. GO is a directed acyclic graph, the root of the graph is the most general term, which means that if one term (gene set) does not satisfy our constraint, than all its descendants will also not satisfy it, because they cover a subset of the genes covered by the given term. In this way we can significantly prune the search space of possible enriched gene sets. Therefore, we first try to construct gene sets from the top nodes of the GO, and if we fail we do not refine the last added term that did not satisfy our constraint.

3 Experiments and Conclusion

We applied the proposed methodology to three classification problems: leukemia [5], diffuse large B-cell lymphoma (DLBCL) [6] and prostate tumor [7]. The data for these three data sets were produced from Affymetrix gene chips and are available at http://www.genome.wi.mit.edu/cancer/. Gene annotations and interaction data was downloaded from Entrez database ftp://ftp.ncbi.nlm.nih.gov

/gene/. Note that this paper does not address the problem of discriminating between the classes. Instead, for the given target class we aim at finding relevant enriched gene sets that can capture the biology characteristic for that class.

To illustrate the straightforward interpretability of the enriched gene sets found by our approach, we provide the best-scoring gene sets for some of the target classes in the mentioned three classification problems (see Table 1). The statistical validation of the discovered gene sets was done by class labeled permutation testing. Table 2 list the most enriched gene sets defined by a single GO term, for the leukemia dataset. We can see that ES of the single GO terms is much smaller then the ES of the newly constructed gene sets, and most importantly, the found gene sets are constructed from not enriched GO terms. Similar results we got for the other two datasets.

The experimental results show that the introduced method improves over existing methods, and we base our conclusion on the following facts:

- Newly constructed sets have higher ES then ES of any single GO terms.
- Newly constructed sets are composed of non-enriched GO terms, which means that we are extracting additional biological knowledge that can not be found by single GO term GSEA.

Table 1. Top most enriched gene sets found in the leukemia, DLBCL and prostate dataset having p-value ≤ 0.001

Gene Set	ES
Enriched in ALL	
1. int(Func(’zinc ion binding’), Comp(’chromosomal part’), Proc(’interphase of mitotic cell cycle’))	0.60
2. Proc(’DNA metabolism’)	0.59
3. int(Func(’ATP binding’), Comp(’chromosomal part’), Proc(’DNA replication’))	0.55
Enriched in AML	
1. int(Func(’metal ion binding’), Comp(’cell surface’), Proc(’response to pest,pathogen,parasite’))	0.54
2. int(Comp(’lysosome’))	0.53
3. Proc(’inflammatory response’)	0.51
4. int(Proc(’inflammatory response’), Comp(’cell surface’))	0.51
Enriched in DLBCL	
1. int(Func(’exonuclease activity’), Comp(’nucleus’))	0.62
2. int(Func(’DNA binding’), Comp(’nucleus’), Proc(’regulation of DNA replication’))	0.61
3. Proc(’DNA replication’)	0.59
4. int(Comp(’chromosomal part’), Proc(’phosphoinositide-mediated signaling’))	0.56
Enriched in FL	
1. Comp(’integral to plasma membrane’), Proc(’cell surface receptor linked signal transd.’)	0.31
2. Comp(’integral to membrane’), Proc(’G-protein coupled receptor protein signal. pathway’)	0.28
Enriched in Tumor tissue	
1. Func(’structural constituent of ribosome’), Proc(’protein biosynthesis’))	0.74
2. Proc(’protein biosynthesis’), Comp(’cytoplasmic part’)	0.70
3. int(Func(’transit. metal ion bind.’), Proc(’protein fold.’), Comp(’intra. memb.bound organelle’))	0.62
Enriched in Normal tissue	
1. int(Func(’receptor binding’), Proc(’regulat. of mitosis’))	0.38
2. int(Func(’heparin binding’), Proc(’cell adhesion’))	0.36
3. int(Func(’growth factor activity’), Proc(’development’), Comp(’membrane’))	0.35

Table 2. Enriched gene sets in leukemia dataset (using single GO terms)

Class	Gene Set	ES
ALL	1. Proc('DNA metabolism')	0.59
	2. Comp('intracellular non-membrane-bound organelle')	0.35
	3. Proc('development')	0.22
	4. Comp('cytoplasmic part')	0.22
AML	1. Proc('inflammatory response')	0.51
	2. Proc('response to chemical stimulus')	0.41
	3. Proc('proteolysis')	0.38
	4. Proc('cell communication')	0.33

We believe that the strength of the proposed method will be even bigger through the expected increase in both the quality and quantity of gene annotations and gene-gene interaction information in the near future.

Acknowledgment

This research was supported by the Slovenian Ministry of Higher Education, Science and Technology. We are grateful to Filip Železný for previous joint work on the use of GO and gene interaction information in the analysis of microarray data which has largely stimulated this work. We are thankful to the reviewers for their constructive and critical comments and also for helpful suggestions.

References

1. Subramanian, A., et al.: Gene set enrichment analysis: A knowledgebased approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. of the U.S.A. 102(43), 15545–15550 (2005)
2. Ashburner, M., et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25(1), 25–29 (2000)
3. Khatri, P., Draghici, S.: Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 21(18), 3587–3595 (2005)
4. Alexa, A., et al.: Improved Scoring of Functional Groups from Gene Expression Data by Decorrelating GO Graph Structure. Bioinformatics 22(13), 1600–1607 (2006)
5. Golub, T.R., et al.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 5439, 531–537 (1999)
6. Shipp, M.A., Ross, K.N., Tamayo, P., et al.: Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine 8, 68–74 (2002)
7. Singh, D., Febbo, P.G., Ross, K., et al.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209 (2002)

Variable Selection for Optimal Decision Making

Lacey Gunter[1,2], Ji Zhu[1], and Susan Murphy[1,2]

[1] Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA
[2] Institute for Social Research, University of Michigan, Ann Arbor, MI 48109, USA

Abstract. This paper discusses variable selection for medical decision making; in particular decisions regarding which treatment to provide a patient. Current variable selection methods were designed for use in prediction applications. These techniques often leave behind small but important interaction variables that are critical when the goal is decision making rather than prediction. This paper presents a new method designed to find variables that aid in decision making and demonstrates the method on data from a clinical trial for treatment of depression.

1 Variable Selection for Decision Making

We consider variable selection in the simplest decision making setting in which one must decide between two actions (usually treatments). Prior to taking an action, we obtain observations about a subject $X = (X_1, X_2, ..., X_p)$, and using this information we choose an action A. We then receive a response, R, an unknown random function based on the action taken and the observations and patient outcomes subsequent to the action. The response, R, gives us some indication of the desirability of the chosen action. A policy, π, is a decision rule mapping the space of observations, X, into the space of the actions, A. The goal is to find a policy π^*, which maximizes the response. Alternate policies are compared via the expected mean response denoted by $V_\pi = E_\pi[R]$. V_π is called the (average) Value for the policy π [1]. The optimal policy, π^*, is defined as

$$\pi^* = \arg\max_{\pi} V_\pi = \arg\max_{\pi} E_\pi \left[R\right].$$

A simple example is a clinical trial to test two alternative treatments. The baseline covariates are the observations, the assigned treatment is the action and the response could be the patient's condition post treatment. The goal is to discover which treatment is optimal for any given patient, using the trial data.

Variable selection is often needed in decision making applications. Currently, variable selection for decision making in artificial intelligence is predominantly guided by expert opinion. In clinical trials, the combination of predictive variable selection methods and statistical testing of a few interaction variables suggested by expert opinion are most commonly used [3].

When selecting variables for decisions making, it is useful to distinguish between variables included merely to facilitate estimation as opposed to variables included in the decision rules. *Predictive* variables are variables used to reduce the

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 149–154, 2007.

variability and increase the accuracy of the estimator. Variables that prescribe the optimal action for a given patient are *prescriptive* variables. For optimal estimation, it is best to select both types of variables. However, only prescriptive variables must be collected when implementing the policy.

For a variable to be *prescriptive*, it must have a qualitative interaction with the action [3]. A variable X_i is said to qualitatively interact with the action, A, if there exists at least two disjoint, non empty sets $S_1, S_2 \subset \mathit{space}(X_j)$ for which

$$\arg\max_a E[R|X_j = x_{1j}, A = a] \neq \arg\max_a E[R|X_j = x_{2j}, A = a],$$

for all $x_{1j} \in S_1$, and $x_{2j} \in S_2$. These variables are useful because they help decipher which action is optimal for each individual patient.

To illustrate this idea, see the plots in Figure 1. Figure 1(a), shows a variable, X_1, that does not interact with the action, A. Figure 1(b) shows a non-qualitative interaction between a variable, X_2, and A. In both plots the optimal action is always $A = 1$. Knowing the value of X_1 or X_2 is useful for predicting R for a given action, but will not effect which action should be chosen. Figure 1(c), shows a qualitative interaction between a variable, X_3, and A. In this plot, the optimal action is $A = 0$, for $X_3 \leq .5$ and $A = 1$ for $X_3 > .5$. Knowing X_3 will impact the choice of action, thus it is useful for decision making.

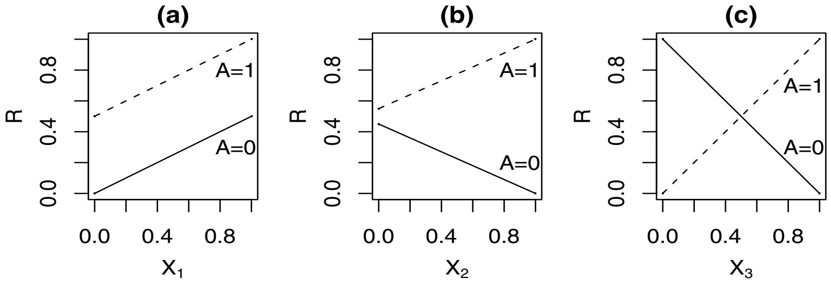


Fig. 1. Plots illustrating interactions. Plot (a) shows no interaction, (b) shows a non-qualitative interaction and (c) shows a qualitative interaction.

The degree to which a *prescriptive* variable is useful depends on two factors:

1. *Interaction*: the magnitude of the interaction between the variable and the action. For an action with two possible values, $A \in \{0, 1\}$, this is the degree to which $E[R|X = x, A = 1] - E[R|X = x, A = 0]$ varies as x varies
2. *Proportion*: the proportion of patients whose optimal choice of action changes given a knowledge of the variable. If $a^* = \arg\max_a E[R|A = a]$, this is the proportion of patients for which $\arg\max_a E[R|X = x, A = a] \neq a^*$

Consider the plots in Figure 2. Figure 2(a) shows the relationship between R, A, and a variable X_4, with an underlying plot of the distribution of X_4. Figures 2(b) and 2(c) are similar to 2(a), but for variables X_5 and X_6. Notice that X_4 and X_5 have the same distribution. However, the interaction between X_4 and A is stronger than the interaction between X_5 and A. So the effect of choosing the optimal action is much greater given X_4 than given X_5. Now notice that X_4 and X_6 have the same relationship with R and A, but are distributed differently. The distribution of X_4 is centered at the intersection, so half of the patients would do better choosing $A = 0$ over $A = 1$. Whereas, the proportion of patients benefiting from choosing $A = 0$ is much smaller with X_6.

In the next section we present a new method that ranks the variables in X based on their potential for a qualitative interaction with A. The method is based upon the *interaction* and *proportion* factors for qualitative interactions.

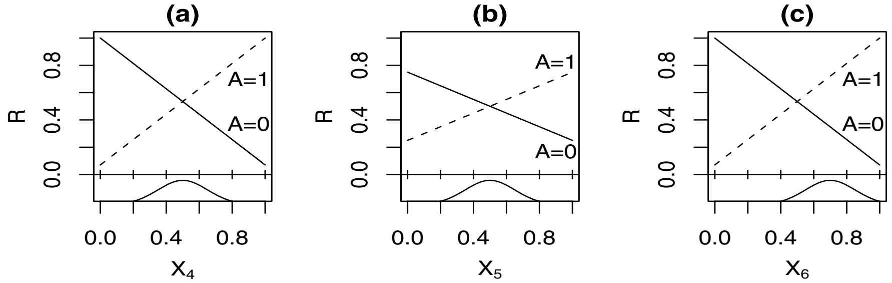


Fig. 2. Plots of qualitative interaction usefulness factors. Plot (a) shows large interaction and proportion, (b) shows smaller interaction, (c) shows smaller proportion.

2 Variable Ranking for Qualitative Interactions

Assume we have a data set of $i = 1, ..., n$ patients, with $j = 1, ..., p$ baseline observations for each patients. Also assume we have an action, $A \in \{0, 1\}$, and a response, R, for each patient. Given an estimator of $E\left[R|X_j = x_{ij}, A = a\right]$ say $\hat{E}\left[R|X_j = x_{ij}, A = a\right]$, define the following two quantities for $j = 1, ..., p$:

$$D_j = \max_{1\le i\le n} D_{ij} - \min_{1\le i\le n} D_{ij}, \qquad \text{and}$$

$$P_j = p_{1j}\left(1 - p_{1j}\right), \text{ with } p_{1j} = \frac{1}{n}\sum_{i=1}^{n} I\{D_{ij} > 0\},$$

where $D_{ij} = \hat{E}\left[R|X_j = x_{ij}, A = 1\right] - \hat{E}\left[R|X_j = x_{ij}, A = 0\right]$. D_j is a measure of the magnitude of the interaction. P_j is a measure of the proportion of patients affected by a change in the optimal choice of action due to the inclusion of X_j.

These two quantities can be combined to make a score:

$$U_j = \left(\frac{D_j - \min_{1\le k\le p} D_k}{\max_{1\le k\le p} D_k - \min_{1\le k\le p} D_k}\right)\left(\frac{P_j - \min_{1\le k\le p} P_k}{\max_{1\le k\le p} P_k - \min_{1\le k\le p} P_k}\right).$$

The score U_j can be used to rank the variables in X. It has been defined generally to allow for different models of the relationship between R, X, and A. In the section that follows, we used a linear model for the estimator $\hat{E}$.

Since variable ranking is more of a first pass method and not a final variable selection method, we suggest a full algorithm for variable selection in [4]. The algorithm is briefly summarized below.

1. Select important predictors of R in X using any predictive variable selection method
   (a) Use cross-validation to choose the tuning parameter value that gives the best predictive model
2. Rank the variables in X using U_j and select the top k in rank
   (a) Use the important predictors of R selected in step 1 to decrease the variance in the estimator $\hat{E}$
3. Use any predictive variable selection method to select from the important predictors of R selected in step 1, A, and the k interactions chosen in step 2
   (a) Use cross-validation to choose the tuning parameter value that yields a policy with the highest estimated Value

In the next section we reference this algorithm as New Method U. We used Lasso [2] for our predictive variable selection method. We tested the performance of this new method on a wide range of realistically simulated data. We compared the new method against a single Lasso fit on X, A and all interactions between X and A with the penalty parameter chosen by cross-validation on the prediction error. We reference this comparison method as 'Standard Lasso'. The new method appeared to find qualitative interactions better than the standard Lasso and selected subsets of variables that produced policies with higher average Values. Please see [4] for detailed simulation design and results.

3 Nefazodone CBASP Trial

To demonstrate the use of this method, we applied it to data from a depression study to determine which variables might help *prescribe* the optimal depression treatment for each patient. We also tried the standard Lasso for comparison.

The Nefazodone CBASP trial was conducted to compare the efficacy of three alternate treatments for chronic depression. The study randomized 681 patients with chronic major depressive disorder (MDD) to either Nefazodone, cognitive behavioral-analysis system of psychotherapy (CBASP) or the combination of the two. For detailed study design and primary analysis see [5]. We considered $p = 64$ baseline variables detailed in [4]; these variables made up the X matrix. The outcome, R, was the last observed 24-item Hamilton Rating Scale for Depression score [6], post treatment. For lack of space we only compared Nefazodone against the combination treatment. We had responses from 440 patients in this subset.

We used bootstrap [2] to reduce the variance obtained from the different cross-validation splits as follows. We took 100 bootstrap samples of the data and ran the new method and standard Lasso on each sample. We recorded the interaction variables selected and the sign of their coefficients for each sample.

Figure 3 shows plots of the absolute value of the percentage of time an interaction was selected with a positive coefficient minus the percentage of time an interaction was selected with a negative coefficient under the standard Lasso and the new method U. We used this percentage adjustment to help reduce the number of spurious selections. The standard Lasso selected a large number of the variables over 20% of the time. The new method selected only 3 variables a substantial percentage of the time. These variables were 2 indicators dealing with *alcohol* and a *somatic anxiety* score. For detailed results see [4].

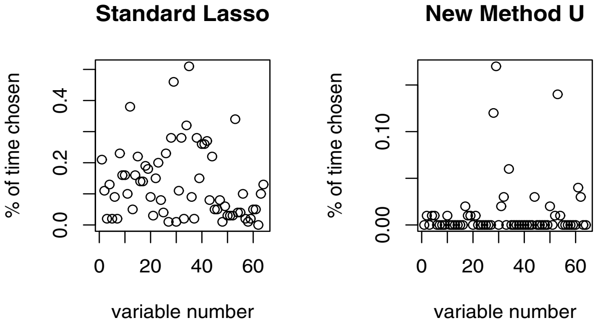


Fig. 3. Variable selection results for Nefazodone CBASP trial. In each plot, the x-axis is the variable number, the y-axis is the adjusted selection percentage.

4 Conclusion

In this paper, we discussed what makes a variable important in decision making and why current variable selection techniques are not designed to find these variables. We presented a new technique explicitly designed to select variables for decision making. The method was tested in simulations and we demonstrated it on a depression data set. The results look promising.

Acknowledgments. We wish to thank Martin Keller and the investigators of [5] for the use of their data, and gratefully acknowledge Bristol-Myers Squibb for helping fund the study. We also acknowledge financial support from NIH grants R21 DA019800, K02 DA15674, P50 DA10075, and NSF grant DMS 0505432 and technical support from A. John Rush, MD, University of Texas Southwestern.

References

1. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA (1998)
2. Hastie, T., Tibshirani, R., Friedman, J.: Elements of Statistical Learning. Springer, New York (2001)

3. Gail, M., Simon, R.: Testing for Qualitative Interactions Between Treatment Effects and Patient Subsets. Biometrics 41, 361–372 (1985)
4. Gunter, L., Murphy, S., Zhu, J.: Variable Selection for Optimal Decision Making. Technical Report 463, Department of Statistics, University of Michigan (2007)
5. Keller, M.B., McCullough, J.P., Klein, D.N., et al.: A Comparison of Nefazodone, the Cognitive Behavioral-analysis System of Psychotherapy, and Their Combination for Treatment of Chronic Depression. N. Engl. J. Med. 342, 331–366 (2000)
6. Hamilton, M.: Development of a Rating Scale for Primary Depressive Illness. Br. J. Soc. Clin. Psychol. 6, 278–296 (1967)

Supporting Factors in Descriptive Analysis of Brain Ischaemia

Dragan Gamberger[1] and Nada Lavrač[2,3,*]

[1] Rudjer Bošković Institute, Bijenička 54,10000 Zagreb, Croatia
dragan.gamberger@irb.hr
[2] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[3] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

Abstract. This paper analyzes two different approaches to the detection of supporting factors used in descriptive induction. The first is based on the statistical comparison of the pattern properties relative to the properties of the entire negative and the entire positive example sets. The other approach uses artificially generated random examples that are added into the original training set. The methodology is illustrated in the analysis of patients suffering from brain ischaemia.

1 Introduction

The task of descriptive induction is to construct patterns or models describing data properties in a symbolic, human understandable form. This work focuses on subgroup discovery whose main component is the construction of rules describing relevant subpopulations of some target concept (class). The SD subgroup discovery algorithm [3] is an example of supervised learning algorithm constructed specially for this purpose. Some other approaches like contrast set mining [1], association rule learning, or Patient Rule Induction Method (PRIM) [2] can be used as well.

The descriptive induction task is not concluded by the construction of rules describing relevant subgroups. The problem is that the rules contain only a minimal set of *principal differences* between the detected subset of target (positive) and the control (negative) class examples. For modeling purposes other properties that characterize the detected subset of positive examples are also relevant. These properties are called *supporting factors* and they are useful for better human understanding of the *principal factors* and for the support in the decision making process [3].

This paper proposes and gives an analysis of two different approaches to the detection of supporting factors used in descriptive induction. Description

[*] This work was supported by Slovenian Ministry of Higher Education, Science and Technology project "Knowledge Technologies", Croatian Ministry of Science, Education and Sport project "Machine Learning Algorithms and Applications", and EU FP6 project "Heartfaid: A knowledge based platform of services for supporting medical-clinical management of the heart failure within the elderly population".

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 155–159, 2007.

Table 1. Five rules induced by the SD algorithm from the original training set with in total 300 patients. Listed are also sensitivity and specificity values for these rules.

Ref.	**Rule**	**Sens.**	**Spec.**
	Original training set		
rn0a	$(fibr > 4.45)$ *and* $(age > 64.00)$	41%	100%
rn0b	$(af = yes)$ *and* $(ahyp = yes)$	28%	95%
rn0c	$(str = no)$ *and* $(alcoh = yes)$	28%	95%
rn0d	$(fibr > 4.55)$	46%	97%
rn0e	$(age > 64.00)$ *and* $(sex = f)$ *and* $(fibr > 3.35)$	44%	90%

of the domain together with the complete list of available attributes can be found at http://lis.irb.hr/AIME2007paper/. Some previous results of descriptive induction in the same domain are presented in [4].

2 Statistical Approach

Table 1 presents five rules defining the most relevant subgroups that can be detected for the brain stroke patients using the SD algorithm. The rules have been induced with a low value of the generalization parameter ($g = 10$) and they have very good specificity (between 90 and 100 percents). Statistical approach to the detection of supporting factors for the first of these rules is illustrated in Table 2.

The supporting factors detection process is repeated for every attribute separately. For numerical attributes we compute their mean values while for categorical attributes we compute the relative frequency of the most frequent or medically most relevant category. The necessary condition for an attribute to be considered to form a supporting factor is that its mean value or the relative frequency must be significantly different between the target pattern and the control set. Additionally, the values for the pattern must be significantly different from those in the complete positive population. The reason is that if there is no such difference then such factor is supporting for the whole positive class and not specific for the pattern. An example is the *ecghlv* attribute in Table 2. Although we have 42% of left ventricular hypertrophy cases in the pattern compared to only 27% in the control set, this is not a good supporting factor because there are 47% such cases in the positive population. This means that we have less LV hypertrophy cases in the pattern than expected and stating it as a supporting factor would be misleading.

The statistical significance between example sets can be for numerical attributes determined using Mann-Whitney test and for categorical attributes using the chi-square test of association. A practical tutorial on using these tests, as well as other potentially applicable tests, can be found in [5] (Ch. 11a and 8, respectively). By setting cut-off values at $P < .01$ for the significance of the difference with respect to the control set and $P < .05$ for the significance with respect to the positive set, attributes glucose, smoking, and stress are detected as relevant supporting factors for the pattern.

Table 2. Illustration of the statistical approach to supporting factor detection for the pattern ($fibr > 4.45$) *and* ($age > 64.00$). Three columns after attribute name are mean values for numerical attributes or relative frequencies of the selected category for the categorical attributes. Presented are values for the control set (91 examples), for the set of patients with confirmed brain stroke (209 examples), and for the positive patients included into the pattern (85 examples). The last two columns present statistical significance of the difference for the pattern versus the negative class and for the pattern versus the positive class. In bold are attribute names selected by the expert as supporting factors for the pattern.

	Control cases	**Brain stroke**	**Pattern**	**Stat. significance**	
	(neg.)	(pos.)	(subgr. of pos.)	patt. ↔ neg.	patt. ↔ pos.
Numerical attributes					
age	59.7	69.5	75.2	$P < .001$	$P < .001$
fibr	3.5	4.6	5.6	$P < .001$	$P < .001$
bmi	27.2	27.3	26.7	...	$P < .05$
gluc	6.7	7.7	8.4	$P < .001$	$P < .05$
Categorical attributes					
sex=fem.	44%	60%	71%	$P < .001$	$P < .1$
smok=yes	62%	50%	34%	$P < .001$	$P < .02$
fhis=yes	49%	59%	58%	...	...
stres=yes	77%	54%	39%	$P < .001$	$P < .02$
af=yes	19%	38%	36%	$P < .01$	...
ecghlv=yes	27%	47%	42%	$P < .05$	...
stat=yes	14%	20%	6%	$P < .1$	$P < .01$

Table 3. Supporting factors selected by the expert for three patterns defined by rules from Table 1

	Pattern		
	rn0a	**rn0b**	**rn0c**
Principal factors	$fibr > 4.45$ $age > 64.00$	$af = yes$ $ahyp = yes$	$str = no$ $alcoh = yes$
Supporting factors	$sex = f.$ $smok = no$ $str = no$ $stat = no$ *gluc* inc. (mean 8.4)	$acoag = yes$ $aarrh = yes$ *ua* inc. (mean 360) *ecgfr* inc. (mean 90)	$fhis = no$ $asp = no$ *gluc* inc. (mean 8.0)
Expert's naming of patterns	**elderly patients**	**patients with serious cardiovascular problems**	**do-not-care patients**

The decision which statistical significance is sufficiently large can depend on the medical context. The selection of supporting factors is an excellent occasion to include expert's domain knowledge into the descriptive induction process. By allowing the domain expert to decide on accepting supporting factors, they may decide to include also attributes with lower statistical significance when it is medically interesting (e.g. sex *female* and statins *no* in Table 2) and to omit some obvious relations like presence of antiarrhytmic therapy for patients with diagnosed atrial fibrillation. Table 3 presents the result of expert's selection of supporting factors for the first three patterns of Table 1. In this process the computed differences among mean values and relative frequencies have been much more important than the actual statistical significance. Detection and recognition of supporting factors enabled expert understanding of patterns induced by the subgroup methodology. The final result are names that have been given to these patterns by the expert.

3 Induction of Coexisting Factors

Another approach to support factor construction is to force rule induction algorithms to include coexisting features into the rules. This can be done by expanding the training set with additional negative examples. These additional examples are obtained by random sampling of real attribute values from the examples of the original training set. In this way they are very similar to the original examples, except that coexisting properties among attributes are destroyed due to randomness in their generation. The only chance that rules can efficiently make the distinction between the real positive examples and the artificially generated negative examples is that most relevant coexisting features are included into the rules. This means that final rules must contain both principal conditions (principal factors) which enable the distinction between positive examples and the real negative examples and coexistence conditions that enable the distinction between the positive examples and the artificially generated negative examples. The later conditions can be interpreted as supporting factors.

Table 4 presents the induced rules obtained by the SD algorithm from the training sets including artificially generated examples. It can be noticed that induced rules are longer and more specific and that they include some relevant supporting factors already detected by the statistical approach. Good is that these supporting factors are in the form of features (attribute value pairs, like glucoses larger than 7.05), but mean attribute values and relative frequencies are now missing. That makes their extraction and interpretation by medical experts not as straightforward as in the previously described statistical approach.

The advantage of this approach is that it is actually very effective in detecting homogenous subgroups that can not be detected using the SD algorithm from the original training set. The approach has successfully detected patterns based on left ventricular hypertrophy (rule $rn300c$) coexisting with very high blood pressures, and based on family history (rule $rn100e$) coexisting with female sex in combination with non-zero fundus ocular. Detected relations can be useful for medical interpretation and decision support purposes. At http://lis.irb.hr/AIME2007paper/ the interested reader may find other, potentially relevant rules induced from larger extended training sets.

Table 4. Rules induced by the SD algorithm from the expanded training sets. The used generalization parameter is the same as used to induce rules presented in Table 1. Sensitivity and specificity values are measured on the original training set.

Ref.	**Rule**	**Sens.**	**Spec.**
Original training set + 100 random negative examples			
rn100a	$(str = no)$ *and* $(smok = no)$ *and* $(fibr > 3.55)$ *and* $(hypo = no)$	20%	99%
rn100b	$(fibr > 4.55)$ *and* $(age > 64.00)$ *and* $(gluc > 7.05)$ *and* $(ahyp = yes)$	16%	100%
rn100c	$(af = yes)$ *and* $(aarrh = yes)$ *and* $(fo > 0.50)$	20%	96%
rn100d	$(str = no)$ *and* $(asp = no)$ *and* $(aarrh = no)$ *and* $(alcoh = yes)$	16%	99%
rn100e	$(fo > 0.50)$ *and* $(sex = f)$ *and* $(alcoh = no)$ *and* $(fhis = yes)$	21%	92%
Original training set + 300 random positive examples			
rn300a	$(af = yes)$ *and* $(aarrh = yes)$ *and* $(fo > 0.50)$ *and* $(ahyp = yes)$	17%	98%
rn300b	$(str = no)$ *and* $(smok = no)$ *and* $(age > 67.00)$ *and* $(stat = no)$ *and* $(dya > 85.50)$ *and* $(hypo = no)$	15%	98%
rn300c	$(ecghlv = yes)$ *and* $(dya > 95.50)$ *and* $(sys > 172.50)$ *and* $(af = no)$ *and* $(asp = no)$	9%	99%
rn300d	$(str = no)$ *and* $(smok = no)$ *and* $(age > 67.00)$ *and* $(fibr > 4.95)$	13%	100%
rn300e	$(acoag = yes)$ *and* $(af = yes)$ *and* $(fo > 0.50)$	8%	99%

4 Conclusions

The significance of the methodology of induction from extended training sets is that it enables effective detection of homogenous subgroups. The subgroups described by rich sets of coexisting factors are intuitive for human interpretation. The approach nicely complements the subgroup discovery process and can be also easily implemented in domains with a very large number of attributes (e.g. gene expression domains). On the other hand, the statistical approach for detection of supporting factors is more systematic. Its main advantage is that it enables effective integration of the available medical knowledge into the descriptive induction process. Both approaches can be combined, resulting in a powerful tool for descriptive induction.

References

1. Bay, S.D., Pazzani, M.J.: Detecting group differences: Mining contrast sets. Data Min. Knowl. Discov. 5(3), 213–246 (2001)
2. Friedman, J.H., Fisher, N.I.: Bump-hunting for high dimensional data. Statisics and Computing 9, 123–143 (1999)
3. Gamberger, D., Lavrač, N., Krstačić, G.: Active subgroup mining: A case study in a coronary heart disease risk group detection. Artificial Intelligence in Medicine 28, 27–57 (2003)
4. Gamberger, D., Lavrač, N., Krstačić, A., Krstačić, G.: Clinical data analysis based on iterative subgroup discovery: Experiments in brain ischaemia data analysis. Applied Intelligence (in press, 2007)
5. Lowry, R.: Concepts and Applications of Inferential Statistics (2007) `http://faculty.vassar.edu/lowry/webtext.html`

Knowledge Acquisition from a Medical Corpus: Use and Return on Experiences

Lina F. Soualmia[1] and Badisse Dahamna[2]

[1] Laboratoire LIM&Bio, UFR SMBH Léonard de Vinci – Université Paris XIII
74, Rue Marcel Cachin, 93017 Bobigny Cedex, France
Lina.Soualmia@gmail.com
[2] Rouen University Hospital – 1, Rue de Germont, 76031 Rouen Cedex, France
Badisse.Dahamna@chu-rouen.fr

Abstract. The present work aims at refining and expanding user's queries thanks to association rules. We adapted the A-Close algorithm to a medical corpus indexed by MeSH descriptors. The originality of our approach lies in the use of the association rules in the information retrieval process and the exploitation of the structure of the domain knowledge to evaluate the association rules. The results show the usefulness of this query expansion approach. Based on observations, new knowledge is modelled as expert rules.

Keywords: Knowledge Discovery, Data Mining, Information Retrieval.

1 Introduction

Internet as source of health information is increasing in preeminence. Information retrieval remains problematic. In specific domains such as medicine, controlled vocabulary in terminologies can help overcome problems with synonymy and ambigity. We propose here to use a new method to refine the users' queries by interactive query expansion founded on a set of association rules extracted from electronic health documents by a data mining technique. The contribution of this paper is twofold. First, we propose to extract new knowledge in the form of association rules from health documents, i.e. previously unknown or not specified, and we give a method to evaluate the quality of the extracted association rules. Second, we propose to use these association rules in information retrieval. As a return on experiences additional knowledge is modelled by the domain expert. In fact, the observations of the association rules patterns allowed the generation of *expert rules*.

The remainder of the paper is organized as follows: in section 2 we start by giving definitions of association rules and we describe different experiences of data mining from a medical corpus with and without categorization; text-mining is also processed. We propose new criteria to evaluate the quality of association rules by using the domain knowledge in section 3. Query expansion is detailed in section 4. We give return on experiences in section 5. Finally, we present directions for further research.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 160–164, 2007.

2 Knowledge Discovery in Medical Documents

Association Rules. Association rules were initially used in data analysis and in data extraction from large relational databases [1]. A Boolean association rule AR states that if an object has the items $\{i_1,i_2\ldots,i_k\}$ it tends also to have the items $\{i_{k+1},\ldots,i_n\}$. It is expressed as: AR : $i_1 \wedge i_2 \wedge \ldots \wedge i_k \Rightarrow i_{k+1} \wedge \ldots \wedge i_{n.}$

The AR *support* represents its utility. This measure corresponds to the number of objects containing at the same time the rule antecedent and consequent. *Support(AR)*=$|\{i_1, i_2\ldots, i_n\}|$. The AR *confidence* represents its precision. This measure corresponds to the proportion of objects that contains the consequent rule among those containing the antecedent. *Confidence (AR)* = $|\{i_1, i_2\ldots, i_n\}|/|\{i_1, i_2\ldots, i_k\}|$. *Exact association rules* have a confidence = 100%, i.e. verified in all the objects of the database and *approximative association rules* have confidence < 100%.

The knowledge extraction process is realized in several steps: the data and context preparation (objects and items selection), the extraction of the frequent itemsets (compared to a minimum support threshold), the generation of the most informative rules using a data mining algorithm (compared to a minimum confidence threshold), and finally the interpretation of the results and deduction of new knowledge [6]. An extraction context is a triplet $C= (O, I, R)$ where O is the set of objects, I is the set of all the items and R is a binary relation between O and I.

Using the semantic based on the closure of the Galois connection, two new bases for association rules are defined by the algorithm A-Close [2]. These bases are generating sets for all valid non-redundant association rules composed by minimal antecedents and maximal consequents, i.e. the most relevant association rules. We adapt the algorithm to the case of a large collection of health documents. Many algorithms exist, but the A-Close association rules caracteristics' are particularily interesting in our application of query expansion [3]. The objects in O are the indexed electronic health documents. Each document has a unique identifier and a set of MeSH descriptors. These descriptors are keywords (e.g. *abdomen, hepatitis...)* and couples of keywords and qualifiers (e.g. the association *hepatitis/diagnostic* where *hepatitis* is a keyword and *diagnostic* a qualifier). The relation R represents the indexing relation. We distinguish two cases of data mining: founded on the conceptual indexing of the health documents and founded on the plain text indexing.

Data Mining from the Database. 11,373 documents are selected at random from the database. The support threshold is *minsup*=20 and the confidence threshold is *minconf*=70% for the approximative association rules. In the case 1, I is the set of keywords used to index the documents. In the case 2, I is the set of keywords and qualifiers associated to the documents. In the case 3, I is the set of couples of keywords and qualifiers related to the documents.

Table 1. Number of rules; ER: Exact rules; AR: Approximative rules

Context	ER	AR	Total
Case 1 : I={keywords}	2 438	9 381	11 819
Case 2: I={keywords}∪{qualifiers}	5 241	11 738	16 976
Case 3: I={keywords/qualifiers}	648	1 917	2 565

In all these contexts, the number of rules is high to be manually analyzed by our expert. Indeed, the number of generated association rules may be high and the results interpretation may become complex and inextricable [4]. Association rules between couples of (keyword/qualifier) are more precise than simple association rules.

Categorizing Documents. To obtain more precise rules we realize experiments on categorized documents according to medical specialities (for example: *cardiology*) to evaluate the influence of the categorization on the association rules generation. The categorization algorithm [5] is processed on the initial collection of 11,373 documents. A document may belong to several specialities. In the case 1, *I={keywords}* and in the case 2, *I ={(keywords/qualifiers)}*. The number of rules (Table 2) is nearly the same before categorization but the rules are not the same as they don't have the same support and confidence measures.

Table 2. Number of rules by speciality; ER: Exact rules; AR: Approximative rules

		Case 1			Case 2		
Speciality	**Docs**	**ER**	**AR**	**Total**	**ER**	**AR**	**Total**
Allergy	509	101	231	332	93	206	299
Cardiology	558	251	542	793	151	332	483
Oncology	644	154	329	483	119	358	477
Psychiatrics	515	76	337	413	57	155	212
Gastroenterology	501	85	300	385	96	248	344
Neurology	1 137	169	520	689	83	285	368
Environment	1 254	257	924	1 181	148	584	732
Diagnosis	883	465	1 218	1 683	112	312	424
Therapeutic	782	555	2 010	2 565	206	562	768
Pediatrics	906	1 116	5 629	6 745	205	634	839
	Total	**3 229**	**12 040**	**15 269**	**1 270**	**3 676**	**4 946**

Weighted Association Rules. In medical documents the descriptors could be *major* (alloted by a star "*") or *minor*. We generate here association rules between major descriptors. In the case 1, *I* is the set of the major keywords *I={keyword*}*. In the case 2, *I* is the set of the major keywords qualifiers couples *I={(keyword/qualifier)*}*.

Table 3. Number of rules; ER : exact rules; AR approximative rules

		Case1			Case2		
Speciality	**Docs**	**ER**	**AR**	**Total**	**ER**	**AR**	**Total**
Allergy	509	4	12	16	2	12	14
Cardiology	558	7	37	44	5	31	36
Oncology	644	2	13	15	0	20	20
Psychiatrics	515	1	8	9	0	3	3
Gastroenterology	501	4	34	38	2	12	14
Neurology	1 137	4	34	38	0	25	25
Environment	1 254	6	85	91	5	53	58
Diagnosis	883	7	36	43	4	36	40
Therapeutic	782	2	32	34	2	18	20
Pediatrics	906	6	90	96	4	61	65
	Total	**43**	**381**	**424**	**24**	**271**	**295**

Text Mining. Text mining differs in the nature of the studied data. We have realized an automatic indexing of the plain text of the documents. The number of rules is high to be evaluated by an expert.

Table 4. Number of rules; I={term}

Speciality	Documents	Exact Rules	Approximative Rules	Total
Neurology	1 137	1 354	105 202	106 556
Environment	1 254	2 815	397 073	399 888

3 Association Rules Evaluation

Unbiassed interest measures are founded on statistical results (support/confidence). Subjective measures are based on the users' knowledge. All the extracted rules of the most expressive context (Table 3.Case 2) were evaluated. an interesting association rule is one that confirms or states a new hypothesis [4]. There may be several cases according to the existing relationships between MeSH terms. It could associate a (in)direct son and its father in the hierarchy; two terms that belong to the same hierarchy (same (in)direct father); a See Also relationship that exists in the thesaurus; a new relationship judged interesting by our domain expert. The rules that describe existing relationships are automatically classified thanks to the MeSH thesaurus (n=52). It remains 242 rules to analyze by the expert. We obtained the following results: New rules: 70.78 % (42.85% of total); See Also: 13.48 % (08.16%); Same hierarchy: 09.55 % (05.78%); Father-Son: 06.18% (03.74%). Some examples of new rules: *breast cancer/diagnosis → mammography*

aids/prevention and control → condom

Turner syndrome ∧ *child → human growth hormone* ∧ *growth disorders*

4 Association Rules in Query Refinement and Expansion

As an output, the association rules may be visualized or automatically added to the database for interactive query expansion [6] (using the approximative association rules) and automatic query expansion (using the exact association rules). The association rule *"breast,cancer"→"mammography"* is a new one. Applying it on a query containning the term "breast cancer", we propose to the user documents related to "mammography" to complete its knowledge. The co-occurrence tools developed for information retrieval bring closer the terms which frequently appear in the same documents and thus which have a semantic proximity. Analogically, association rules can be exploited. It is useful in the case of nonprecise information needs: association rules are indication on the possible definition of a term or its environment. An expanded query by association rules contains more related terms. By using the vectorial model, more documents will be located increasing recall. Interactive expansion requires user implication. We developed an evaluation tool used by a set of 500 users (http://chu-rouen.fr/enquete). The results (76% of the users are satisfied) show the usefulness of this approach.

5 Return on Experiences

Indexing Correction. According to the indexing policy, the more precise descriptor should be used to index a document, i.e. the descriptor in the lower level in hierarchy. However, 1,466 documents contain descriptors that have father-son relathionships. For example, a document is indexed by the keywords *trisomy* and *chromosome aberrations*, whereas *trisomy* is a *chromosome aberrations*. 478 documents are indexed by qualifiers that have a relathionship and are associated to the same keyword. This can explain the proportion of the 29.21% of existing associations. An indexing correction is thus proposed to the indexing team.

Modelling Expert Rules. The other return on experience is based on the observations of the expert and consists in modelling and formalizing rules between couples of (keyword/qualifier): the pattern of (*hepatitis/prevention → hepatitis vaccines*) is used to model *tuberculosis/prevention → BCG vaccine*; *dysentery bacillary/prevention → shigella vaccines*. 456 rules are modelled using this pattern. Different cases are possible: $K_1/Q_1 \xrightarrow{-} K_2$ states that K_1/Q_1 should be replaced by K_2 and $K_1/Q_1 \xrightarrow{++} K_2/Q_2$ states that the couple K_2/Q_2 should be added to the couple K_1/Q_1 in retrieval and indexing.

6 Conclusion and Future Work

We apply data mining to extract interesting associations rules, some of which previously unknown, from the documents. We proposed to exploit these association rules in information rerieval process. Association rules link conceptual structures of the documents. Indeed, the descriptors are represented by concepts organized in a hierarchy on which it is possible to make generalization. We plan to exploit hierarchies in order to generate generalized association rules. A minimal support is fixed to eliminate the very rare rules. We plan to fix a maximum support threshold in order to avoid very frequent association rules and to use other statistical indices.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In: VLDB, pp. 478–499 (1994)
2. Pasquier, N., et al.: Generating a Condensed Representation of Association Rules. Intelligent information systems (2004)
3. Soualmia, L.F., Darmoni, S.J.: Combining Knowledge-based Methods to Refine and Expand Queries in Medicine. In: Christiansen, H., Hacid, M.-S., Andreasen, T., Larsen, H.L. (eds.) FQAS 2004. LNCS (LNAI), vol. 3055, pp. 243–255. Springer, Heidelberg (2004)
4. Fayyad, U.M., et al.: Advances in Knowledge Discovery and Data Mining. AAAI Press, Stanford (1996)
5. Névéol, A., et al.: Using MeSH Encapsulated Terminology and a Categorization Algorithm for Health Resources. International Journal of Medical Informatics 73(1), 57–64 (2004)
6. Magennis, M., Van Rijsbergen, C.J.: The Potential and Actual Effectiveness of Interactive Query Expansion. In: SIGIR, pp. 324–332 (1997)

Machine Learning Techniques for Decision Support in Anesthesia

Olivier Caelen[1], Gianluca Bontempi[1], and Luc Barvais[2]

[1] Machine Learning Group, Département d'Informatique,
Université Libre de Bruxelles, Bruxelles, Belgium
[2] Service d'Anesthésiologie-Réanimation, Faculté de Médecine,
Université Libre de Bruxelles, Bruxelles, Belgium

Abstract. The growing availability of measurement devices in the operating room enables the collection of a huge amount of data about the state of the patient and the doctors' practice during a surgical operation. This paper explores the possibilities of generating, from these data, decision support rules in order to support the daily anesthesia procedures. In particular, we focus on machine learning techniques to design a decision support tool. The preliminary tests in a simulation setting are promising and show the role of computational intelligence techniques in extracting useful information for anesthesiologists.

1 Introduction

Machine learning and data mining are key technologies in order to transform data into useful information for better diagnosis, event detection and decision aid. This paper deals with the anesthesia domain where several platforms have recently been made available to support the anesthesiologist in the operating room. An example is the TOOLBOX software [1] which has been used for several years by the group of anesthesiology of the ULB Erasme Hospital[1]. This software monitors the patient's state and acts as a servo-controller on the multiple intravenous drug infusions, whose setting is regularly adjusted by the anesthesiologist, by simultaneously using pharmacokinetic and pharmacodynamic principles [2] (Figure 1). Before and

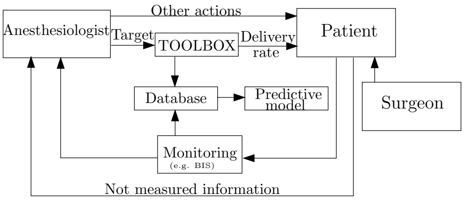


Fig. 1. The TOOLBOX software and the anesthesia procedure

[1] L'Hôpital Erasme is the university hospital of Université Libre de Bruxelles (ULB), Brussels, Belgium.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 165–169, 2007.

during the operation, TOOLBOX stores necessary statistics and monitoring information like: (i) basic details regarding the doctor, the patient and his general state, (ii) the type of surgery, (iii) the evolution of the hemodynamic and physiological parameters (e.g. the BIS) of the patient, (iv) the evolution of the drugs concentration levels chosen by the anesthesiologist. In this study, 910 surgical intervention sessions are used to build the database.

This paper discusses and assesses the role of machine learning techniques in extracting useful information from the database generated by TOOLBOX in order to test and develop a decision support tool to assist the anesthetist during his routine procedure. In particular, we will focus on the impact of the brain concentration of Propofol on the hypnosis upon monitoring with the bispectral index. The bispectral index (BIS) [3,4] is a well-known measure adopted by anesthesiologists to rate the depth of the hypnosis. The BIS index represents the electro-encephalographic signal in a normalized range from 100 to 0, where 100 stands for the "awake" status and 0 stands for electrical silence. Propofol is a short-acting intravenous hypnotic agent used for the induction and maintenance of general anesthesia. According to some information about the patient and the target Propofol brain concentration, the decision support tool will give information on the future BIS value of the patient that will help the anesthesiologist to take the best decision regarding the drugs modification. A correlation technique is used to estimate the time between the Propofol target brain concentration modification and its impact on the BIS index.

In this paper, we will assess and compare a linear model and a local learning approach called the lazy learning [5]. The learning procedure is supported by a forward feature selection procedure to reduce the input dimensionality of the prediction problem. This step is very important since there is a large number of variables (e.g. the patient age, the surgery type, the phase of the operation, etc.) which could influence the value of the BIS signal.

The main contributions of this paper are (i) the application of a system identification procedure on a huge database concerning the relation between the drug modification and the impact on the BIS index, (ii) the comparison of the accuracy between a classical linear model and a local linear model (lazy learning) (iii) the execution of a forward variable selection to extract the most relevant input variables.

2 Learning the Predictive Model

The goal of our decision support architecture is to assist the anesthetist in adjusting the concentration of the Propofol drug in order to let the BIS of the patient attain the desired level. Suppose that the dynamics of the BIS index can be described by a single-input single-output (SISO) NARMAX (Nonlinear AutoRegressive Moving Average with eXternal input) discrete-time dynamic system [6]

$$B(t + \Delta t^{*}) = f(B(t), tpo(t), tpn(t), \Delta timeP, tr(t), a, w, h, s, lbm) + \epsilon(t) \qquad (1)$$

where, at time t, $B(t)$ is the BIS value[1], $tpo(t)$ (in $\mu g/ml$) is the old concentration of Propofol, $tpn(t)$ is the new concentration of Propofol (action of the anesthetist), $\Delta timeP$ is the time between t and the previous Propofol modification and $tr(t) \in [4, 6]$ (in ng/ml) is the concentration of Remifentanil. Also, a, w, h, s, lbm are, respectively, the age, the weight, the height, the sex and the lean body mass of the patient. $\epsilon(t)$ is random noise and Δt^* is the time delay which maximizes the correlation between drug modification and BIS variation.

We apply a system identification procedure [6] to the samples collected by TOOLBOX to estimate the model

$$\widehat{B}(t + \Delta t^*) = \widehat{f}(B(t), tpo(t), tpn(t), \Delta timeP, tr(t), a, w, h, s, lbm, \alpha_N) \qquad (2)$$

where α_N is a vector containing the parameters of the model.

Let us define as query point q the vector containing all the input variables. We identify the system by using a *training set* of $N = 1702$ measures $\{B_i(t + \Delta t^*), q_i(t)\}$, $i = 1, \ldots, N$. The sample $(B_i(t + \Delta t^*), q_i(t))$ means that (i) we observed at time $t + \Delta t^*$ the BIS value $B_i(t + \Delta t^*)$, (ii) the target concentration of Propofol was set at time t and (iii) no other modification of the Propofol target occurred during the interval $[t, t + \Delta t^*]$.

The simplest learning approach boils down to a conventional linear identification [7]. However, when linear identification does not return a sufficiently accurate prediction, the designer may want to use alternative methods for learning non-linear relationships. This paper adopts a method of local modeling, called lazy learning, which proved to be successful in many problems of non-linear modeling [8] and in two international competitions on data analysis and time series prediction [9].

The learning procedure is preceded by a feature selection step in order to reduce the dimensionality of the problem. We use a *sequential forward selection* [10] where a leave-one-out cross-validation procedure is used to assess the accuracy of the input sets. This procedure is useful both for statistical reasons and to return to the anesthesiologist high-level information about which variables play a role on the evolution of the patient physiological parameters.

3 Results

This section summarizes the results of the different BIS predictors assessed during the forward selection procedure. Three leave-one-out criteria are used to assess the accuracy of the predictive models. Let $\widehat{E}_i^{loo} = \widehat{B}_{(-i)}(t+\Delta t^*) - B_i(t+\Delta t^*)$ be the leave-one-out error made on the sample i where $\widehat{B}_{(-i)}(t + \Delta t^*)$ is the prediction for the sample i returned by a model trained on all the samples except i. The first criterion is the *normalized mean squared error* $NMSE = \frac{\sum_{i=1}^{N}\left(\widehat{E}_i^{loo}\right)^2}{\sum_{i=1}^{N}(\widehat{\mu}_b - B_i(t+\Delta t^*))^2}$ where $\widehat{\mu}_b = 1/N \sum_{i=1}^{N} B_i(t+\Delta t^*)$ is the average of the future

[1] In order to smooth the fluctuations, $B\,(t)$ is the time average of the BIS over the interval $[t - 30, t]$ and $B\,(t + \Delta t)$ is the time average over the interval $(t + \Delta t - 30,\ t + \Delta t + 30)$.

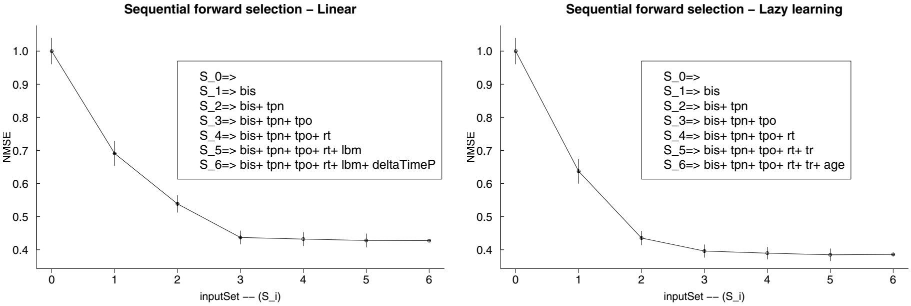


Fig. 2. Results of the sequential forward selection process for a linear model (left) and a lazy learning predictor (right)

BIS index. This quantity is greater than zero and normalizes the performance of the predictor with respect to the variance of the signal to be predicted. A value of NMSE= 1 means that we are simply predicting the average of the BIS series. The second criterion is the *mean of the absolute errors* $MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{E}_i^{loo}\right|$ which returns an indication of the average magnitude of the errors made by the decision support system. The last criterion returns the percentage (P) of times that the variation of the BIS is wrongly predicted.

$$P = 100\left(1 - \frac{\sum_{i=1}^{N} I\left[\left(\widehat{B}_{(-i)}(t+\Delta t^*) - B_i(t)\right)\cdot\left(B_i(t+\Delta t^*) - B_i(t)\right)\right]}{N}\right) \tag{3}$$

where $I[A] = \begin{cases} 1 & \text{if } A \geq 0, \\ 0 & \text{if } A < 0. \end{cases}$.

In both cases (linear and lazy) the forward selection procedure confirms the importance of taking into account the value of the current BIS index as well as the new and the old targets of Propofol. Note that most of the other variables (age, sex, ...) are integrated in the pharmacokinetic model used by TOOLBOX and this could explain the fact that these variables are discarded by the selection procedure.

The last experiment compares a linear predictor $\Lambda_{S_3}^{lin}$ and a lazy predictor $\Lambda_{S_3}^{lazy}$ both taking as inputs the current BIS value, the previous and the current Propofol target. Table 1 reports the three criteria accuracy figures for the

Table 1. Three measures of BIS prediction error for the linear ($\Lambda_{S_3}^{lin}$) and the lazy ($\Lambda_{S_3}^{lazy}$) model

model	$NMSE$	MAE	P
$\Lambda_{S_3}^{lin}$	0.437	6.35	18.2
$\Lambda_{S_3}^{lazy}$	0.395	5.91	16.9

linear model $\Lambda_{S_3}^{lin}$ and the lazy predictor $\Lambda_{S_3}^{lazy}$. According to a paired-t test all the differences are significant. This means that the lazy predictor significantly outperforms the linear one and suggest the existence of a nonlinear relationship linking the target of Propofol and the BIS signal.

4 Conclusion and Future Work

This paper compares conventional linear and machine learning prediction techniques in a predictive modeling task. The encouraging results show that predictive models can extract useful information from historical data and provide support to the decisions of anesthetists during surgical operations. Future work will focus on the implementation of a prototype to be tested, in real conditions, during daily operations in the operating room.

Acknowledgement. The work was funded by the FIRST Europe project number EP1A320501R059F/415717 of the Region Wallonne, Belgium. The authors thank their colleagues A. A. Miranda and G. Bejjani for their valuable comments.

References

1. Cantraine, F., Coussaert, E.: The first object oriented monitor for intravenous anesthesia. Journal of Clinical Monitoring and Computing 16(1), 3–10 (2000)
2. Bailey, J.M., Haddad, W.M., Hayakawa, T.: Closed-loop control in clinical pharmacology: Paradigms, benefits and challenges. In: Proceedings of the 2004 American Control Conference, pp. 2268–2277 (2004)
3. Sigl, J.C., Chamoun, N.G.: An introduction to bispectral analysis for the electroencephalogram. Clin Monitor 10, 392–404 (1994)
4. Gentilini, A., Frei, C.W., Glattfedler, A.H., Morari, M., Sieber, T.J., Wymann, R., Schnider, T.W., Zbinden, A.M.: Multitasked closed-loop control in anesthesia. IEEE Engineering in Medicine and Biology 39–53 (2001)
5. Bontempi, G., Birattari, M., Bersini, H.: Lazy learning for modeling and control design. International Journal of Control 72(7/8), 643–658 (1999)
6. Söderström, T., Stoica, P.: System Identification. Prentice-Hall, Englewood Cliffs (1989)
7. Myers, R.H.: Classical and Modern Regression with Applications. PWS-KENT, Boston, MA (1990)
8. Bontempi, G.: Local Learning Techniques for Modeling, Prediction and Control. PhD thesis, IRIDIA- Université Libre de Bruxelles (1999)
9. Bontempi, G., Birattari, M., Bersini, H.: Lazy learners at work: the lazy learning toolbox. In: Proceeding of the 7th European Congress on Intelligent Techniques and Soft Computing EUFIT '99 (1999)
10. Aha, D.W., Bankert, R.L.: A comparative evaluation of sequential feature selection algorithms. Artificial Intelligence and Statistics 5 (1996)

Learning Decision Tree for Selecting QRS Detectors for Cardiac Monitoring

François Portet[1], René Quiniou[2], Marie-Odile Cordier[2], and Guy Carrault[3]

[1] Department of Computing Science, University of Aberdeen, Aberdeen, AB24 3UE, UK
fportet@csd.abdn.ac.uk
[2] Irisa, INRIA, Université de rennes 1, Campus de Beaulieu, 35042, Rennes, France
{quiniou,cordier}@irisa.fr
[3] LTSI, Université de rennes 1, Campus de Beaulieu, 35042, Rennes, France
guy.carrault@univ-rennes1.fr

Abstract. The QRS complex is the main wave of the ECG. It is widely used for diagnosing many cardiac diseases. Automatic QRS detection is an essential task of cardiac monitoring and many detection algorithms have been proposed in the literature. Although most of the algorithms perform satisfactorily in normal situations, there are contexts, in the presence of noise or a specific pathology, where one algorithm performs better than the others. We propose a combination method that selects, on line, the detector that is the most adapted to the current context. The selection is done by a decision tree that has been learnt from the performance measures of 7 algorithms in various instances of 130 combinations of arrhythmias and noises. The decision tree is compared to expert rules tested in the framework of the cardiac monitoring system IP-CALICOT.

1 Introduction

The QRS complex is the main wave in the ECG as it reflects the ventricular activity of the heart. Its automatic detection is an essential task for cardiac monitoring systems that has been studied for several decades and has resulted in a large number of methods [1-5]. But, each method has situations where it fails as each QRS detector reacts differently to the large number of different QRS waveforms and noises.

In this paper, we propose to combine the strength of several algorithms to detect the QRS complex even in difficult situations. The approach is not to fuse the detectors outputs but to select, on-line, the best detector from among a set of algorithms according to an evaluation of the current context of the chunk of ECG to process. The selection is done by decision tree (DT) which is learnt to select 7 QRS detectors according to various situations, called *contexts*, representing 130 combinations of arrhythmia and clinical noise. The learning is detailed in section 2 and its results is analysed in section 3. This method is then compared with expert rules previously acquired in section 4. Finally, the paper ends with a short discussion.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 170–174, 2007.

2 Learning Method

Selecting the QRS detector the most suited to the specific context of some ECG chunk is a difficult task. We advocate the use of selection rules but these rules must be acquired. In a previous experiment [6], we described an expert acquisition method. This experiment emphasized the complexity of the task. That is why an automatic approach is considered. Many methods could be used to learn selection rules but decision tree learning presents several advantages in our application: (1) the ECG contexts are composed of both nominal and categorical data that are easily handled by the learning method; (2) rules derived from decision tree are explicit and checkable by human experts; and (3) the learned decision tree can processes large volume of data in a short time, which is mandatory in ICU monitoring.

Succinctly, a decision tree (DT) consists of several test nodes and class (or decision) leaves. It classifies an input by executing the tests in the tree beginning at the root and going down the tree until a leaf is reached which gives the class of the input (or the decision to be taken). The C4.5 algorithm of Quinlan [7] has been used to learn the tree. In our application, the DT input is an ECG context described by a set of attributes (i.e. properties of an ECG chunk) and the output is the algorithm to apply to the input context (i.e. the decision). To learn the DT, a training data set has been created. This has been achieved by (1) generating all possible contexts that can be found in an ECG, (2) applying all the QRS detectors to these contexts and (3) deciding what the best detector to apply to a given context is.

2.1 Definition of the Context

A context is defined as the combination of a *rhythm context* and a *noise context*. Indeed, in clinical practice, an ECG is composed of the original ECG —the rhythm context— which is usually corrupted by noise —the noise context.

An ECG is composed of different QRS waveforms and the variation of waveforms inside an ECG signal disturbs the detection. To the best of our knowledge, the influence of QRS waveform variation on the QRS detection has been studied only in very few papers [6]. The rhythm contexts have been extracted from the MIT-BIH Arrhythmia Database [8]. 10 rhythms that are representative of normal rhythms and arrhythmic situations have been chosen to assess the detectors on sequences of identical QRSs as well as on sequences of non identical QRSs.

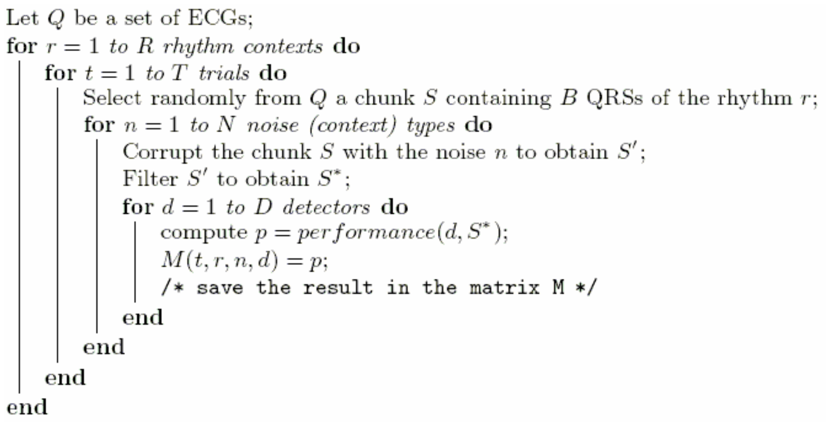
The noise used to corrupt the ECG comes from several sources. Few studies have analyzed the influence of noise on QRS detectors. Most of them used composite noise that is not representative of real clinical situation. In our study, clinical additive noise was extracted from the MIT-BIH Noise Stress Test Database[9], which contains three noise records, lasting 30 mins each, predominantly composed of baseline wander (*bw*), muscle artifact (*ma*), and electrode motion artifact (*em*). We used these three types of noise at four Signal-to-Noise Ratios (SNR). Thus the attributes of a context are: rhythm context type, noise context type, and the SNR of the noise context.

2.2 Selected QRS Detectors

Many QRS detection schemes have been described in the literature for the last 30 years; however, since the 90s the performance has improved slowly in non noisy situations. For example, the Pan and Tompkins detector (1985) [1] (ER=0.68%) performs slightly less than the Christov's one (2004) [5] (ER=0.44%) in uncorrupted situations. That is why selection rules are learnt in noisy situations in order to emphasize the difference between detectors. Seven algorithms were selected. They were those used in [6] (*pan, gritzali, af*2, and *df*2) plus those proposed by Benitez *et al.*[4] (*benitez*), Suppappola and Sun[2] (*mobd*), and Kadambe *et al.*[3] (*kadambe*).

2.3 Computation of the Training Set for Decision Tree Learning

The decision rules for selecting some algorithm in a specific context are given by a decision tree previously learned. The training set was computed by the algorithm given Fig. 1.



```
Let $Q$ be a set of ECGs;
for $r = 1$ to $R$ rhythm contexts do
    for $t = 1$ to $T$ trials do
        Select randomly from $Q$ a chunk $S$ containing $B$ QRSs of the rhythm $r$;
        for $n = 1$ to $N$ noise (context) types do
            Corrupt the chunk $S$ with the noise $n$ to obtain $S'$;
            Filter $S'$ to obtain $S^*$;
            for $d = 1$ to $D$ detectors do
                compute $p = performance(d, S^*)$;
                $M(t, r, n, d) = p$;
                /* save the result in the matrix M */
            end
        end
    end
end
```

Fig. 1. Algorithm used for the computation of the results

This algorithm computes a matrix **M** where the performance ***p*** of each QRS detector is related to the context and the trial. ***p*** is composed of 3 values: True Positives (TP -- correct detections), False Negatives (FN -- missed detections) and False Positives (FP -- false alarms). These values are then used to compute: the Error Ratio, ER=(FN+FP)/(TP+FN), the Sensitivity, Se=TP/(TP+FN), the Positive Predictivity, PP=TP/(TP+FP) and the F-Measure, FM=2*PP*Se/(PP+Se).

The training set was composed using the following parameters: chunk length **B=10** beats (QRSs), **N=13** noise values, **R=10** rhythm types, **T=100** instances of each context type, **D=7** detectors. In all, 9,500 QRSs were used representing 8.70% of the whole database and leading to a training set composed of T*N*R (100*13*10) =13,000 individuals. The class (i.e. algorithm that should be chosen in this context) of each individual was found by selecting the QRS detector with the best FM value, for each instance of context.

3 Learning Results

A tree of size 69 for 48 leaves was obtained from the training set. *benitez* was used in a majority of the contexts, particularly in the contexts *no_noise* and *bw*. These two contexts do not perturb the signal as *bw* can be easily removed by the input filter of the detectors. Then, for the *ma* noise it alternates between *kadambe* and *benitez*. The choice of *kadambe* comes from its wavelet filtering, which appears to be the most able to deal with high-frequency noise. For the *em* context, the choice alternates between all the detectors of the set except *kadambe*. The *em* noise is composed of high and low frequency components which can affect the detectors very much. Moreover, in this branch, the rhythm type rather than the level of noise is used to distinguish the different cases. This demonstrates the value of using the rhythm information as a factor which influences the detectors performance.

4 Experiments and Comparison with Expert Rules

A test set of 11 uncorrupted ECG records including 10 different rhythm contexts and representing 5 hours and 30 mins, was extracted from the MIT-BIH Arrhythmia database. The clinical noise (*bw, ma, em* at 4 SNRs) has been added randomly to the uncorrupted ECGs in order to control the SNR. The learned decision tree was translated into production rules and loaded into the cardiac monitoring system IP-Calicot [10]. IP-Calicot is a piloting system which enables the selection of signal processing algorithms on line, to treat chunks of ECG according to a context analysis. For comparison, expert rules acquired following the method described in Portet *et al.* [6] have also been tested on the dataset. To assess the maximum performance reachable with the selection rules, the best detector performance (the detector with maximum FM) for each chuck of ECG has also been retained. The results collected are used as gold standard and are grouped under the name *idealSelection*.

Table 1. Performance of the selected detectors

detector	$ER^{\pm STD}$ (%)	$Se^{\pm STD}$ (%)	$PP^{\pm STD}$ (%)	$FM^{\pm STD}$ (%)	nb of switch
af2	$51.82^{\pm 22.17}$	$92.16^{\pm 3.19}$	$67.69^{\pm 8.67}$	$78.05^{\pm 6.62}$	-
benitez	$27.87^{\pm 8.60}$	$96.46^{\pm 1.62}$	$79.85^{\pm 4.85}$	$87.38^{\pm 3.37}$	-
df2	$37.20^{\pm 23.78}$	$78.93^{\pm 13.43}$	$83.03^{\pm 11.66}$	$80.93^{\pm 11.94}$	-
gritzali	$52.10^{\pm 10.68}$	$86.66^{\pm 4.64}$	$69.09^{\pm 4.95}$	$76.89^{\pm 4.12}$	-
kadambe	$22.24^{\pm 7.77}$	$93.13^{\pm 2.99}$	$85.83^{\pm 4.80}$	**89.33**$^{\pm 3.37}$	-
mobd	$55.70^{\pm 13.69}$	$95.56^{\pm 2.51}$	$65.09^{\pm 5.37}$	$77.43^{\pm 4.26}$	-
pan	$34.90^{\pm 12.52}$	$78.20^{\pm 11.06}$	$85.65^{\pm 5.48}$	$81.76^{\pm 7.38}$	-
expert rules	$20.60^{\pm 6.62}$	$93.72^{\pm 2.62}$	$86.75^{\pm 4.12}$	**90.10**$^{\pm 3.05}$	623
DT	$22.68^{\pm 7.95}$	$94.44^{\pm 2.48}$	$84.66^{\pm 4.81}$	**89.28**$^{\pm 3.46}$	542
idealSelection	$14.38^{\pm 4.86}$	$94.85^{\pm 1.93}$	$91.13^{\pm 2.92}$	**92.96**$^{\pm 2.32}$	2119

The result synthesized in Table 1 shows that, according to the FM, expert rules outperform all the other methods (*idealSelection* is used as gold standard). The best algorithm is *kadambe* with FM=89.33% and ER=22.24%. However, *benitez* shows a Se = 96.46% superior to *kadambe*. These two algorithms outperform the others with an FM greater by 5.6%. Among the algorithm selection methods, only expert rules (ER=20.60%) outperform *kadambe* (ER=22.24%) reducing ER by 1.64%. Moreover, expert rules obtain the lowest standard deviation for FM. This shows that the selection method by expert rules is more stable than the DT and, thereby is more reliable. DT is slightly below the results of *kadambe*. *idealSelection* shows that the upper bound for FM is 92.96%. Thus, expert rules contribute to fill 21.2% of the gap between the best algorithm *kadambe* and the gold standard. This shows that the algorithm selection strategy can be greatly improved with more accurate rules. Selection rules are also interesting because they switch algorithms fewer times than *idealSelection* (2119).

5 Discussion

This experience showed that the selection of algorithms rests mainly on the acquisition of good selection rules. The experience undertaken with QRS detection algorithms shows that there remains some room for improvement. According to the gold standard, the maximum reachable FM is 92.96%. Expert rules reached 90.10%, improving the best algorithm by 0.77% in noisy contexts. This is a good score according to the current literature studies in which the sensitivity is improved typically by less than 1% [5] even in non noisy situation. This method will be applied to the other kinds of signal processing algorithms used in cardiac monitoring, such as QRS classification and P wave detection, which are less developed fields than the QRS detection and for which more significant results are expected.

References

1. Pan, J., Tompkins, W.J.: A real-time QRS detection algorithm. IEEE Trans. Biomed. Eng. 32(3), 230–236 (1985)
2. Suppappola, S., Sun, Y.: Nonlinear transforms of ECG signals for digital QRS detection: a quantitative analysis. IEEE Trans. Biomed. Eng. 41(4), 397–400 (1994)
3. Kadambe, S., Murray, R., Boudreaux-Bartels, F.: Wavelet transform-based QRS complex detector. IEEE Trans. Biomed. Eng. 47(7), 838–848 (1999)
4. Benitez, D., et al.: The use of the Hilbert transform in ECG signal analysis. Comput. Biol. Med. 31, 399–406 (2001)
5. Christov, I.: Real time electrocardiogram QRS detection using combined adaptive threshold. Biomed. Eng. Online. 3(28), 1–9 (2004)
6. Portet, F., Hernández, A., Carrault, G.: Evaluation of real-time QRS detection algorithms in variable contexts. Med. Biol. Eng. Comput. 43(3), 381–387 (2005)
7. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
8. Mark, R., Moody, G.: MIT-BIH arrhythmia data base directory. MIT Press, Cambridge (1988)
9. Moody, G., Muldrow, W., Mark, R.: A noise stress test for arrhythmia detectors. Comput. Cardiol. (1984)
10. Portet, F., et al.: Piloting signal processing algorithms in a cardiac monitoring context. In: Miksch, S., Hunter, J., Keravnou, E.T. (eds.) AIME 2005. LNCS (LNAI), vol. 3581, Springer, Heidelberg (2005)

Monitoring Human Resources of a Public Health-Care System Through Intelligent Data Analysis and Visualization

Aleksander Pur[1], Marko Bohanec[2], Nada Lavrač[2,3], Bojan Cestnik[2,4], Marko Debeljak[2], and Anton Gradišek[5]

[1] Ministry of Interior Affairs, Ljubljana, Slovenia
[2] Jožef Stefan Institute, Ljubljana, Slovenia
[3] University of Nova Gorica, Nova Gorica, Slovenia
[4] Temida, d.o.o., Ljubljana, Slovenia
[5] Dagra d.o.o., Ljubljana, Slovenia
aleksander.pur@policija.si, {marko.bohanec,nada.lavrac, marko.debeljak}@ijs.si, bojan.cestnik@temida.si, dagra@siol.net

Abstract. A public health-care system (HCS) is a complex system that requires permanent monitoring. This paper focuses on the Slovenian national HCS subsystem consisting of a network of health-care professionals at the primary care level. The challenge addressed in this paper is the development and application of intelligent data analysis, decision support and visualization methods aimed to improve the monitoring of human resources of this network. The main outcome is a set of proposed performance indicators and the developed model for monitoring the network of primary health-care professionals of Slovenia. The model enables improved planning and management through data analysis and visualization modules developed for the monitoring of physicians' qualification, age, workload and dispersion.

Keywords: intelligent data analysis, decision support systems, visualization, primary public health-care system.

1 Introduction

According to the World Health Report [1], a health-care system (HCS) is a system composed of organizations, institutions and resources that are devoted to producing a health action. Human resources are one of the main parts of this system. The main subject of this paper is the model that we have developed for monitoring and planning the network of physicians at the primary health care (PHC) level, taking into the account the physicians' qualifications, their geographic and work-time dispersion, their age and their availability for patients. The motivation for this development came from the Ministry of Health of the Republic of Slovenia, who need a holistic overview of the PHC network in order to make short- and long-term management decisions and apply appropriate management actions, as well as evaluate PHC target achievements.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 175–179, 2007.

2 Methodology

Despite many frameworks related to performance and activity monitoring (Data-driven Decision Support System (DSS) [5], Performance Monitoring, Business Performance Management (BPM), Business Activity Monitoring (BAM) [4] etc.), there is a lack of methodologies for representing the concept of monitoring based on different data analysis methods [2]. In this section, we present our approach to monitoring the physicians at the PHC level in Slovenia.

2.1 Approach to Human Resources Monitoring

Our model for monitoring the network of primary-care professionals in the Slovenian HCS consists of hierarchically connected modules. Each *module* is aimed at monitoring some aspect of the physicians at the PHC network, which is of interest for decision-makers and managers of the network (Fig. 1).

Each module involves a number of monitoring processes, which are gathered according to a given monitoring goal. Each *monitoring process* is characterised by: monitoring objectives, input data, data collecting methods, constraints on data, data dimensions, data analysis methods, output data, target criteria or target values of outputs, output data representation and visualisation methods, security requirements and the users of the monitoring system. Among these components, the *data analysis methods* transform the *input data* to *output data* represented using some *data representation formalism* according to the given *monitoring objectives*. The *target* is a level of performance that the organization aims to achieve for a particular activity. Information about *data collection* shows how and how often the data has been collected. The *constraints* define the valid input and output data. *Security requirements* define the use and management of the monitoring processes and of the data.

This approach is not limited to any particular *data analysis method*. In principle, any methods can be used, such as Structured Query Language (SQL) procedures, On-Line Analytical Process (OLAP) techniques for interactive knowledge discovering, as well as knowledge discovery in data (KDD) and data mining methods [6] for discovering important but previously unknown knowledge. The same holds for *data representation* methods, which can include pivot tables, charts, network graphs and maps.

In order to improve the comprehensibility of the model, its modules are hierarchically structured. The modules at the top level represent the main objectives. Usually all the main objectives can be incorporated in a single top-level module. The modules at a lower level are connected to the one at a higher level. Each connection represents a data channel that connects outputs of the lower level module with the inputs of a higher-level module. In principle, the hierarchy is constructed so that the results of lower-level processes could help to explain the results of monitoring processes at a higher level. For example, the module for the assessment of HCS responsiveness could be composed of the physical accessibility of Health Services, availability of resources of Health Services and the rate of visits of population to health care provider.

2.2 The Model of Monitoring of Human Resources in a HCS

The model for monitoring of the network of physicians at the PHC level is made in accordance with the above described methodology. The main concept is described by the hierarchically connected modules, shown in Fig. 1. The module *human resources* represents the main aspect of the monitoring. The lower-level modules intend to provide detailed explanations of the main aspect.

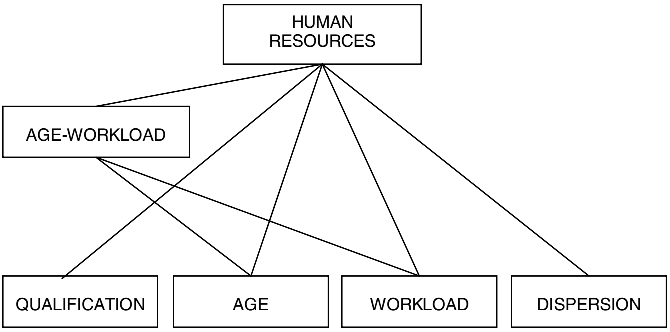


Fig. 1. The structure of the model for monitoring the network of physicians

3 Description of Individual HCS Modules

3.1 Monitoring of Human Resources

The main module *human resources* is aimed at a holistic monitoring of physicians' performance. The monitoring processes in this module intend to represent the main aspects of physicians characterised by their *qualification*, *age*, *gender*, *workload* and *dispersion*. Detailed information about these interesting aspects could be found in lower level modules. These aspects are presented by the multidimensional charts based on the OLAP techniques, and association rules discovering relations between the aspects [7]. These methods can show outliers and anomalies in the HCS.

3.2 Qualification of Physicians

The aim of the module qualification is to enable monitoring of physicians' and dentists' qualification for the job they actually perform. The main monitoring process is based on the social network visualization technique available in program named Pajek ("Spider" [3]). The monitoring of physicians' suitability is achieved by the monitoring of three variables: SPEC (specialization), LIC (licence), and OPR (the type of patients that the physician is in charge of, categorized by patient type). The Pajek diagram (Fig. 2) shows well the typical (thick lines – a high number of physicians) and atypical (thin lines – a low number of physicians) cases, which enable abnormality detection and further analysis of individual discovered anomalies.

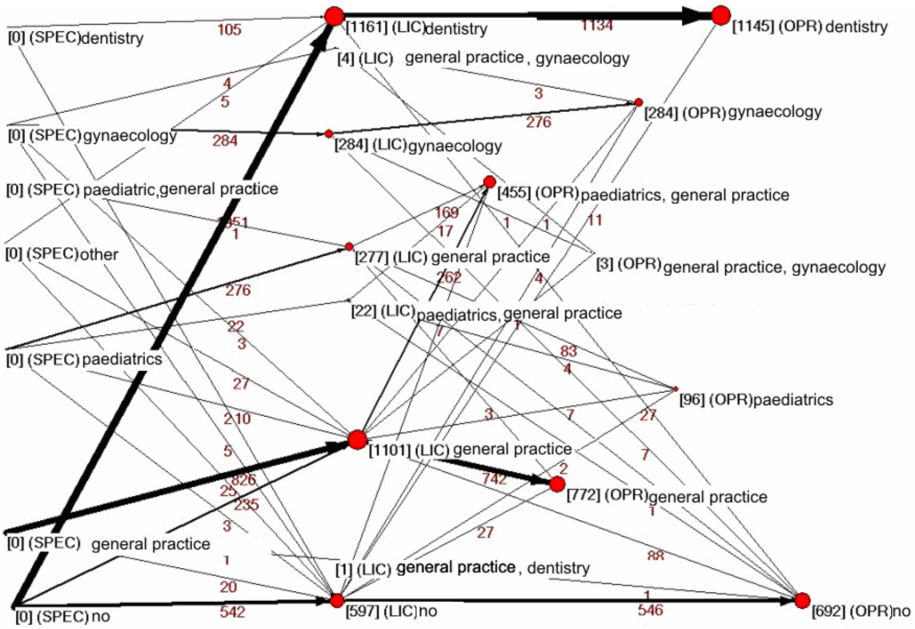

[0] (SPEC)dentistry
105
[1161] (LIC)dentistry
1134
[1145] (OPR) dentistry
[4] (LIC) general practice , gynaecology
4
5
3
[284] (OPR) gynaecology
276
[0] (SPEC) gynaecology
284
[284] (LIC) gynaecology
[0] (SPEC) paediatric,general practice
[455] (OPR) paediatrics, general practice
169
17
1
1
11
951
1
[0] (SPEC)other
[277] (LIC) general practice
[3] (OPR) general practice, gynaecology
262
4
276
[22] (LIC) paediatrics, general practice
22
3
7
1
83
4
[0] (SPEC) paediatrics
27
[96] (OPR)paediatrics
210
3
7
27
[1101] (LIC) general practice
5
7
742
2
826
25
235
[772] (OPR) general practice
1
3
27
[0] (SPEC) general practice
88
1
[1] (LIC) general practice , dentistry
20
1
[0] (SPEC) no
542
[597] (LIC) no
546
[692] (OPR) no

Fig. 2. The qualifications of physicians for the job they are performing

3.3 Short Description of Other Modules

The module *age* is aimed at monitoring the influence of physicians' age on health care network. The main monitoring process in the module is based on the OLAP model. The dimensions of this OLAP model are *age*, *gender*, *specializations* and *locations* where they work. The main monitored quantity in the facts table is the sum of physicians.

The module *workload* is aimed at monitoring of the physicians' workload. Considering the available data, the main monitoring process provides the assessment of workload based on age-adjusted listed patients per working time of physician. Where, the patients' age groups are weighted according to use of health care resources. The monitoring process is based on the OLAP model with dimensions: *time*, *specializations*, and *locations*.

The module *age-workload* are aimed at the combined analyses of physicians' age and their workload. The main monitoring process provides information of the number of physicians and regions where they have to be provided in the next years considering the physicians' age, their registered patients and working locations. This information is presented by GIS technology.

The module *dispersion* is aimed at monitoring the dispersion of locations where physicians work. Depending on the requirements, a physician may work on more than one location, but this dispersion usually means additional workload for physicians and their lower availability for patients at some location. The monitoring process provides the number of locations where physicians work, which are shown by GIS technology.

4 Conclusion

The aim of the presented human resource monitoring system is to assess performance and provide information necessary for the planning and management of primary health-care network in Slovenia. At the general level, the approach is based on a carefuly designed hierarchy of modules, each monitoring a specific aspect of the network related to human resources. In principle, the higher levels of the model provide holistic information, while the lower levels provide more details that are useful for the explanation of observed phenomena. This approach is not limited to any particular data analysis or KDD method. In this way, the monitoring system provides an information-rich picture of the network and its performance, and also helps detecting its critical aspects that require short- or long-term management actions.

The monitoring system has been developed in collaboration with the Ministry of Health of the Republic of Slovenia. Currently, it is implemented as a prototype and has been tested with real data for the year 2006. For the future, we wish that it becomes a regular tool for monitoring the health-care network in Slovenia. We also envision the application of the same methodology in other public networks, such as education and police.

Acknowledgments. We gratefully acknowledge the financial support of the Slovenian Ministry of Education, Science and Sport, and the Slovenian Ministry of Health which is the end-user of the results of this project.

References

1. World Health Organization: World Health Report 2000: Health Systems. Improving Performance (Accessed January 26, 2007) (2000) http://www.who.int/ whr/2000/en/ whr00_en.pdf
2. Bird, M.S. (ed.): Performance indicators good, bad, and ugly, Working Party on Performance Monitoring in the Public Services. J. R. Statist. Soc. A, pp. 1-26 (2005)
3. Batagelj, V., Mrvar, A.: Program for Analysis and Visualization of Large Networks. Reference Manual, University of Ljubljana, Ljubljana (2006)
4. Dresner, H.: Business Activity Monitoring, BAN Architecture, Gartner Symposium ITXPO, Cannes, France (2003)
5. Power, J.D.: Decision Support Systems. Concepts and Resources for Managers, Quorum Books division Greenwood Publishing (2002) ISBN: 156720497X
6. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco (2001)
7. Srikant, R., Agrawal, R.: Mining Quantitative Association Rules in Large Relation Tables, IBM Almaden Research Center, San Jose (1996)

An Integrated IT System for Phenotypic and Genotypic Data Mining and Management

Angelo Nuzzo[1], Daniele Segagni[1], Giuseppe Milani[1],
Cinzia Sala[2], and Cristiana Larizza[1]

[1] Department of Computer Science and Systems, University of Pavia, Pavia, Italy
[2] DIBIT, San Raffaele Scientific Institute, Milan, Italy
angelo.nuzzo@unipv.it

Abstract. This paper describes the application of an information technology infrastructure aimed at supporting translational bioinformatics studies which need the joint management of phenotypic and genotypic data. The system provides an integrated and easy to use software environment, based on data warehouse and data mining tools, to discover the most frequent complex phenotypes and search their penetrance and heritability by mapping them on the population pedigree. We first use a logical formalization to define phenotypes of interest in order to retrieve individuals having that phenotype from the electronic medical record. We then use an open-source Web-based data warehouse application for analyzing phenotypic data and presenting the results in a multidimensional format. Relationships between the selected individuals are automatically visualized by integrating in the system an ad-hoc developed pedigree visualization tool. Finally, the application of the system to support a genetic study of an isolated population, the Val Borbera project, is presented.

Keywords: genotype-phenotype association, clinical data warehouse, data mining, Web-based systems.

1 Introduction

A specific characteristic of the post-genomic era will be the correlation of genotypic and phenotypic information [1][2]. In this context, the studies aimed at the so-called genetic dissection of complex traits represent a first crucial benchmark for Biomedical Informatics and for Translational Bioinformatics. The definition of an Information Technology infrastructure is crucial to support this kind of studies in both phenotype discovering and genotypic traits mapping. This kind of studies is based on different types of data, i.e. clinical, genetic and genealogical data. Thus the final objective of an IT-based data management system is to be able of combining such different sources of information in an integrated framework, in order to make data investigation more efficient and easier for the final user and to improve the knowledge discovery process.

In this paper we describe the architecture of an integrated application which embeds different software tools with the aim of exploiting and integrating automatically clinical and genealogical data to support genetic studies. In particular, we are using

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 180–184, 2007.

such system in the Val Borbera Project [3] to explore the phenotypic and genealogic data of an Italian isolated population.

2 Methods

The overall strategy of geneticists' analysis is made of 3 main steps: i) discover the phenotype or clinical condition to be investigated, given the clinical data of the population, ii) search relationships between individuals showing the same phenotype (if any, genetic causes may be supposed), iii) choose appropriate loci to be genotyped to identify genotype-phenotype association. Thus the final purpose of the system that we have developed is to support each of the previous steps. In particular, the system provides tools for the following main tasks:

1. formally defining the phenotypes to be investigated using a graphical user interface;
2. exploring clinical data to extract and analyze individuals with the same phenotypes (as they have been previously defined);
3. mapping the individuals extracted onto the population pedigree, to analyze their relationships and find possible heritability paths.

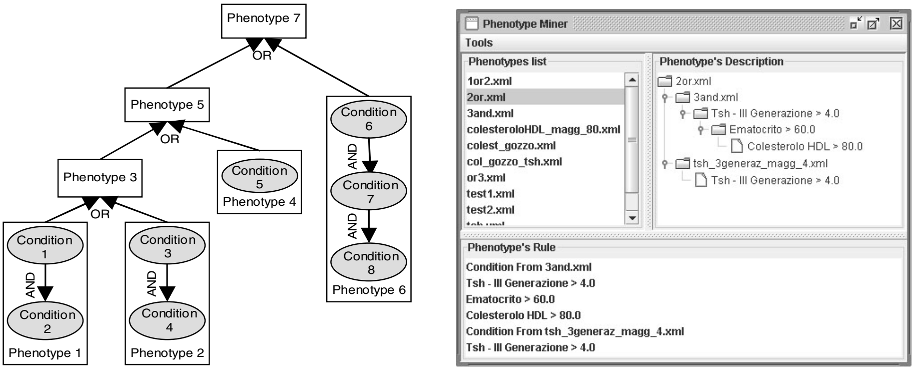
2.1 Phenotype Definition Tool

A first crucial task that hampers the development of automated IT solutions in genetic studies is an appropriate definition and identification of the phenotypes that geneticists want to investigate. We have defined and provided a formalization of the concept of phenotype to represent, maintain and manage its definition into the framework. Our basic assumption is to consider a phenotype as a set of conditions in the form of attribute/value pairs, combined with logical operators (AND, OR) that allow the definition of more and more complex phenotypes. In particular, the AND operator allows the specialization of a phenotype, while the OR operator is used to merge different phenotypes into a single more comprehensive one (fig.1, left). With this assumption, it has been possible to develop a dynamic query generator, the "Phenotype Editor", easily exploitable via a graphical user interface (fig. 1, right), so that no technical skills about query languages are required to the user.

2.2 OLAP Engine

Dealing with clinical data to analyze phenotypic information implies to take into account their heterogeneity and provide a browsing interface that allows their investigation as they were homogeneous. Our proposed solution is based on the use of a tool for the multidimensional inspection of the dataset [4], [5]. The technique of multidimensional analysis is implemented with software tools called Online Analytical Processing (OLAP) engines. This kind of tools has been developed for business and economical data mining: for the first time we use them in a clinical context.

In our system we use an OLAP engine written in Java programming language, called Mondrian [6]. It executes queries written in the MDX language (that has become a standard for data warehouse applications) [7], reads data from a relational



Phenotype Miner

Tools

Phenotypes list

1or2.xml
2or.xml
3and.xml
colesteroloHDL_magg_80.xml
colest_gozzo.xml
col_gozzo_tsh.xml
or3.xml
test1.xml
test2.xml

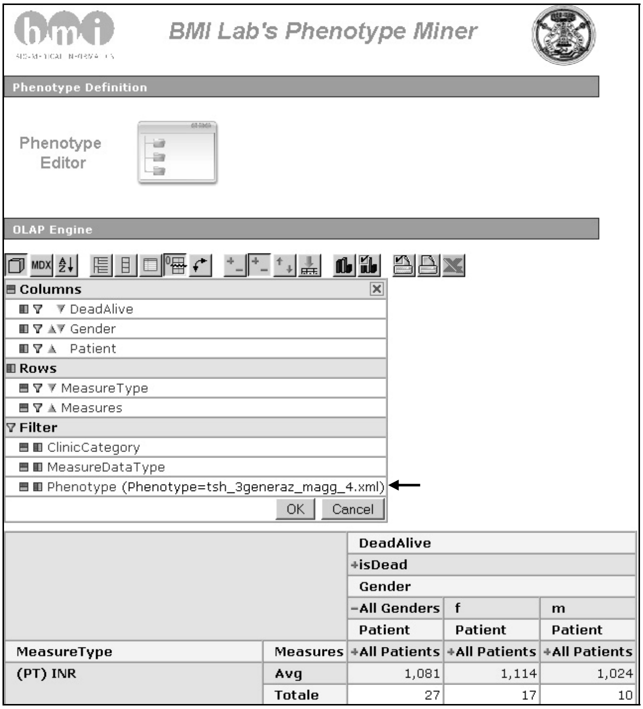
Phenotype's Description

2or.xml
3and.xml
Tsh - III Generazione > 4.0
Ematocrito > 60.0
Colesterolo HDL > 80.0
tsh_3generaz_magg_4.xml
Tsh - III Generazione > 4.0

Phenotype's Rule

Condition From 3and.xml
Tsh - III Generazione > 4.0
Ematocrito > 60.0
Colesterolo HDL > 80.0
Condition From tsh_3generaz_magg_4.xml
Tsh - III Generazione > 4.0

Fig. 1. Left: graphical representation of some phenotypes organized in a hierarchical structure reflecting the formal phenotype definition. Right: A screenshot of the "Phenotype Editor" tool showing a real case of phenotype definition based on the previous schema.



Columns
DeadAlive
Gender
Patient
Rows
MeasureType
Measures
Filter
ClinicCategory
MeasureDataType
Phenotype (Phenotype=tsh_3generaz_magg_4.xml)

		DeadAlive		
		+isDead		
		Gender		
		−All Genders	f	m
		Patient	Patient	Patient
MeasureType	Measures	+All Patients	+All Patients	+All Patients
(PT) INR	Avg	1,081	1,114	1,024
	Totale	27	17	10

Fig. 2. Mondrian section which allows phenotype selection (indicated by the arrow) and other dimensions for data investigation

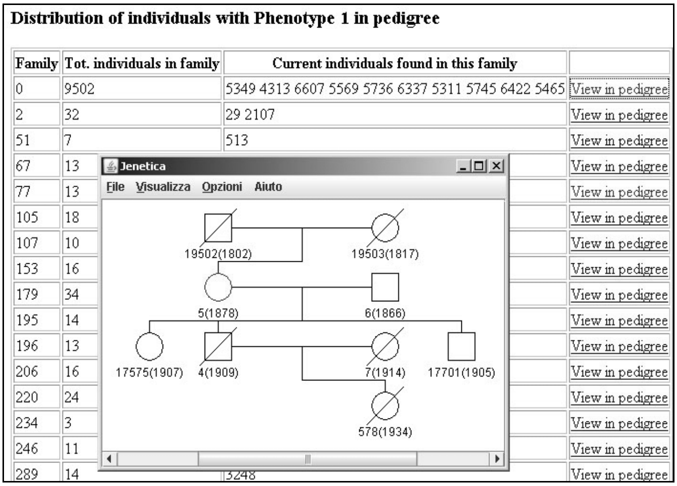
database, and presents the results in a multidimensional format through a Java API, so that the final user may choose the preferred presentation layer. We developed a Web application based on JSP pages to integrate it with the Phenotype Editor as well as the pedigree visualization tool (described in next paragraph) deployed as Java Web Start applications.

The phenotype investigated is modeled as a dimension of the multidimensional data structure. Once the user selects a phenotype, the information about individuals having that phenotype are provided as input to the OLAP engine to be summarized in a unique table, which values may be inspected by expanding or collapsing its rows and columns (figure 2).

2.3 Pedigree Visualization

As described above, finding phenotypes of interest is the first step of a genetic study. The following step is to discover which of the most common phenotypes may be influenced by genetic factors. This means that the phenotype must be mapped over the individuals' pedigree, so that, comparing its distribution with the genetics markers' one, it could be possible to highlight common heritability paths suggesting genotype-phenotype association.

To support this task, we have developed a "search engine" that automatically groups into families the individuals with the same phenotype (discovered using the OLAP engine), and shows them with a pedigree visualization tool (figure 3). The family search is based on the output files generated by PedHunter [8], an open-source pedigree analysis tool, while the pedigree visualization is provided by a Java based graphical tool we developed on top of the open-source library Pelican [9], called Jenetica.



Distribution of individuals with Phenotype 1 in pedigree

Family	Tot. individuals in family	Current individuals found in this family	
0	9502	5349 4313 6607 5569 5736 6337 5311 5745 6422 5465	View in pedigree
2	32	29 2107	View in pedigree
51	7	513	View in pedigree
67	13		View in pedigree
77	13		View in pedigree
105	18		View in pedigree
107	10		View in pedigree
153	16		View in pedigree
179	34		View in pedigree
195	14		View in pedigree
196	13		View in pedigree
206	16		View in pedigree
220	24		View in pedigree
234	3		View in pedigree
246	11		View in pedigree
289	14	3248	View in pedigree

Fig. 3. The page in the background shows how the individuals with the selected phenotype are distributed into the population families. Clicking on the right column, the family structure is shown by the pedigree visualization tool (in foreground).

3 Results and Conclusion

The system described in the previous section is still under development, but it has been already used for several tests on a real dataset, the clinical database of the Val

Borbera isolated population study [3]. Geneticists can dynamically compose queries on the dataset using the graphical Phenotype Editor: we verified that the resulting initial report table is always the same as the one obtained by a technician performing a SQL statement (which is typically quite a lot of instruction lines long). They can next perform a real-time research of which phenotypes could be of particular interest for the population exploring data at different levels of detail using the OLAP engine.

Then the automatic mapping of the selected phenotype on the pedigree allows the geneticists to do hypotheses on the genetic origin of that phenotype and make suitable choices for the following genotyping and genetic analysis. The system has the advantage of automating and make transparent for the end-user the access to several tools heterogeneous as regards their implementation, their interface and the input/output data format. The integrated environment, by allowing a simple access to them through a unique interface, could be particularly useful for doing efficiently and, if necessary, iteratively, the various steps of the phenotypic data analysis, before planning the more expensive step of individuals genotyping.

References

1. Lander, E.S., Schork, N.: Genetic dissection of complex traits. Science 265(5181), 2037–2048 (1994)
2. Botstein, D., Risch, N.: Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nature Genetics 33, 228–237 (2003)
3. Sala, C., Bione, S., Crocco, L., Gatti, M., Poggiali, E., Bellazzi, R., Buetti, I., Rognoni, C., Camaschella, C., Toniolo, D.: The Val Borbera Project: epidemiological and genealogical analysis of an isolated population in Northern Italy. European Society of Human Genetics (submitted, 2006)
4. Kinball, R., Ross, M.: The Data Warehouse Toolkit, 2nd edn. Wiley and Sons Inc., Chichester (2002)
5. Wyderka, K.: Data Warehouse Technique for Outcomes Management. Health Management Technology 20(10), 16–17 (1999)
6. Hyde, J.: Mondrian OLAP project, Pentaho Analysis Service http://mondrian.pentaho.org/
7. Spofford, G., Harinath, S., Webb, C., Huang, D.H., Civardi., F.: MDX Solutions, 2nd edn. Wiley Publishing Inc., Chichester (2006)
8. Agarwala, R., Biesecker, L.G., Hopkins, K.A., Francomano, C.A., Schaffer, A.A.: Schaffer-Software for Constructing and Verifying Pedigrees Within Large Genealogies and an Application to the Old Order Amish of Lancaster County. Genome Research 8, 211–221 (1998)
9. Dudbridge, F., Carver, T., Williams, G.W.: Pelican: pedigree editor for linkage computer analysis. Bioinformatics 20(14), 2327–2328 (2004)

Automatic Retrieval of Web Pages with Standards of Ethics and Trustworthiness Within a Medical Portal: What a Page Name Tells Us

Arnaud Gaudinat, Natalia Grabar, and Célia Boyer

Health on the Net Foundation, SIM/HUG, Geneva, Switzerland
name.surname@healthonnet.org

Abstract. The ever-increasing volume of health online information, coupled with the uneven reliability and quality, may have considerable implications for the citizen. In order to address this issue, we propose to use, within a general or specialised search engine, standards for identifying the reliability of online documents. Standards used are those related to the ethics as well as trustworthiness of websites. In this research, they are detected through the URL names of Web pages by applying machine learning algorithms. According to algorithms used and to principles, our straightforward approach shows up to 93% precision and 91% recall. But a few principles remain difficult to recognize.

1 Introduction

The issue related to the quality of online information is important, specially in the medical area, as eight Internet users out of ten look for health information [1] and as often such searches are directly linked to their own health condition or to their relatives. But the quality and reliability of proposed online health documents are uneven and we assume that this should be clearly indicated to users. Various initiatives exist for the quality control assessment of health information on Internet [2]. At the Health on the Net Foundation (*www.hon.ch*), we have adopted an accreditation program through the third party evaluation of health website's reliability done according to the Ethical Code of Conduct HONcode [3]. The Code is composed of eight ethical principles, namely *authority*, *complementarity*, *privacy*, *reference*, *justifiability*, *transparency*, *sponsorship* and *advertising*. Each website, which asks for the accreditation, is evaluated by HON's experts in order to check whether it provides clear statements for these principles. Up to now, the HONcode accredited database contains over 1'200'000 Web pages in 32 languages. When perfomed manually the accreditation process guarantees high quality results but must cope with the increasing number of online health information. In this work, we want to take advantage of the database with quality annotated websites and to propose a method and data suitable for the automatic detection of health websites' quality.

We apply supervised learning methods: they allow to better characterise and constrain expected categories related to the eight HONcode criteria. In previous

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 185–189, 2007.

Table 1. Learning data: numbers of URLs used for generation of the language model in English

Principle	Meaning	Total	Learning	Evaluation
HC1	Authority	2843	2571	272
HC2	Complementary	2470	2218	252
HC3	Privacy	2374	2115	259
HC4	Reference	1855	1674	181
HC5	Justifiability	460	407	53
HC6	Transparency	2539	2275	264
HC7	Sponsorship	2088	1893	195
HC8	Advertising	1545	1389	156
HC9	Date	1545	1378	167

research, regular expression [4] or presence of HONcode label [5], have been used. Comparing to these, supervised learning methods allow to formalize textual events with more precision, and to capture events which would be not detected by humans. Moreover, categorisation methods shown to be helpful in automatic systems working with textual documents [6,7]. In previous work, we proposed an automatic tool for the categorization of Web pages according to the HONcode principles on the basis of documents' content [8]. We propose now to apply similar method for the categorization of documents through their URL addresses.

2 Material

A key component of any system for the automatic text categorisation is a knowledge base with positive examples. In this work, we use the name of URL Web page. URL is the *Uniform Resource Locator* which indicates the location of a Web page on Internet. Each URL is unique. URL begins with the scheme name that defines its namespace, while the remaining part of the URL corresponds to the hierarchical structure of website and to the name of file. The reason to use URLs as material for the categorisation of Web pages is that they can be composed with keywords related to the HONcode principles. Here, a few examples of URL names registered for the *privacy* principle within the HONcode accredited database:

anatome.ncl.ac.uk/tutorials/privacy.html
www.wmcnet.org/workfiles/media/noticeofprivacyplan.pdf
parathyroid.com/disclaimer.htm, *www.vh.org/welcome/help/vhpolicies.html*

The learning dataset is composed of over 12'623 URLs of some HONcode accredited English sites. Table 1 indicates number and title of principles, and number of the URLs used in our work. As the principle HC4 *reference* covers heterogeneous information (reference to date and reference to statement on clinical trials, etc.), it has been separated into two sets, and *date* has been exported. Additionally, notice that some of the URLs recorded can cover up to 5 principles.

3 Method

Pre-processing of material. For the detection of significant parts of URLs, for instance, *www.hon.ch/confidentiality_page/privacy_disclaimer.html*, we split them two parts, *inurl* and *endurl*:

- *inurl*, when exists, corresponds to the directory names in which the file is located. It includes the entire directory pathway except the domain and file names. In the example above, *inurl* is *confidentiality_page*
- *endurl*, corresponds to the file name: *privacy_disclaimer*

inurl and *endurl* are segmented on non alphanumeric characters, *ie.* _ - ? / =

Training step. Machine learning algorithms used are those proposed by our learning framework [9]: Naive Bayes (NB), Support Vector Machine (SVM), k-Nearest Neighbours (kNN) and Decision Tree (DT). Different combinations of features and categorisation algorithms have been applied to data in English. *Features* tested within the learning process are the following: (1) word combination (*e.g.* n-grams of 1 to 4 words); and (2) character combination (*e.g.* n-grams of 1 to 5 characters). *Unit weight* is defined by three elements [10]: term frequency, inverse document frequency and length normalisation. *Features selection*, which aims at reducing vector-space dimension through selection of the most discriminatory features, is performed with document frequency (DF) [11].

Evaluation. We used 10% of our corpora for the evaluation task, the 90% being used exclusively for the learning task. These two corpora are independent. Evaluation is performed with four measures in their micro and macro versions: precision, recall, F-measure and error rate. Macro precision (*maP*) is representative of the distribution of features in each category (principle), and micro precision (*miP*) in each processed unit (URLs).

4 Results and Discussion

The method has been applied to three sets of material:

- *inurl*: learning and evaluation performed on *inurl* parts of URLs;
- *endurl*: learning and evaluation performed on *endurl* parts of URLs;
- *red*: learning and evaluation performed on reduced set of 6 principles: *authority*, *complementary*, *privacy*, *transparency*, *sponsorship* and *advertising*.

Among all the methods, features and weightings indicated, we present only those which show significant differences between them, namely: two learning methods (*NB* and *SVM*); two features (single words *w1* and 5-grams of characters *c5*). Figure 1 presents figures for the average precision and recall obtained with these three sets, and we can distinguish three clusters: (1) *inurl* set, which provides the less performing results for both recall and precision; (2) *NB* method generates good recall figures (with *endurl* and *red* sets); (3) *SVM* method generates good precision figures (with *endurl* and *red* sets). We assume that merging these two

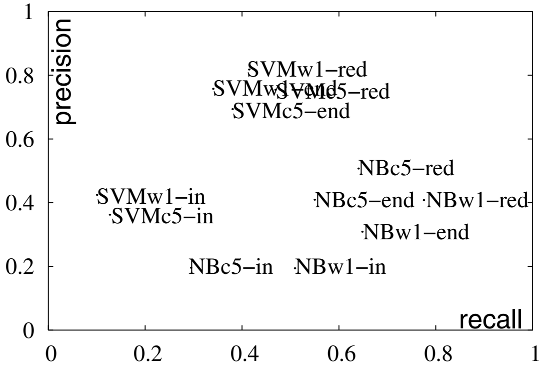


Fig. 1. Micro precision and micro recall of various methods applied

Table 2. Categorisation of URL names (*endurl*). Recall/Precision contingency of quality criteria. System setting: method *SVM*, language *English*, single word *w1.*

Auto/Man	Authority	Compl.	Privacy	Refer.	Justif.	Trans.	Spon.	Adver.	Date
Authority	**68/84**	7/6	2/1	2/5	0/0	7/5	11/36	2/1	1/9
Compl.	5/4	**64/38**	7/4	1/3	5/29	5/3	1/3	12/33	0/0
Privacy	2/2	2/2	**93/87**	0/0	0/0	0/1	1/3	2/7	0/0
Refer.	4/2	4/1	4/1	**51/62**	9/29	0/0	2/3	0/0	24/48
Justif.	25/1	25/1	0/0	0/0	**25/7**	0/0	0/0	0/0	25/4
Trans.	1/2	1/1	3/3	1/5	1/7	**91/91**	1/3	1/3	1/9
Spon.	8/2	4/1	0/0	8/5	0/0	4/1	**76/53**	0/0	0/0
Adver.	8/2	15/3	12/2	8/5	4/7	0/0	0/0	**50/47**	0/0
Date	5/1	5/1	5/1	26/14	16/21	5/1	0/0	0/0	**37/30**

methods, SVM and NB, can be interesting: NB can guarantee better recall and SVM better precision. Furthermore, we can observe that, not surprisingly, *red* set allows to generate better results than *endurl*.

Table 2 indicates the precision/recall confusion matrix obtained with *SVM* algorithm applied to single words *w1* from *endurl* with no weighting. We can observe that out of nine criteria the following ones could be processed with good results: *transparency* (91%/91%), *privacy* (93%/87%), *authority* (68%/84%), and *sponsorship* (76%/53%). Web Pages related to principles *complementarity* and *advertising* are categorised with mean performances. As for three remaining principles (*justifiability*, *date* and *reference*) they show scarce performances, and other approaches should be applied for their detection.

5 Conclusion and Perspectives

In this paper, we presented a novel approach for the categorisation of Web pages according to the quality and announced ethical policy of the websites. We exploit for this the HONcode accredited database and two machine learning algorithms (SVM and Naive Bayes). These algorithms have been applied to URL names of Web pages and show competitive results, up to 93% precision and 91% recall according to principles. URL-based categorisation can be thus run separately or in combination with the content analysis. In our further work, our special attention should be given to the combination of both approaches (URL and content based), to two hard to modelise principles (*reference* and *justification*), and to the visualisation of the quality information within a search engine.

Acknowledgement. This work has been realised as part of the PIPS (Personalised Information Platform for Life & Health) project funded by the European Commission programme *Information society technology*, contract nb 507019.

References

1. Fox, S.: Online Health Search 2006. Most Internet users start at a search engine when looking for health information online. Very few check the source and date of the information they find. Technical report, Pew Internet & American Life Project, Washington DC (2006)
2. Risk, A., Dzenowagis, J.: Review of internet information quality initiatives. Journal of Medical Internet Research 3(4), e28 (2001)
3. Boyer, C., Baujard, O., Baujard, V., Aurel, S., Selby, M., Appel, R.: Health on the net automated database of health and medical information. Int J Med Inform 47 (1-2), 27–29 (1997)
4. Wang, Y., Liu, Z.: Automatic detecting indicators for quality of health information on the web. International Journal of Medical Informatics (2006)
5. Price, S., Hersh, W.: Filtering web pages for quality indicators: an empirical approach to finding high quality consumer health information on the world wide web. In: AMIA 1999, pp. 911–915 (1999)
6. Vinot, R., Grabar, N., Valette, M.: Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'internet. In: TALN, pp. 257–284 (2003)
7. Wang, Y.: Automatic recognition of text difficulty from consumers health information. In: IEEE. (ed.) Computer-Based Medical Systems (2006)
8. Gaudinat, A., Grabar, N., Boyer, C.: Machine learning approach for automatic quality criteria detection of health webpages. In: McCray, A. (ed.) MEDINFO 2007, Brisbane, Australia (to appear, 2007)
9. Williams, K., Calvo, R.A.: A framework for text categorization. In: 7th Australian document computing symposium (2002)
10. Salton, G.: Developments in automatic text retrieval. Science 253, 974–979 (1991)
11. Koller, D., Sahami, M.: Toward optimal feature selection. In: International Conference on Machine Learning, pp. 284–292 (1996)

A Mixed Data Clustering Algorithm to Identify Population Patterns of Cancer Mortality in Hijuelas-Chile*

Eileen Malo[1], Rodrigo Salas[2], Mónica Catalán[3,4], and Patricia López[1]

[1] Universidad de Valparaíso; Departamento de Computación; Chile
eileen.malo@gmail.com, patricia.lopez.campos@gmail.com
[2] Universidad de Valparaíso; Departamento de Ingeniería Biomédica; Chile
rodrigo.salas@uv.cl
[3] Universidad de Valparaíso; Departamento de Estadística; Chile
monica.catalan@uv.cl
[4] Universidad de Carlos III de Madrid; Departamento de Estadística; España

Abstract. The cancer disease in Hijuelas-Chile represents the 45% of the population deaths in the last decade. This high mortality rate have concerned the sanitary authority that lacks of information to identify the risk groups and the factors that influence in the disease.

In this work we propose a clustering algorithm for mixed numerical, categorical and multi-valued attributes. We apply our proposed algorithm to identify and to characterize the common patterns in people who died of cancer in the population of Hijuelas between 1994 and 2006. As a consequence of this research, we were able to characterize the people who died of Cancer in Hijuelas-Chile.

1 Clustering Algorithm for Cancer Data

In Chile, according to the Chilean Ministry of Health [1] and the study of E. Medina and A. Kaempffer [6], the cancer is responsible for 24.2% of total death at the year 2000 and shows an upward trend with increasing mortality rates from 99 to 122.8 per 100000 habitants in the period 1980-2002.

Hijuelas is a small area located in the Aconcagua valley in the 5th Region, central part of Chile, and it is characterized by its agricultural activity. Unfortunately the cancer diseases in Hijuelas represent the 45% of the population deaths in the last decade. This high mortality rate has concerned the sanitary authority that lacks of information to identify the risk groups and the factors that influence in the disease. Unfortunately, Hijuelas does not account with a cancer registry and the data had to be collected in terrain from the death certificates and the clinical cards.

The cluster analysis is a descriptive method that could be used to studying mortality in epidemiology. A conglomerate of cases may be defined as the onset

* This work was supported by Research Grant Fondecyt 1061201 and DIPUV-REG. N32/2005.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 190–194, 2007.

of a number of cases of a disease larger than expected for a certain population group, geographical area or timeframe. In the case of cancer, the study of conglomerates entails a number of specific characteristics as compared to other groups of diseases that will support new prevention plans to reduce the mortality rates or to outline new investigations (see [4]).

We propose a novel technique called *C-Prototypes* that improves the K-prototypes clustering algorithm for mixed attributes [2]. The algorithm was inspired in the population cancer mortality data that is composed with numerical, categorical and multi-valued attributes. The multi-valued attributes correspond to variables that could take several values in one realization. In the Cancer data the *cause of death* can take several values at a time, because a person can have several causes of death in no particular order.

The aim is to find a partition $P = \{P_1, ..., P_K\}$ such that: $P_i \neq \phi$, $i = 1..K$; $\bigcup_{i=1}^{K} P_i = \mathcal{X}$; and $P_i \bigcap P_j = \phi, \forall i \neq j$. The algorithm 1 shows the procedure of the C-*Prototype* clustering algorithm to find the centroids of the K clusters.

Algorithm 1. C-Prototype Algorithm

```
Given is a training data set $\mathcal{D}$ with $N$ elements.
Pick $K$ the number of partitions. Let the set of prototypes $Q = \{\mathbf{q}_1, ..., \mathbf{q}_K\}$,
initialized with a random sample of $K$ elements taken from $\mathcal{D}$.
repeat
    Let the clusters $P_1, .., P_K$ be empty sets, $P_k = \phi, k = 1..K$
    for all $\mathbf{x} \in \mathcal{D}$ do
        Find the closest prototype $\mathbf{q}^*$ to the sample $\mathbf{x}$ as $\mathbf{q}^* = \arg\min_{\mathbf{q} \in Q} d(\mathbf{x}, \mathbf{q})$, where
        $d(\mathbf{x}, \mathbf{q})$ is calculated with equation (1).
        Add the sample $\mathbf{x}$ to the cluster $P^*$, $P^* = P^* \bigcup \{\mathbf{x}\}$.
    end for
    Update all the prototypes as the centroids of the samples assigned to their respective clusters with equations:

        $\mathbf{q}_k^n(l) = \frac{1}{\#(P_k)} \sum_{\mathbf{x} \in P_k} \mathbf{x}^n(l)$    $l = 1..d_n$
        $\mathbf{q}_k^c(l) = mode_1\{\mathbf{x}^c(l), \mathbf{x} \in P_k\}$    $l = 1..d_c$
        $\mathbf{q}_k^m(l) = \{mode_1\{\mathbf{x}^c(l), \mathbf{x} \in P_k\}, ..., mode_T\{\mathbf{x}^c(l), \mathbf{x} \in P_k\}\}$    $l = 1..d_m$

    where $mode_j\{\mathbf{x}^c(l), \mathbf{x} \in P_k\}$ returns the $j-$th most frequent value of the $l-$th
    attribute of the vector $\mathbf{x}$ in the set $P_k$.
until The values of the prototypes do not change
Return the set of prototypes $Q$.
```

Let the data sample $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, where each sample is a vector $\mathbf{x} = [(\mathbf{x}^n)^T, (\mathbf{x}^c)^T, (\mathbf{x}^m)^T]^T$ that corresponds to a composition of numerical $\mathbf{x}^n = [x^n(1), ..., x^n(d_n)]^T$, categorical $\mathbf{x}^c = [x^c(1), ..., x^c(d_c)]^T$ and multi-valued $\mathbf{x}^m = [x^m(1), ..., x^m(d_m)]^T$ attributes.

To measure the difference between two elements we need to define a dissimilarity function. This is accomplished by defining distances for each type of attributes and then combining them in a lineal convex combination:

$$d(\mathbf{x},\mathbf{y}) = \alpha d_{num}(\mathbf{x}^n,\mathbf{y}^n) + \beta d_{cat}(\mathbf{x}^c,\mathbf{y}^c) + \gamma d_{mult}(\mathbf{x}^m,\mathbf{y}^m) \tag{1}$$

where $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$, $0 \leq \gamma \leq 1$ and $\alpha + \beta + \gamma = 1$. The values α, β y γ will correspond to the degree of importance that are given to the numerical, categorical and multi-valued attributes respectively. E. Malo [5] demonstrated that the previous distance functions satisfy the metric conditions.

The distance of the numerical attributes is $d_{num}(\mathbf{x}^n,\mathbf{y}^n) = \frac{d_{euclidean}(\mathbf{x}^n,\mathbf{y}^n)}{1+d_{euclidean}(\mathbf{x}^n,\mathbf{y}^n)}$, where $d_{euclidean}(\mathbf{x}^n,\mathbf{y}^n) = \sqrt{\sum_{l=1}^{d_n}(x^n(l) - y^n(l))^2}$ is the euclidean distance. The distance of the categorical attributes is $d_{cat}(\mathbf{x}^c,\mathbf{y}^c) = \frac{1}{d_c}\sum_{l=1}^{d_c}\delta(x^c(l),y^c(l))$, $\delta(x^c(l),y^c(l)) = \begin{cases} 1 & \text{if } x^c(l) \neq y^c(l) \\ 0 & \text{if } x^c(l) = y^c(l) \end{cases}$ is the Hamming distance. Finally, the distance of the multi-valued attributes is $d_{mult}(\mathbf{x},\mathbf{y}) = \frac{1}{d_m}\sum_{l=1}^{d_m}\frac{d_{\mathcal{M}}(x^m(l),y^m(l))}{1-d_{\mathcal{M}}(x^m(l),y^m(l))}$, where both $x^m(l)$ and $y^m(l)$ consist in a set of categorical values and
$d_{\mathcal{M}}(x^m(l),y^m(l)) = \frac{1}{2}\left(\#(x^m(l)\bigcup y^m(l)) - \#(x^m(l)\bigcap y^m(l))\right)$

2 Descriptive Analysis of the Population Cancer Mortality Data

In this section a descriptive analysis of the Hijuelas population cancer mortality data is made with the C-Prototype algorithm (further results can be found in [5]). The aim of this analysis is to find patterns that characterize and give a profile of the people who died of Cancer in Hijuelas, Chile. The data was collected in terrain from papers files retrieved from the Hijuelas Civil Registry and from the Hijuelas Medical Primary Attentions. We were able to register the death certificates and the medical cards of only 189 persons that died during the period 1994-2006, unfortunately, several persons (about 200) were not included because of missing data.

The data collected consist in the personal information and the cause of death. The cause of death could be cancer or another pathology and is at least one and at most five. We use the ICD-10 standard to identify the type of cancer [3]. The collected and selected attributes for each person are: **the numerical attributes:** [n1] age at the diagnostic day (in years), [n2] age at the day of death (in years), [n3] survival time (in months), [n4] weight at the diagnostic day (in kilograms); **the categorical attributes:** [c1] gender (Male (M), Female (F)), [c2] marital status (single (S), married (M) and widow (W)), [c3] living area (Ocoa (O), Petorquita (P), Romeral (R)), [c4] occupation (farmer (F), teacher (T), worker (W), housewife (H), pensioner (P), seasonal worker (S)), [c5] health assurance (Fonasa A (FA), Fonasa B (FB), Fonasa C (FC), Fonasa D (FD), Isapre (I), None (N)), [c6] smoking (Yes (Y), No (N)), [c7] place of living (high residential (HR), low residential (LR), rural (R), industrial (I), centric(C)); and

the multi-valued attribute: [m1] Cause of death (according to the ICD-10 standard).

The dataset was separated according to the living area attribute [c3] and the gender attribute [c1] was not considered. We selected the three most frequent cause of death ($T = 3$) for the multi-valued attribute of the prototypes, where m1-1 is the most frequent, m1-2 is the second most frequent and m1-3 is the third most frequent value. The result of the descriptive analysis is shown in table 1, and shows that the most common causes of death that appear in the clusters were: the malignant neoplasms that affect the Digestive organs **C15-26**, the respiratory and intrathoracic organs **C30-39**, the central nervous system **C69-72**; diseases of the circulatory system **I00-52**; and, diseases of the respiratory system **J80-99**.

Table 1. Prototypes description found by considering the dataset divided according the area. P_1, P_2 and P_3 correspond to the clusters of the Petorquita area; O_1 and O_2 correspond to the clusters of the Ocoa area; R_1, R_2 and R_3 correspond to the cluster of the Romeral area.

Attribute	P_1	P_2	P_3	O_1	O_2	R_1	R_2	R_3
n1	45.4	52.3	77.6	48.6	65.0	65.2	48.5	57.0
n2	51.4	55.8	78.8	49.1	65.2	67.0	49.0	63.0
n3	65.6	32.5	12.3	9.8	12.0	15.1	6.5	65.0
n4	59.3	48.2	58.0	60.1	63.7	63.4	60.2	57.2
c2	M	M	M	M	M	S	M	M
c4	H	F	F	H	F	F	F	F
c5	N	N	N	N	N	N	FA	N
c6	N	N	N	N	N	N	N	N
c7	R	R	R	R	R	R	R	LR
m1-1	C15-26	C15-26	C15-26	C15-26	C15-26	C15-26	I00-52	C15-26
m1-2	J80-99	I00-52	I00-52	I00-52	I00-52	I00-52	C15-26	C30-39
m1-3	C30-39	C30-39	C30-39	C30-39	C30-39	C30-39	C69-72	J80-99

Note that in most of the clusters found are composed mostly with persons that lived in rural areas that were married, they worked as a farmer or they were housewife, they did not have a health assurance and they did not smoke. The persons did not have a health assurance due, probably, that they were independents workers with low income, and they did not care for their health. The fact that this persons lived in rural areas and they are mostly farmers make us suspect that they were using forbidden pesticides in their farms, they did not have the adequate implements to avoid the exposure and they consumed the contaminated products.

The Petorquita area has the higher level of industry, urbanization and habitants. On the other hand, the industrial activity in Ocoa is rather low, while the Romeral area is mostly agricultural with almost no industrial activity. It is important to note that despite the level of industrialization; most of the clusters correspond to rural areas. Most of the people are non smoking that lived in a rural area however one of the highest cause of death is the pulmonary cancer C30-39 this make us reaffirm our suspicion in the use of forbidden pesticides.

Furthermore note that the survival time of the cluster conformed with third age persons (≥ 65) ranges between 11.9 and 20 months while the cluster conformed with adult persons (< 65) the survival time ranges between 32.5 and 65.6 months with the exception of the cluster O_1 from Ocoa and R_2 from Romeral with a survival time of 9.8 and 6.5 respectively. This phenomenon is due probably because the habitants have less access to medical facilities than the people living in Petorquita.

3 Concluding Remarks

This paper has shown the importance of the application of data mining techniques to extract information and to support the decision making of the pertinent authorities. We have proposed a novel clustering algorithm called *C-prototypes* that has been designed inspired in the structure of the population Cancer mortality dataset.

In the descriptive analysis of the population cancer mortality in Hijuelas we were able to conclude that the group of people were characterized as farmers with low income who died due to the stomach and pulmonary cancer. This draw the hypothesis for further studies about the environmental influence of forbidden pesticides used in the agriculture, the lack of proper equipment to avoid the exposure and the lack of access to medical facilities in the rural areas.

Acknowledgement. The authors wishes to thank Sra. Verónica Rossat, Dr. Orlando Orrego, Dr. Alfredo Aroca and Sra. María Elena Bitterlich for their valuable assistant and help.

References

1. de Salud de Chile, M.: Programa nacional del cáncer (2004) http://www.minsal.cl
2. Huang, Z.: Clustering large data sets with mixed numeric and categorical values. In: Proceedings of First Pacific-Asia Conf. Knowledge Discovery and Data Mining vol. 1, pp. 21–34 (1997)
3. Karhalainen, A.: International statistical classification of diseases and related health problems (icd-10) in occupational health, Publications WHO/SDE/OEH/99.11, World Health Organization (1999)
4. Lertsundi-Manterola, A., Saez, M., Marcos-Gragera, R., Izquierdo, A., Pibernat, N., Sala, E., Camps, N.: Análisis de conglomerados de cáncer. el caso del barrio de campdora, girona. Revista Espaola de Salud Publica, vol. 79, pp. 443–452 (2005)
5. Malo, E.: Modelo de clustering para el análisis de los datos poblacionales del cáncer registrados en la comuna de Hijuelas, Memoria de Título de Ingeniería en Informática Aplicada, Universidad de Valparaíso (January 2007)
6. Medina, E., Kaempffer, A.: Cancer mortality in Chile: epidemiological considerations. Revista Médica de Chile 129(10), 1195–1202 (2001)

Novel Features for Automated Lung Function Diagnosis in Spontaneously Breathing Infants

Steffen Leonhardt and Vojislav Kecman

Philips Chair of Medical Information Technology, Helmholtz Institute for Biomedical Engineering, RWTH Aachen University, Germany
medit@hia.rwth-aachen.de
The University of Auckland, Auckland, New Zealand
v.kecman@auckland.ac.nz

Abstract. A comparative analysis of 14 classic and 23 novel mathematical features for diagnosing lung function in infants is presented. The data set comprises tidal breathing flow volume loops of 195 spontaneously breathing infants aged 3 to 24 months, with 9 known breathing problems (diseases). The data set is sparse. Diagnostic power was evaluated using support vector machines featuring both polynomial and Gaussian kernels in a rigorous experimental setting (100 runs for random splits of data into the training set (90% of data) and test set (10% of data)). Novel features achieve lower error rates than the old ones.

Keywords: Lung function diagnosis, support vector machines, novel breathing features.

1 Introduction

Tidal breathing flow volume (TBFV) loops are a known technique to evaluate the performance of spontaneous breathing in infants [1] and to detect especially obstructive lung disorders at an early stage. Regarding their discriminative power, several reports on single TBFV features and specific time relationships have been published, but so far the focus has usually been on investigation of single symptoms at a time. In addition, the significance of several of these classic (i.e., old) TBFV loop features seemed questionable as they were found not to be a proper description of clinically observed loop patterns. We therefore aimed at developing more descriptive novel TBFV loop features [2]. Tidal breathing spirometry is done on the sleeping or sedated babies.

Infants aged 3 to 24 months with either suspected obstructive disorders, stenosis or tracheo-broncho-malacias were eligible for the study. In all patients except the normal group, routine diagnosis was confirmed by bronchoscopy serving as a Gold Standard (reference). If available, medical history on allergies or gastroesophageal reflux was considered. To obtain a TBFV loop, flow and corresponding volume have to be plotted against each other (so-called phase plane diagram). For a healthy infant who breathes rather sinusoidally an ellipsoid shape can be expected. As shown in Fig. 1, these loops significantly change shape with specific pathologies.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 195–199, 2007.

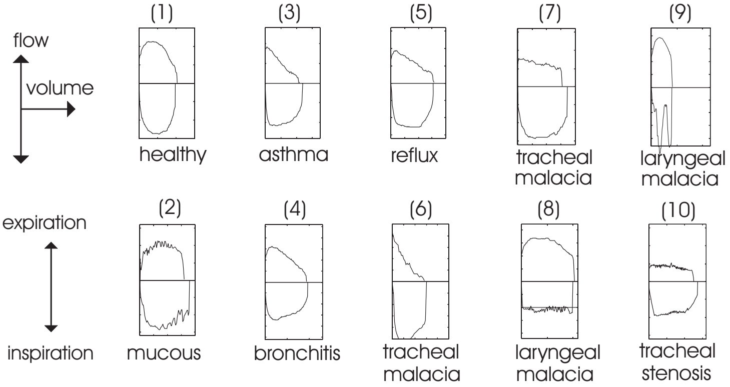


Fig. 1. Typical TBFV loops obtained during tidal breathing spirometry of sedated infants. By definition, inspiratory flow is counted negatively.

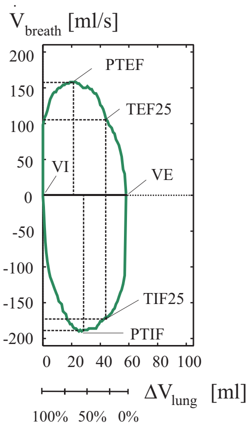
2 Classic and Novel Loop Features

Classic features: In the past [1], the analysis of TBFV loops was performed by observing classic characteristic flow or volume measurements, see Fig. 2. To avoid dependencies on size or body weight, most of these classical features have usually been normalized by relating them to tidal volume or similar variables. In addition, specific timing relationships were measured. A list of a classic features is as follows - PTEF/VE (max. exp. flow related to VE), TEF50/VE (exp. flow at 50% volume related to VE), TEF25/VE (exp. flow at 25% volume related to VE), PTIF/VI (max. insp. flow related to VI), TIF50/VI (insp. flow at 50% volume related to VI), TIF25/VI (insp. flow at 25% volume related to VI), TEF50/TIF50 (exp. flow related to insp. flow, both at 50% volume), TEF50/TEF10 (exp. flow at 50% volume related to exp. flow at 10% volume), TEF25/PTEF (exp. flow at 25% volume related to max. exp. flow), PTEF/PTIF (max. exp. flow related to max. insp. flow), Veto PTEF (volume fraction when reaching max. exp. flow), TE/TI (exp. time related to insp. time), TI/Tt (fraction of insp. time related to total breath time Tt), TetoPTEF (fraction of time to reach PTEF related to total breath time Tt), TitoPTIF (fraction of time to reach PTIF related to total breath time Tt). Details can be seen in [2].

Novel features: One important goal of this paper has been to come up with linguistically complex, yet quantifiable TBFV loop features that extract more information from the loop and, hence, can better distinguish various pathologies from the healthy norm as compared to the classical TBFV loop features. Their detailed description is given in [2] and here we just list few of them:

Sphericity: The roundness of a TBFV loop may be quantified by dividing the radius of the inscribed half-circle by the radius of the circumscribed half-circle. The symbol for roundness shall be R. For expiration, this leads to

$$o_{\exp} = r_{inscribed\,\exp} / r_{circumscribed\,\exp} \tag{1}$$



VE	volume at expiration [ml]
PTEF	maximal flow during expiration [ml/s]
TEF25	expiratory flow at 25 % volume [ml/s]
VI	volume at inspiration [ml]
PTIF	maximal flow during inspiration [ml/s]
TIF25	inspiratory flow at 25 % volume [ml/s]

Fig. 2. Examples of absolute classical features

Triangularity: One possible way to quantify triangularity is to relate the area of the inscribed triangle (connection between the normalized PTEF and the starting and end points of the loop) to the area under the TFBV loop. For expiration (and similarly for inspiration) this leads to equation (2) for

$$\nabla_{exp} = 2\int_{0}^{VE} \bar{V}_{breath}(\Delta V)dV/\left(PTEF * VE\right) \qquad (2)$$

Rectangularity: To quantitatively capture the amount of rectangularity, a definition analogous to the roundness can be employed. In doing so, the TBFV loop has to be normalized and then the inscribed and the circumscribed rectangles have to be computed. However, it quickly becomes evident that there is an additional degree of freedom in how to choose the height h and the width w. This was solved by setting the area $A = w * h$ maximal. Based on this boundary condition, the expiratory width and height relations are given by

$$\Box_{exp} = \frac{w_{exp,inscribed}}{w_{\mathrm{exp},circumscribed}} \frac{h_{exp,inscribed}}{h_{\mathrm{exp},circumscribed}} \qquad (3)$$

Waviness: To quantify collapse phenomena in TBFV loops, one possibility is to approximate the normalized loop with high order polynomials and then analytically compute the location of the different extremes of this approximation. The waviness may then be computed by summing the absolute differences in height between the different extremes. For expiration, this leads to

$$\sim_{\exp} = \sum\nolimits_{i=0}^{n} \left|\bar{V}_{breath}\left(V_{extr,[i+1]}\right)\right| - \bar{V}_{breath}\left(V_{extr,[i]}\right) - 2 \qquad (4)$$

Approximation with Polynomials: Another way to analyze TBFV loops is to approximate the normalized loops by low (first and second) order polynomials given as

$$\hat{\dot{V}}_{breath} = a_1 \cdot \varDelta V_{lung} + b_1 \ , \hat{\dot{V}}_{breath} = a_2 \cdot \left(\varDelta V_{lung}\right)^2 + b_2 \cdot \varDelta V_{lung}. \qquad (5)$$

3 Classification and Results

The lung data described was created and analyzed in [2] where the SPSS tool [3] has been used for all the statistical analysis. Here, the *support vector machines* (SVMs) ([4], [5]) are used and the results are compared to the ones given in [2]. The basic aim of investigations here is to show the ability of SVMs as well as to investigate the efficiency of both the old and new features in diagnosing breathing problems at infants. Here, the nonlinear SVMs have been created by using Gaussian and polynomial kernels. SVMs create a discriminant function $f(\mathbf{x})$ by using a training data set $D = \left\{[\mathbf{x}(i), y(i)] \in \Re^n \times \Re^1,\ i = 1, ..., P\right\}$ consisting of P pairs $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_P, y_P)$ where the inputs $\mathbf{x}_i$ are n-dimensional vectors and the labels (or system responses) y_i are discrete values for classification problems going from 1 (healthy) to 10 (stenosis of major bronchus) here. The final classifier is given as

$$o = f(\mathbf{x}) = \sum\nolimits_{i=1}^{P} v_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{6}$$

where o is the output of SVM (which is the value of the function f for a given input x). v_i are the weights of the expansion, K is the (scalar) value of a given kernel for a given input x, x_i are the training data values and the bias term b is an offset parameter determined during the training. All the results shown here are obtained on the test data pairs. 100 experimental runs have been performed here, with data sets randomly split up into the training sets (90% of data pairs) and the test ones (10% data pairs). Table 1 shows comparisons of performances of the SPSS and SVMs on selecting 'the best' feature by using *old features for SPSS* and both *old* and *new ones for SVM.* The results shown are obtained by classifying disease vs. healthy symptoms based on a single feature. SVM shows superior performance on all diseases in respect to the SPSS discriminating capacities. Also, based on the single best feature only, new features are showing stronger discriminant capacities than the old ones in the case of 6 out of 9 diseases. The bold values of error in % shown in Table 1 indicate the winners, i.e., the smallest error achieved in diagnosing certain disease, and it is easy to see that all the winners belong to the SVMs' models. There are few more interesting outcomes while classifying healthy symptoms vs. disease for all the features (inputs). The error percentages clearly indicate which diseases are easy to diagnose and which are not. Thus, diagnosing Asthma, Mild tracheal malacia and Laryngeal malacia is the easiest task, while Bronchitis and Gastroesophageal reflux diagnoses are much more difficult and prone to higher errors averaging over all 23 new features. Finally, the more data mining oriented readers may be interested in the algorithm and software used for designing the SVM classifier here. Data set having 195 24-dimensional training data pairs falls into category of low-size problem which can be readily solved within the matlab environment by using an active-set algorithm [5]. A routine used has been implemented as C MEX-file for classification by M. Vogt [4], and it has been adapted for solving a multi-class breathing diseases problems by the second author.

Table 1. Minimal error achieved by classifying disease vs. health while implementing SPSS and SVMs and the single best feature selected for the old and new features

Disease Error in %	SPSS, *old* best	SVM, *old* best	SVM, *new* best
Asthma	5.5/$b_{1,\mathrm{exp}}$	4.0/TETI	**3.2/**$c_{2,\mathrm{exp}}$
Bronhitis	19.7/$c_{2,\mathrm{exp}}$	15.1/Teto-PTEF	**14.3/**$c_{2,\mathrm{exp}}$
Gastroesophageal reflux	17.0/$c_{2,\mathrm{exp}}$	17.3/TEF25-PTEF	**12.9/**$c_{2,\mathrm{exp}}$
Mild tracheal malacia	12.1/$b_{1,\mathrm{exp}}$	**6.0/TETI**	6.8/∇_{exp}
Severe tracheal malacia	16.4/$\sim_{exp}$	15.3/Teto-PTEF	**9.7/**$\sim_{exp}$
Laryngeal malacia	5.0/o_{insp}	5.6/Tito-PTIF	**2.4/**o_{insp}
Tracheal stenosis	7.5/o_{exp}	**4.0/TEF50-VE**	6.0/o_{exp}
Laryngeal stenosis	17.3/$c_{2,\mathrm{exp}}$	12.0/TETI	**9.2/**$b_{1,\mathrm{exp}}$
Stenosis of major bronchus	21.7/$c_{2,\mathrm{exp}}$	**11.6/TETI**	12.0/$\Box_{exp}$

4 Conclusions

The paper shows an improvement of diagnosis accuracy for a breathing problems at infants by using a SVM. It has been shown that the SVM classifier performs better than a Bayes classifier implemented in SPSS. It is also pointed at an increased efficiency by using new features proposed in [2]. The SVM was trained by using an active set method for solving SVMs' QP based learning problem. The results shown are the best known to date for diagnosing several lung diseases at infants. The algorithm is able to point to difficult diagnosing tasks and as such it can also be used for a training of younger specialists.

References

1. Operators Manual: 2600 Pediatric Pulmonary Function Laboratory, SensorMedics Corporation, Yorba Linda, CA 92687, USA (1991)
2. Leonhardt, S.: Automatisierte Lungenfunktionsdiagnose bei spontanatmenden Säuglingen (Automated Lung Function Diagnosis in Spontaneously Breathing Infants, in German), Shaker Verlag, Aachen, Germany (2001)
3. Buehl, A., Zoefel, P.: SPSS fuer Windows, Version 6.1, 2. Auflage. Addison-Wesley Publishing Company, Bonn (1995)
4. Kecman, V.: Learning and Soft Computing, Support Vector Machines, Neural Networks, and Fuzzy Logic Models, The MIT Press, Cambridge, MA (2001) http://www.support-vector.ws
5. Vogt, M., Kecman, V.: Chapter Active-Set Methods for Support Vector Machines. In: L. Wang (ed.) a Springer-Verlag book, Support Vector Machines: Theory and Applications, Series: Studies in Fuzziness and Soft Computing, vol. 177, pp. 133–158 (2005)

Multi-level Clustering in Sarcoidosis: A Preliminary Study

V.L.J. Karthaus[1], H.H.L.M. Donkers[2], J.C. Grutters[1], H.J. van den Herik[2], and J.M.M. van den Bosch[1]

[1] St. Antonius Hospital, P.O. Box 2500, 3430 EM Nieuwegein
[2] MICC / IKAT, Universiteit Maastricht, P.O. Box 616, 6200 MD Maastricht
v.karthaus@antonius.net

Abstract. Sarcoidosis is a multisystem disorder that is characterized by the formation of granulomas in certain organs of the body. The exact cause of sarcoidosis is unknown but evidence exists that sarcoidosis results from exposure of genetically susceptible hosts to specific environmental agents. The wide degree of clinical heterogeneity might indicate that sarcoidosis is not a single polymorphic disease but a collection of genetically complex diseases. As a first step to identify the hypothesized subcategories, large amounts of multidimensional data are collected that are divided into distinct levels. We investigated how clustering techniques can be applied to support the interpretation of sarcoidosis and subsequently to reveal categories of sarcoidosis data. An attempt is made to relate multiple clusters between the different data levels based on validation criteria.

Keywords: Sarcoidosis, Clustering.

1 Introduction

Sarcoidosis is a multisystem disorder that is found throughout the world within almost all races and ages [1]. The disease is distinctly characterized by the formation of non-caseating granulomas in certain organs of the body. Most frequently, the lungs or the lymph nodes are affected but the inflammation can occur in almost any part of the body. The range and severity of symptoms associated with sarcoidosis vary significantly, depending on the specific organs involved and the degree of the involvement. Prognosis of sarcoidosis may be difficult to establish because of its unpredictable course.

The exact cause of sarcoidosis is not known. A common theory suggests that sarcoidosis is likely to be triggered by a combination of environmental and genetic factors. Moreover, the heterogeneous nature of sarcoidosis lead us to believe that there is more than one cause, each leading to possibly different manifestations of the disease [2]. From this wide degree of heterogeneity we may hypothesize that sarcoidosis is not a single disease but a collection of genetically complex diseases. Identification and classification of categories of sarcoidosis may provide researchers with more homogenous groups of sarcoidosis patients. These groups can significantly contribute to investigating the cause as well as to a better treatment and prognosis of

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 200–204, 2007.

the individual patients. The application of machine-learning techniques and, in particular, unsupervised clustering algorithms may help to reveal categories of sarcoidosis patients. In this paper, we will investigate how clustering techniques can be applied to support the (biological) interpretation of sarcoidosis and subsequently to reveal categories of sarcoidosis data.

In the St. Antonius Hospital, a large database disposed for specific categories of lung diseases including sarcoidosis is being built. To organize this high-multidimensional data and to facilitate translational research, the data sets are divided into four levels based on the distinction between genetic, protein, cellular, and clinical data (see Fig. 1). Each level contains different amounts and types of data originating from possibly overlapping patient groups.

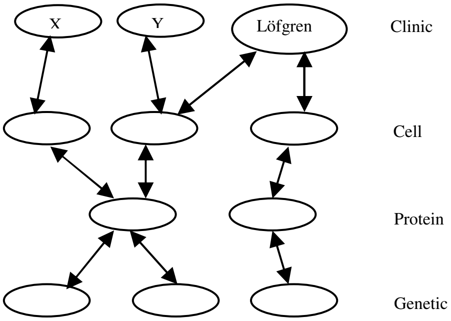


Fig. 1. Different clusterings at different levels and possible interactions

At present, a total number of 861 sarcoidosis patients are stored in the database. The clinical data (with the exception of mainly the majority of lung function, laboratory and data related to the radiological stage) of 319 sarcoidosis patients in total are covered. Moreover, the cellular data of 745 sarcoidosis patients are collected. Of 247 patients both clinical and cellular data are available.

The data analysis is expected to reveal different clusterings at different levels. We hypothesize that sarcoidosis consists of a collection of separate diseases. However, two diseases might be indistinguishable at one level and at the same time dissimilar at another level. For instance, one generally accepted constellation of sarcoidosis symptoms is Löfgren's syndrome, which is a form of acute sarcoidosis. It occurs for a short period with the specific combined clinical symptoms of erythema nodosum and joint problems. In addition, it shows a distinct genotype different from other sarcoidosis types. At the cellular level we see a pattern characterized by elevated levels of lymphocytes, which is also seen in other individuals with active sarcoidosis. At present, the existence and number of clusters at each level are uncertain. The key problem is therefore to discover and describe distinct clusters at each level and their relation with clusters at other levels.

The paper is organized as follows. Section 2 presents the clustering approach. Section 3 shows the preliminary experimental results. An interpretation together with a tentative conclusion is provided in Section 4. The conclusion and directions for future research are mentioned in Section 5.

2 Clustering Approach

A fundamental issue in the effort to detect clusters is how to measure the (dis)similarity between patients. Many distance measures exist and to a large extent the choice depends on the types of variables present. We used Gower's similarity coefficient [3] because it has two typical features. First, it calculates the distance between two patients by taking into account mixed-mode data. This way of measuring also allows exclusion of negative matches of particular attributes. A second typical feature of Gower's similarity measure is its implicit strategy for dealing with missing values. If an individual has a missing value for a particular attribute, the similarity with another patient is calculated based on the remainder of the attributes.

It is difficult to assess which clustering method is most appropriate in a particular situation. The goal of a clustering algorithm is to expose the underlying structure by discovering the 'natural' grouping in the data. One of the fundamental difficulties is, however, to capture this intuitive description by an explicit definition [4]. In literature, numerous clustering techniques have been proposed [5, 6]. Discriminating criteria are, e.g., the algorithm, the nature of clusters: crisp or fuzzy, the type of data used, and the clustering criterion that governs the clusters shape. We used the following three algorithms: Hierarchical Agglomerative Nesting (AGNES, five linkage methods were considered), Hierarchical Divisive Analysis (DIANA), and Partitioning Around Medoids (PAM) [7].

Besides the selection of the appropriate cluster method, two issues arise when applying clustering algorithms [8]: (1) the result of a clustering algorithm does not tell whether its structure is actually present in the data (2) the algorithms tend to bias towards partitions consistent with their own clustering criterion such as spherical or elliptical structures. As a consequence, these algorithms find clusters of that type, regardless whether they are present or not. Therefore we would like to know whether the groups obtained are the result of the clustering method imposing a pattern rather than discovering something that is actually present. To this end, we used cluster-validation techniques, especially since there is no a-priori information about the real number of clusters. In Halkidi et al. an extensive review of validation methods is given [9]. Distinction is made between external and internal cluster criteria that fundamentally differ in their nature. External criteria evaluate the clustering against an existing structure, for example a known set of class labels. In contrast, internal criteria assess the quality of a cluster based on the intrinsic information of the data themselves. Measures in this context are based on, e.g., compactness, connectness, separation, and combinations of these. Examples of the latter type are the silhouette width [10], modified Hubert Γ statistic [9], and the Dunn index [11].

In order to validate and compare clusterings at different levels, we used a two-stage validation process. First, we estimated for a certain cluster algorithm the optimal partition at each level using the suggested internal criteria. Next, we used external criteria to make a comparison between the clusterings on different levels. To start with, we utilized the adjusted Rand Index as the external criterion [12]. This index measures the similarity between two clusters by looking at pairs of points and how they are partitioned in these clusters.

3 Preliminary Experimental Results

We analyzed the data of 247 patients of whom both clinical and cellular data are available. In Fig. 2, Dunn's index and the silhouette width for the cellular and clinical data are depicted. Kaufman and Rousseeuw [7] describe a subjective interpretation: a silhouette width greater than 0.5 should be interpreted as a reasonable to strong partition whereas a value less than 0.25 indicates cluster absence. According to this assertion, the data exhibit a weak clustering structure (cellular) to barely any structure (clinical). Furthermore, considering the silhouette widths only, the figure seems to indicate that a three-cluster partition is the best solution produced by most cluster algorithms, while a singe link performed worse than all other methods. The results are confirmed by Dunn's index. Given the poor cluster results at each level, we expect that the adjusted Rand Index hardly shows similarity between the structures on both levels. This expectation is in accordance with the calculated figures. That is, the adjusted Rand indices do not indicate significantly better agreement between both levels than what is expected by chance alone [13].

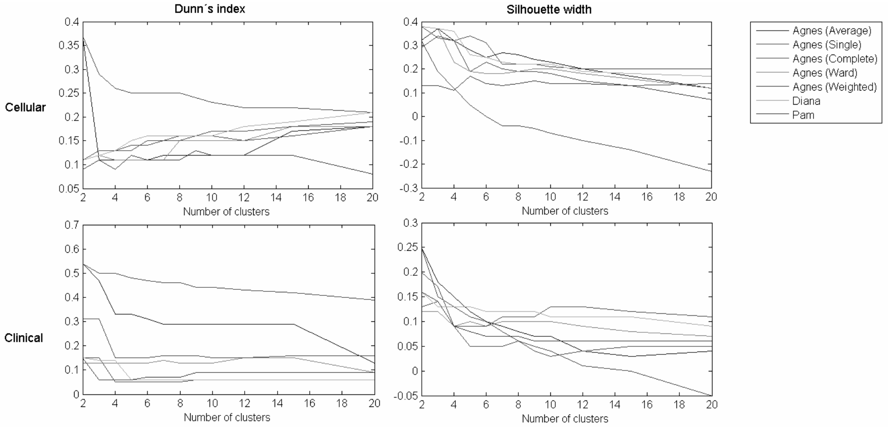


Fig. 2. Dunn's index and silhouette width for cellular and clinical data

4 Interpretation and a Tentative Conclusion

The fact that we have only slightly better results than random agreement between both levels was to a certain extent to be expected. We offer three possible explanations for this observation. The first one is that the current amount of patients is not sufficient to find clear structures, in particular when compared to the variety and number of existing attributes. The data collection is, however, an ongoing process and we anticipate more patients to be included in the study. Besides, several important clinical characteristics, in particular radiological findings and lung function and other laboratory tests, were not available at the time of the experiments. We foresee a considerable improvement because of the inherent clinical quality of these findings.

The second possible explanation is that there is no or only little underlying structure at each level and we can only discover clusters at the combined cellular and clinical level. This would be opposed to current belief within translational sarcoidosis research, but more analyses are required to examine this possibility after inclusion of additional data. Finally, the third possible explanation is that the cluster algorithms and distance measure used are not fully applicable for this specific task and data.

5 Conclusion and Future Directions

The outcomes so far were statistically not significant, but there is indication that further research will provide us with improved results. Therefore, we tentatively may conclude that unsupervised learning is a viable way to pursue research in this direction. We will continue this study by including more patients and attributes at all levels. The validation will be augmented by including the expert to assess the outcome and the relation between the clusters. Finally, the comparison of clusters will be approached by two novel methods: (1) the use of belief networks and (2) meta clustering with the clusters found as input.

References

1. Demedts, M., Wells, A.U., Anto, J.M., Costabel, U., Hubbard, R., Cullinan, P., Slabbynck, H., Rizzato, G., Poletti, V., Verbeken, E.K., Thomeer, M.J., Kokkarinen, J., Dalphin, J.C., Taylor, A.N.: Interstitial lung diseases: an epidemiological overview. Respir J Suppl 32, 2s–16s (2001)
2. Newman, L.S., Rose, C.S., Maier, L.A.: Sarcoidosis. N Engl J Med 336, 1224–1234 (1997)
3. Gower, J.C.: A general coefficient of similarity and some of its properties. Biometrics 23, 623–628 (1971)
4. Estivill-Castro, V.: Why so many clustering algorithms: A position paper. SIGKDD Explorations 4, 65–75 (2002)
5. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys 31, 264–323 (1999)
6. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley-Interscience, New York (2000)
7. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley-Interscience, New York (1990)
8. Handl, J., Knowles, J., Kell, D.B.: Computational cluster validation in post-genomic data analysis. Bioinformatics 21, 3201–3212 (2005)
9. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems 17, 107–145 (2001)
10. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65 (1987)
11. Dunn, J.: Well separated clusters and optimal fuzzy partitions. J. Cybernetics 4, 95–104 (1974)
12. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification 2, 193–218 (1985)
13. Steinley, D.: Properties of the Hubert-Arabie Adjusted Rand Index. Psychological Methods 9, 386–396 (2004)

Part IV

Text Mining, Natural Language Processing and Generation

An Experiment in Automatic Classification of Pathological Reports

Janneke van der Zwaan, Erik Tjong Kim Sang, and Maarten de Rijke

ISLA, University of Amsterdam, Amsterdam, The Netherlands
{jvdzwaan,erikt,mdr}@science.uva.nl

Abstract. Medical reports are predominantly written in natural language; as such they are not computer-accessible. A common way to make medical narrative accessible to automated systems is by assigning 'computer-understandable' keywords from a controlled vocabulary. Experts usually perform this task by hand. In this paper, we investigate methods to support or automate this type of medical classification. We report on experiments using the PALGA data set, a collection of 14 million pathological reports, each of which has been classified by a domain expert. We describe methods for automatically categorizing the documents in this data set in an accurate way. In order to evaluate the proposed automatic classification approaches, we compare their output with that of two additional human annotators. While the automatic system performs well in comparison with humans, the inconsistencies within the annotated data constrain the maximum attainable performance.

1 Introduction

Increasing amounts of medical data are stored in electronic form. Medical data contains a lot of information that can be used for many different purposes, such as decision support, epidemiological research, quality control, etc. Medical reports are predominantly written in natural language, and as such they are not computer-accessible. Currently, a common way to make medical narrative accessible to automated systems is by assigning 'computer-understandable' keywords. Experts usually perform this task by hand. In this paper, we investigate methods to support or automate medical classification.

The PALGA foundation is the Dutch national network and registry of histo- and cytopathology (in Dutch: Pathologisch Anatomisch Landelijk Geautomatiseerd Archief, http://www.palga.nl). Since 1971, the PALGA foundation has been maintaining a database with abstracts of all histo- and cytopathological examinations that take place in Dutch hospitals. For every examination, a report is written and the conclusion of the report is sent to the PALGA database. The database can be used to find information about the history of a single patient, but it is also available for research and national health care projects.

A key part of the reports in the PALGA database is the set of diagnosis lines which contain a standardized summary of the report. Diagnosis lines are used for indexing and retrieval of the conclusions. They contain a limited number of

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 207–216, 2007.

fields (four) with terms from a restricted vocabulary. The diagnosis lines have been created over a long period of time by a large group of doctors and there is some uncertainty about the consistency and the quality of their contents. The size of the database does not permit a thorough and complete manual quality check.

In this paper we investigate the potential of machine learning approaches for automatically generating diagnosis lines from summaries of pathological reports. Similar tasks have, of course, been considered before in the literature; see e.g., [1] for a (somewhat dated) overview. Three things set our setting apart from settings reported on in the literature. First, in our case, the task is to generate diagnosis lines from conclusion texts—*summaries* of reports and often very incomplete. Second, we are working with a very large data set (over 14 million records, out of which we use close to .5 million for our experiments; see below for details), while most studies in the literature are based on far smaller data sets. Third, because of its size, its usage across hospitals all over the Netherlands, and its age, the data set is full of errors and inconsistencies, unlike most data sets used for medical coding in the literature. Against this challenging background, we are interested in finding answers to two research questions. First, what level of accuracy can automatic classification obtain for this task? Specifically, what type of feature representation is most effective? Second, and motivated by the bottlenecks that we ran into while trying to increase the recall scores of automatic classifiers, what performance levels do humans attain when given the exact same task as the automatic classification system?

This paper contains five sections. After this introduction, we describe the classification problem in more detail, outline our learning approach and discuss related work. In Section 3, we present our experiments and their results. Section 4 provides an elaborate comparison between the best automatic learner and two domain experts. We conclude in Section 5.

2 Method

In this section, we describe our data, the machine learning method that was applied to the data, the techniques used for evaluation, as well as related work.

2.1 The Palga Data

The PALGA data set consists of over 14 million reports. It contains all histological examinations that were performed in the Netherlands from 1990 up to and including 2004. A sample record is shown in Figure 1. The reports contain three parts: main, conclusion and diagnosis lines. The terms in the diagnosis lines are restricted to a set of 14,000 terms, each of which is represented by a code. The coding system used by the PALGA foundation is based on an early (1982) version of the Systemized Nomenclature Of MEDicine (SNOMED). Different types of terms exist; in SNOMED these are called *axes*. The first character of a code represents the axis to which the code belongs. For instance, the term *biopt* is

Record ID	39319785
Patient ID	PATIENT-23 m
Date	07 - 1990
Conclusion	Huidexcisie para-orbitaal links: basaalcelcarcinoom van het solide type. Tumorcellen reiken tot in de excisieranden.
Diagnosis line	huid * gelaat * links * excisie * basaalcelcarcinoom

Fig. 1. Example record from the PALGA data set. Pathologists use everyday terms to code diagnoses. These terms are linked to codes in the PALGA coding system.

linked to code P11400, where P indicates a Procedure. The PALGA coding system consists of the axes Topography, Procedure, Morphology, Etiology, Function and Disease. The coding system is ordered hierarchically.

Every code is linked to one or more terms. A thesaurus is available for enabling pathologists to use everyday language rather than codes in the report writing process. Codes are linked to at most one *preferred term*. Terms linked to identical codes are synonyms. For instance, both *colon* (preferred term) and *dikke darm* are linked to code T67000.

Diagnosis lines, the computer-readable summaries of the contents of pathological reports, consist of PALGA terms only. The lines are used for indexing and retrieving PALGA reports. The quality of the retrieval results is obviously dependent on the quality of the diagnosis lines. Detailed guidelines exist for assuring that the lines are accurate. Among others, these guidelines state that diagnosis lines should be complete and that the report conclusion should contain all relevant information for a pathologist to assign correct diagnostic terms [13]. The diagnosis fields contain three compulsory diagnostic fields: the two axes *Topography* and *Procedure*, and *Diagnosis*, which may contain terms of the other four axes. Conclusions are coded with one to four diagnosis lines.

Diagnosis lines in the PALGA database contain a lot of noise. Creating accurate and precise diagnosis lines is not the main task of a pathologist and recommendations of the type-checking tools which became available in recent years, are often ignored. We are interested in working with a data set which was as clean as possible, and therefore we have restrict ourselves to reports containing a conclusion of at least six characters and a single diagnosis line with valid singular terms. Additionally, we restrict the reports to those that contain the term *colon* or one of its descendants. This results in a data set of 477,734 conclusions with associated diagnosis lines, hereafter called the *Colon data set*. Diagnosis lines in the Colon data set contain 3.4 terms on average. The data set was randomly divided into 75% of training data and 25% of test data.

2.2 Support Vector Machines

We decided to use Support Vector Machines (SVMs) for our medical text classification task, as they belong to the best performing learning algorithms currently available [7]. Fast implementations of SVMs exist; we used SVM^{light} [8] (with default settings) for our experiments.

One of the strengths of SVMs is that the standard linear kernel can be replaced by a non-linear one, e.g., a polynomial or radial basic function (RBF). Most of our experiments are conducted with linear kernels, but we also experiment with polynomial ones.

Another property of SVMs is that a classifier is trained independently of the number of features of the data samples. This is particularly useful for text classification, where the dimensionality of the feature space generally is high with few relevant features [7]. Most other machine learning approaches to text classification apply some sort of feature reduction or transformation. By using SVMs there is no need to reduce the number of features.

2.3 Evaluation

We are interested in classifier performance for individual terms as well as prediction accuracy for complete diagnosis lines. The notion of equivalence of diagnosis lines is problematic, because the coding system allows something being said in different ways. Among other things, the level of detail can differ; if a finding is specified using a (slightly) more general (or specific) term, it is not necessarily wrong. Still, we decided to use a simple notion of 'exact' equivalence to assess the agreement between diagnosis lines. Thus, our evaluation results will underestimate the actual performance. For individual term prediction we treat diagnosis lines as a bags of terms and perform evaluation with precision, recall and $F_{\beta=1}$.

2.4 Related Work

Medical coding is the task of assigning one or more keywords to medical text. Reasons to code medical documents include data reduction, standardization, quality control, being able to compare individual cases, and making data available for research. In general, medical coding is considered a difficult and time-consuming task [15,4,9,19]. Ever since the introduction of formal coding systems, attempts have been made to automate the coding process [16,20,11]. In [1] a distinction is being made between coding systems that abstract clinical data (such as ICD and MeSH) and those that preserve clinical details. [14] lists a number of information types that medical coding systems can (or should be able to) represent. In [5] manual coding errors in two British hospitals were investigated and compared. It appeared that 'many of the errors seem to be due to laziness in coding, with failure to consult the appropriate manual and reliance on memory for common codes.' A recent study indicates that data quality does indeed improve after the adoption of automatic encoding systems [10].

The PALGA foundation has been involved in an earlier research project regarding medical text classification [2]. In this particular project, complete pathological reports were used to predict appropriate diagnosis lines; 7500 histology reports from two different hospitals were considered and three different document representations were compared. Using the best performing representation—uniform words—a correct diagnosis line could be found within the first five suggestions for 844 of 952 reports. Other experiments showed that a representation based on words performed better than (character) n-grams with $n > 4$, and that performing training and testing with data from the same site allowed for a better performance than when test data came from another site than the training data. In an additional evaluation, three human experts rating the automatically produced diagnosis lines on a three-point scale, reached an agreement kappa score of only 0.44, which shows that coding pathological reports is not a trivial task.

Recently, Gerard Burger, a pathologist associated with the PALGA foundation, created a term extractor for PALGA conclusions. The program, called AutoDiag, uses domain knowledge and ad hoc rules to propose terms for diagnosis lines associated to the input document. AutoDiag extracts terms from the conclusion part of the documents, keeping terms it considers useful while ignoring other terms. Prior to this paper, AutoDiag had not been properly evaluated. We use it in our work and compare its output with that of our system.

3 Experiments and Results

We treat the task of coding of diagnoses as a text classification task and train binary Support Vector Machines to predict individual diagnostic terms. Below, we describe the experiments that were performed. First, we discuss experiments with different feature representations. After that, we evaluate other variations, changing output class representations, machine learning parameters or data sets. We conclude the section with a discussion.

3.1 Feature Engineering

To create a baseline, we chose SVMs with bag of words (bow) for the feature representation; this is a common representation for text classification. In the bow representation, a document is represented as a feature vector, where each element in the vector indicates the presence or absence of a word in the document. In order to reduce the size of the vector, we discard the least infrequent words (frequency < 2) as well as the most frequent words (so-called stop words; we use a list from the Snowball Porter stemmer for Dutch [18]). The dimensionality of the feature vectors for the *bow* experiment is 21,437. The bow features allowed for a reasonable performance on the Colon test data (section 2.1): precision 83.28%, recall 72.92% and $F_{\beta=1}$ 77.76%.

Since in the baseline results, recall was considerably lower than precision, we focused on improving recall. We evaluated several variations on the bow feature representation to accomplish this:

- replacing binary feature values by *tf-idf* (term frequency-inverse document frequency) weights
- replacing word unigrams by word bigrams (19,641 features)
- adding to the features, terms identified from the conclusion texts (1,286 extra features), and/or their parents according to the thesaurus (1,810)

Table 1. Influence of different feature representations on term identification; highest scores in boldface

	Representation	**Precis.**	**Recall**	$\mathbf{F}_{\beta=1}$		**Represent.**	**Precis.**	**Recall**	$\mathbf{F}_{\beta=1}$
a	baseline	83.28%	72.92%	77.76%	j	tf-idf	83.14%	73.71%	78.14%
b	+terms	83.49%	73.16%	77.98%	k	+terms	83.24%	73.94%	78.31%
c	+terms+parents	83.53%	73.22%	78.04%	l	+te+parent	83.35%	74.09%	78.45%
d	terms	79.92%	60.91%	69.13%	m	+te+pa+pr	83.28%	73.69%	78.19%
e	+parents	80.12%	66.41%	72.62%	n	+prep	83.10%	73.26%	77.87%
f	+parents+prep	79.80%	67.92%	73.39%	o	bigrams	84.68%	74.83%	79.45%
g	+prep	80.55%	62.05%	70.10%	p	+terms	**84.85%**	**75.56%**	**79.94%**
h	stems	83.23%	72.71%	77.61%	q	+te+parent	84.83%	75.40%	79.84%
i	split compounds	83.18%	72.37%	77.40%	r	+te+prep	84.84%	75.58%	79.94%
					s	+prep	84.70%	74.88%	79.49%

- reducing the number of features by using stems rather than words (19,098 features) or by splitting compound words (19,006 features)
- preprocessing the input text with Gerard Burger's AutoDiag rule-based term-identification tool (section 2.4)

A summary of the results of the experiments can be found in Table 1. Adding term features and parent features to the baseline set, led to small but significant improvements of both precision and recall (a-c). Replacing the baseline features with term features, had a negative influence on performance (d-g). The stem and the split compound features proved to be worse than the baseline set (h-i). Replacing binary weights by tf-idf weights, resulted in significantly better recall scores (j-n). All experiments with bigram features (o-s) reached significantly better precision and recall scores than the baseline. Bigram features with additional term features (p) reached the highest precision (84.85%) and recall (75.56%) scores. The experiments were inconclusive with respect to preprocessing the input texts. In the terms group, the effect was positive (f-g). With bigrams, scores decreased (r-s) and with tf-idf, performance did not change (m-n).

3.2 Changing Learning Parameters and Output Classes

We performed three alternative experiments to see if an additional performance gain could be obtained. First we, evaluated the influence of an important parameter of the machine learning algorithm: the kernel type. In the previous experiments we used a linear kernel. For the next experiment we tested using a polynomial kernel with three different degrees: 1, 2 and 3. With the baseline feature representation, bag of words, the best results were obtained with degree value 2: precision 84.89% and recall 76.11%, both of which outperform the results of the previous section. However, the performance gain came with a price: the polynomial kernels take much more time to train than the linear ones.

In the next two experiments we attempted to take advantage of the assumption that the terms appearing in diagnosis lines are dependent. First, we trained

SVMs to predict bigrams of terms rather than unigrams. However, for both feature representations that we tested, the baseline set and bigram features plus terms, the recall scores decreased significantly when predicting term bigrams. Next, we evaluated a cascaded learner: first train SVMs to perform the classification task with baseline features and then train a second learner with additional features from the output of the first system. The results were similar to the previous experiment: improved precision scores (85.23%) but lower recall (72.40%).

3.3 Changing Data Sets

Additional experiments were performed to determine whether training and testing with data from different time-periods affects performance. The data was divided into three periods of five years (1990–1994, 1995–1999, 2000–2004). From each period 75,000 records were available for training and 24,995 for testing. Best results were obtained with training and test sets from the same time-periods (bag of words: $F_{\beta=1}$ 77.98%, 78.61% and 77.22% respectively). Performance was significantly worse for experiments with training and test sets from different time-periods (on average, 75.41%). When training and test sets were ten years apart, performance was even lower, 74.09%.

3.4 Discussion

The experiments revealed that compared to precision, the recall scores are rather low. I.e., if a classifier assigns a term to a conclusion, it is probably correct, but many positive instances are missed. Despite several attempts to increase recall, it was mainly precision that went up and recall remained relatively low.

Several reasons can be given for explaining why recall scores are lower than precision scores. First, many terms in the diagnosis terms are infrequent and it is hard to train classifiers for classes with a small amount of positive samples. Second, the information needed in a diagnosis line case might not always be available in the corresponding conclusion or that information might be lost in the conversion to features. And third, low recall scores might be caused by the incorrect or inconsistently tagged data.

So there are different possible causes for low recall. But can we expect to attain higher recall scores, or is the problem simply very hard? How well do experts perform if they only have access to the conclusions (instead of the complete report) for coding purposes? Do experts consistently assign the same codes to conclusions? These matters will be investigated in the next section.

4 A Comparison with Domain Experts

In this section, we compare two of the annotation approaches discussed in the previous section with human expert annotators. Based on earlier work, we created a balanced corpus with 1000 texts of which 35% were records for which the baseline obtained a high score, 19% were records with a low score while the

Table 2. Precision, recall, and $F_{\beta=1}$ of new expert ratings compared to the diagnosis lines in the corpus and Kappa agreement scores between the automatic systems, the humans and the corpus, where "P A" ("P B") stands for "Pathologist A" ("Pathologist B"). Scores have been averaged over terms suggested by raters in the first column.

				Kappa scores		
	Precision	**Recall**	$\mathbf{F}_{\beta=1}$	Corpus	P A	P B
bow	83.62%	72.87%	77.87%	0.65	0.58	0.61
bigrams+terms	84.88%	75.47%	79.90%	0.80	0.52	0.56
Pathologist A	71.75%	72.54%	72.14%	0.44		0.65
Pathologist B	66.75%	67.33%	67.04%	0.42	0.55	
AutoDiag	53.10%	62.19%	57.28%	0.22	0.31	0.46

remaining 46% had a medium classification score of the baseline system (details can be found in [21]). Next, two experts were invited to re-annotate the texts based on only the conclusion part. Even though each text had already been coded by experts, it is not obvious that their ratings are correct (or optimal) or that conclusions contain sufficient information for coding. Comparing the new expert ratings to the corpus will enable us to identify differences in term assignments.

We created a web interface which presented a conclusion text to the annotator together with terms predicted by the baseline system and terms that were extracted from the conclusion with a basic term extractor. Terms were grouped into the three main parts of the diagnosis lines: Topography, Procedure and Diagnosis. The two annotators from different hospitals had the opportunity to suggest alternative terms when they regarded the suggested terms as incomplete. Each of the two pathologists took over four and a half hours to complete this task (about sixteen seconds per conclusion text).

The diagnosis lines created by the two experts were compared with the lines in the corpus. Table 2 lists the results as well as the scores of the baseline system, the best bigram system and the rule-based term extractor AutoDiag. The classifiers proved to be better at reproducing the corpus' term assignments than the experts. Another aspect worth noting is that our human annotators score better on recall than on precision, suggesting that the classification task is inherently hard (and not "just" a recall problem).

These are our explanations for the differences between humans and systems:

- some (complex) terms consist of multiple simple terms, and replacing one by the other results in an error
- often when there is a mismatch between two terms, one is just higher or lower in the same hierarchy
- human annotations proved to be more elaborate than system annotations
- humans also had a larger number of conclusion texts with multiple diagnosis lines (10% versus 0)
- while systems always assign terms to conclusion texts, humans frequently assigned the term *unknown* (18% compared with 1% in the corpus)

As an aside, in our evaluation we also included AutoDiag, the rule-based term extractor mentioned before. With an F-score of 57.28% it performed worse than all other methods we considered.

In general, large differences exist between the diagnosis lines in the corpus and the new expert ratings. Amongst themselves the experts also disagree about the terms that should be assigned to conclusions. So, again, the task of assigning diagnostic terms to PALGA conclusions is hard, and it is not just recall that is a problem. At higher levels in the hierarchy of terms, agreement seems to be much better. These results (on the PALGA data set) confirm findings of earlier studies investigating the reliability of coded diagnoses [12,3], and more general work on the selection of search terms [17,6].

5 Concluding Remarks

We have examined the potential of machine learning approaches for automatically generating diagnosis lines from summaries of pathological reports. We found that automatic systems perform well in predicting individual diagnosis line terms from text conclusions (precision 85% and recall 75%). However, it proved to be difficult to attain performance levels that were distinctively higher than the baseline scores (83% and 73%).

In a follow-up study, we found that, when restricting access to only the conclusion part of the texts, human experts perform worse than the automatic systems when tested on their ability to reproduce the exact diagnosis lines of the evaluation corpus. This is partly caused by conclusions being incomplete. However, there was also a lack of agreement between the two expert annotators, for example on term specificity. Assigning diagnosis lines to text conclusions proves to be a difficult task.

We conclude that machine learning approaches can achieve good performances in predicting diagnosis lines. By selecting pairs of text conclusions and diagnosis lines for which they perform less well, they can be applied for spotting mismatches between such pairs. Using the predicted diagnosis lines of the systems without an additional manual check would be less appropriate given the machine learner's inability to identify incomplete conclusions. As to supporting the coding task of pathologists, we expect the best results from systems trained on documents of individual doctors, as personal coding assistants.

Acknowledgements

We are very grateful to Ton Tiebosch, Gerard Burger, Arjen van de Pol, and Loes van Velthuysen for data, feedback, and the research problem.

This research was supported by various grants from the Netherlands Organisation for Scientific Research (NWO). Erik Tjong Kim Sang was supported under project number 264.70.050. Maarten de Rijke was supported by NWO under project numbers 017.001.190, 220.80.001, 264.70.050, 354.20.005, 600.065.120, 612.13.001, 612.000.106, 612.066.302, 612.069.006, 640.001.501, and 640.002.501.

References

1. Cimino, J.J.: Review paper: coding systems in health care. Methods Inf Med 35(4-5), 273–284 (1996)
2. de Bruijn, B.: Automatic Classification of Pathology Reports. PhD thesis, Maastricht University (1997)
3. Dixon, J., Sanderson, C., Elliott, P., Walls, P., Jones, J., Petticrew, M.: Assessment of the reproducibility of clinical coding in routinely collected hospital activity data: a study in two hospitals. J Public Health Med 20(1), 63 (1998)
4. Franz, P., Zaiss, A., Schulz, S., Hahn, U., Klar, R.: Automated coding of diagnoses three methods compared. In: Proc AMIA Symp, pp. 250–254 (2000)
5. Hall, P.A., Lemoine, N.R.: Comparison of manual data coding errors in two hospitals. J Clin Pathol 39(6), 622–626 (1986)
6. Iivonen, M.: Consistency in the selection of search concepts and search terms. Information Processing and Management 31, 173–190 (1995)
7. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Proceedings of the European Conference on Machine Learning, Springer, Heidelberg (1998)
8. Joachims, T.: Making large-scale SVM Learning Practical. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning, MIT Press, Cambridge (1999)
9. Letrilliart, L., Viboud, C., Boëlle, P.Y., Flahault, A.: Automatic coding of reasons for hospital referral from general medicine free-text reports. In: Proc AMIA Symp, pp. 487–491 (2000)
10. Lorence, D.P., Jameson, R.: Managers reports of automated coding system adoption and effects on data quality. Methods Inf Med 42(3), 236–242 (2003)
11. Moskovitch, R., Cohen-Kashi, S., Dror, U., Levy, I., Maimon, A., Shahar, Y.: Multiple hierarchical classification of free-text clinical guidelines. Artificial Intelligence in Medicine 37(3), 177–190 (2006)
12. Nilsson, G., Petersson, H., Ahlfeldt, H., Strender, L.E.: Evaluation of three Swedish ICD-10 primary care versions: reliability and ease of use in diagnostic coding. Methods Inf in Med 39(4-5), 325–331 (2000)
13. PALGA. Thesaurus coderen (PALGA) (2005)
14. Rector, A.L.: Clinical terminology: why is it so hard? Methods Inf Med 38(4- 5), 239–252 (1999)
15. Ribeiro-Neto, B., Laender, A.H.F., de Lima, L.R.S.: An experimental study in automatically categorizing medical documents. Journal of the American Society for Information Science and Technology 52(5), 391–401 (2001)
16. Sager, N., Friedman, C., Margaret, S.: Medical language processing: computer management of narrative data. Addison-Wesley, Reading (1987)
17. Saracevic, T., Kantor, P.B.: A study of information seeking and retrieving. III. searchers, searches, overlap. Journal of the American Society for Information Science and Technology 39, 197–216 (1988)
18. Snowball. Porter stemmer for Dutch http://www.snowball.tartarus.org/.
19. Surján, G.: Questions on validity of International Classification of Diseases-coded diagnoses. Int J Med Inform 54(2), 77–95 (1999)
20. Wingert, F.: Automated indexing based on SNOMED. Methods Inf Med 24(1), 27–34 (1985)
21. van der Zwaan, J.: Development and evaluation of a method for the automatic coding of pathology report conclusions. Masters thesis, Faculty of Science, University of Amsterdam (2006)

Literature Mining: Towards Better Understanding of Autism

Tanja Urbančič[1,2], Ingrid Petrič[1], Bojan Cestnik[2,3], and Marta Macedoni-Lukšič[4]

[1] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia
[2] Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[3] Temida, d.o.o., Dunajska 51, 1000 Ljubljana, Slovenia
[4] University Children's Hospital, University Medical Center, 1000 Ljubljana, Slovenia
tanja.urbancic@p-ng.si, ingrid.petric@p-ng.si,
bojan.cestnik@temida.si,
marta.macedoni-luksic@mf.uni-lj.si

Abstract. In this article we present a literature mining method RaJoLink that upgrades Swanson's ABC model approach to uncovering hidden relations from a set of articles in a given domain. When these relations are interesting from medical point of view and can be verified by medical experts, they represent new pieces of knowledge and can contribute to better understanding of diseases. In our study we analyzed biomedical literature about autism, which is a very complex and not yet sufficiently understood domain. On the basis of word frequency statistics several rare terms were identified with the aim of generating potentially new explanations for the impairments that are observed in the affected population. Calcineurin was discovered as a joint term in the intersection of their corresponding literature. Similarly, NF-kappaB was recognized as a joint term. Pairs of documents that point to potential relations between the identified joint terms and autism were also automatically detected. Expert evaluation confirmed the relevance of these relations.

Keywords: literature mining, knowledge discovery, biomedical literature, autism.

1 Introduction

The amount and the speed of growth of scientific information available online have strongly influenced the way of work in the research community which calls for new methods and tools to support it. Biomedical field is a very good example, with MEDLINE database, the primary component of PubMed (the United States National Library of Medicine's bibliographic database), which covers more than 5.000 journals published in more than 80 countries, contains more than 15 million citations from the mid-1950's to the present, and increases for more than 1.500 complete references daily [15]. Knowledge technologies, especially knowledge discovery based on data mining and text mining, offer new possibilities by their ability to uncover hidden relationships in data [7]. Several examples of the applications in the biomedical field are included into a presentation of European data mining projects given in [20]. When

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 217–226, 2007.

a set of articles serves as a source of data, the process is typically called literature mining.

An early and very illustrative example of literature mining goes back into 1990, when Swanson presented his ABC model for discovering complementary structures in disjoint journal articles, leading to new hypotheses about diseases [19]. In his work he investigates whether an agent A influences a phenomenon C. To do this, he looked for interconnecting Bs, such that A causes phenomenon B (as reported in an article in literature about A) and the same B influences C (as reported in another article in literature about C). If articles about A (called also A literature) and articles about C (C literature) have few or no published papers in common, such discovered connections can turn out to be previously unknown. If they are also interesting from a medical point of view and can be verified by medical experts, they represent new pieces of knowledge and contribute to better understanding of diseases. This is particularly important in the case of complex pathological conditions, not yet sufficiently understood. To facilitate the discovery of hypotheses by linking findings across literature, Swanson and his colleagues designed a set of interactive software that is available on a web-based system called Arrowsmith [18].

Swanson's work inspired several researchers that continued his line of research. Pratt and Yetisgen-Yildiz [14] designed LitLinker that uses data mining techniques to identify correlations among concepts and then uses those correlations for discovery of potential causal links between biomedical terms. Weeber et al. [22] experimented with Swanson's idea of searching the literature for generating new potential therapeutic uses of the drug thalidomide with the use of a concept-based discovery support system DAD on the scientific literature. Another example of discovering new relations from bibliographic database according to Swanson's model is identification of disease candidate genes by an interactive discovery support system for biomedicine Bitola [9].

In Swanson's approach, either a specific agent A or a more general A category has to be determined in advance, so a target relationship that will be checked as a hypothesis has to be set before the process. In our work described in this article, we wanted to broaden the usability of the approach by suggesting how candidates for A can also be determined in a semi-automatic way. Our approach is based on identification of rare terms that could provide new explanations for the observed phenomena. The system automatically produces intermediate results, but is driven by human choices which rely on background knowledge of the domain. Expert's explicit involvement in the process enables for more focused and faster obtainment of results that experts find interesting and meaningful.

For our testing domain we chose autism. Autism belongs to a group of pervasive developmental disorders that are characterized by early delay and abnormal development of cognitive, communication and social interaction skills of a person. In most cases, these disorders have unclear origin. According to the Autism Society of America, autism is now considered to be an epidemic. The increase in the rate of autism revealed by epidemiological studies and government reports implicates the importance of external or environmental factors that may be changing. There's also an enormous increase of information in the field of autism research, which too often seems a fragmented tapestry stitched from differing analytical threads and theoretical patterns. This is the reason why studies seeking for factors that can help to put pieces together into a single, coherent object are so important.

In the next section we give a brief overview of data and tools that were used in our experiments. In section 3 we present a method RaJoLink which for a given phenomenon C generates candidate agents A that could contribute to better understanding of C and gives pairs of articles connecting C and A via linking terms B as a basis for expert evaluation of discovered relations. Results in the autism domain are also presented. In Section 4 we apply the same method on a domain, restricted as suggested by a medical expert. Finally, we summarize the most important findings.

2 Brief Overview of Experimental Set-Up: Data and Tools

In our study, articles about autism in the PubMed database served as a source of data. Among 10.821 documents (found till August 21, 2006) that contained derived forms of autis*, the expression root for autism, there were 354 articles with their entire text published in the PubMed Central database. Due to a noticeable shift in the focus of investigations about autism, we further restricted this set of articles to those published in the last ten years, which resulted in a set of 214 articles used as an initial source of data in our study. Later in the experiment, other subsets of PubMed articles were selected as a source of data, as described in sections 3 and 4.

In our work we used OntoGen [8], the interactive tool for semi-automatic construction of ontologies, accessible with detailed description at http://ontogen.ijs.si/index.html. OntoGen is based on machine learning and text mining techniques that automatically extract topics covered within the input documents and thus support the user of the system to organize those documents into a topic ontology. This primary functionality of OntoGen helped us to obtain an overview of the fundamental concepts of autism domain knowledge, but was in our case not crucial for the sake of knowledge discovery. We rather used OntoGen's other functionalities, such as generation of word frequency statistics and determination of document similarity on the basis of bag-of-words representation and cosine similarity.

3 Method RaJoLink

The proposed method for uncovering relations that could contribute to understanding of a phenomenon C (in our case, autism) consists of the following steps:

1. Identification of *n* interesting rare terms C_R_1, C_R_2, …C_R_n in literature about C.
2. Search for a joint term A in the literatures about C_R_1, C_R_2, …C_R_n.
3. Search for linking terms B_1, B_2, …., B_m such that for each B_i there exists a pair of articles, one from literature A and one from literature C, both mentioning Bi.
4. Expert evaluation in which it is checked whether obtained pairs of articles contribute to better understanding of C.

We call the method RaJoLink after its key elements: *Ra*re terms, *Jo*int terms and *Link*ing terms. In the following sections we describe these steps in more detail and illustrate them by the autism example.

3.1 Identification of Interesting Rare Terms in the Domain Literature

From the processed text file of autism articles we obtained also a *.txt.stat file with statistical incidence of terms as they appeared in the input documents collection. Initially we took the autism.txt.stat file that OntoGen formed while constructing autism ontologies and allocated our attention to the rarest records. This simple retrieval technique enables quick identification of interesting items.

Focusing on interesting terms that are very rarely mentioned in articles about autism is in our view more promising for new discoveries than exploring terms that are more frequent. With the proposed strategy we individuated a list of rare words that were, in regard of our autism background knowledge, viewed as promising in the sense that they could provide potential explanations of autistic disorders. From the 2192 rarest terms (terms that appeared once) all terms that were not domain specific were left out. From the remaining rare terms, 3 of them were chosen for further investigation: *lactoylglutathione, synaptophysin* and *calcium_channels*. Note that at the point of selecting interesting rare terms an expert's opinion might be very valuable.

To confirm the rarity of the chosen terms in the autism context, we searched the PubMed database for documents that contain the specific rare term together with the term autism. In fact, the term *lactoylglutathione* appeared only once together with the term autism, *synaptophysin* twice, and *calcium_channels* seven times.

3.2 Search for Joint Terms in the Literature About Rare Terms

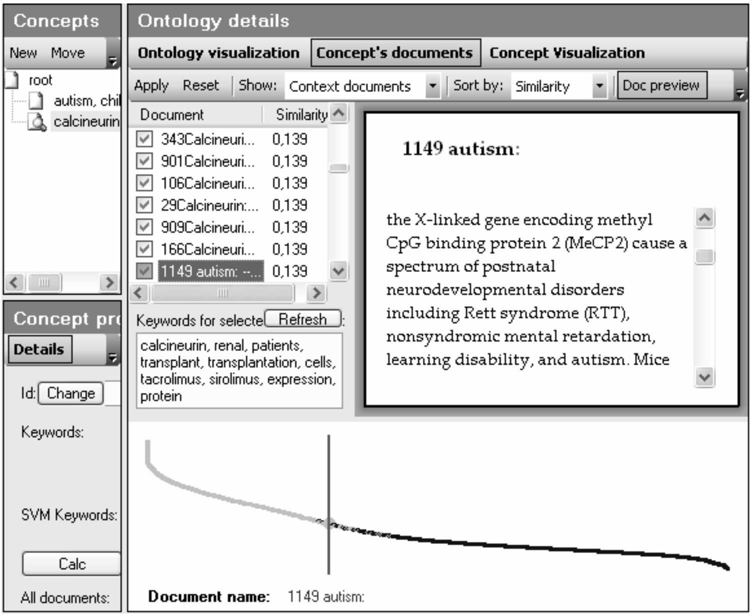
Rare terms identified in the previous step served as a starting point for our deeper investigations of some pathological mechanisms that may lead to the autistic-like manifestations. Therefore we decided to search for the biomedical literature about *lactoylglutathione*, about *synaptophysin* and about *calcium_channels* that is publicly accessible in the PubMed database. As a result, we got three different sets of biomedical articles that were converted into three separate text files: the first one containing abstracts of *lactoylglutathione* articles, the second one with abstracts of *synaptophysin* articles and the third text file with abstracts of *calcium_channels* articles from PubMed. Further search in these domains enabled us to determine joint terms that appeared in all of them.

To find joint terms, we again used OntoGen, which besides constructing ontologies on the three input files of abstracts, created also three *.txt.stat files. Each of the three *.txt.stat files contained the statistical incidence of terms as they appeared in the processed documents from each of the input datasets. From the statistical files taken together, we identified joint terms that appeared in all of them. In other words, these joint terms appear in the lactoylglutathione literature, in the synaptophysin literature as well as in the calcium_channels literature. From several joint terms that were found automatically, *calcineurin* was chosen for further investigation.

The reasoning behind this step is the following: If there are some rare terms that appear in the autism literature and they all have a joint term in their intersection, it is worthwhile checking if this joint term has some connections to autism. If the autism literature and the joint term literature have few or no published papers in common, such an uncovered connection might contribute to better understanding of autism.

3.3 Search for Linking Terms and the Corresponding Pairs of Articles

For detecting some pairs of calcineurin-autism articles that would help us to build prominent relationships between the domain of calcineurin and the domain of autism, we used a set of calcineurin articles abstracts and a set of autism articles abstracts, which we extracted from PubMed. We united the two sets of abstracts in a single dataset and thus generated a database of calcineurin+autism domain. We used this combined dataset as input for OntoGen and built a new ontology on it in order to obtain the information about the similarity between documents from the input dataset.



Document	Similarity
☑ 343Calcineuri...	0,139
☑ 901Calcineuri...	0,139
☑ 106Calcineuri...	0,139
☑ 29Calcineurin:...	0,139
☑ 909Calcineuri...	0,139
☑ 166Calcineuri...	0,139
☑ 1149 autism: -...	0,139

calcineurin, renal, patients, transplant, transplantation, cells, tacrolimus, sirolimus, expression, protein

1149 autism:

the X-linked gene encoding methyl CpG binding protein 2 (MeCP2) cause a spectrum of postnatal neurodevelopmental disorders including Rett syndrome (RTT), nonsyndromic mental retardation, learning disability, and autism. Mice

Document name: 1149 autism:

Fig. 1. OntoGen's representation of the set of autism and calcineurin articles according to their similarities. Two main topics (*autism and calcineurin*) are listed on the left side of the OntoGen's window. As the calcineurin topic is selected, the list of documents that are in the relationship with it is presented in the central part of the window. An outlying autism article (*1149 autism*) can be viewed inside the calcineurin context documents due to its similarity with the neighboring documents.

From the OntoGen's similarity graph (Figure 1) we could quickly notice, which documents are semantically strongly related to each of our research domains, autism and calcineurin, respectively, because they were clearly positioned on the two opposite sides of the similarity curve. However, regarding our goal of looking for relations between our two domains of research, the most prominent examples from the input dataset should be positioned on those graph sides, where the autism articles lay near the calcineurin articles. Therefore we focused our attention on the groups of the calcineurin-autism articles that were positioned in the vicinities according to their similarity. In this way we obtained pairs of calcineurin-autism articles containing terms with similar meanings. We used such terms as a hypothetical conjunct of calcineurin and autism domain. As the candidate hypotheses for calcineurin and autism relationship we found thirteen pairs of PubMed articles that, when put together, could connect the two categories, autism and calcineurin, respectively. Three of such pairs of articles are listed in Table 1.

This step of the method was inspired by the Swanson's ABC approach. Note that instead of determining agent A in advance, in our approach it is generated as a joint term in the second step of the method.

Table 1. Hypotheses for autism and calcineurin relationship

Autism literature	Calcineurin literature
Fatemi et al. [6] reported a reduction of *Bcl-2* (a regulatory protein for control of programmed brain cell death) levels in autistic cerebellum.	Erin et al. [5] observed that calcineurin occurred as a complex with *Bcl-2* in various regions of rat and mouse brain.
Belmonte et al. [3] reviewed neuropathological studies of cerebral cortex in autism indicating abnormal *synaptic* and columnar structure and neuronal migration defects.	Chen et al. [4] reported about the decrease in protein ubiquitination in synaptosomes and in nonneuronal cells that may play role in the regulation of *synaptic* function by a calcineurin antagonist FK506.
Huber et al. [10] showed evidences about an important functional role of fragile X protein, an identified cause of autism, in regulating activity-dependent *synaptic plasticity* in the brain.	Winder and Sweatt [23] described the critical role of protein phosphatase 1, protein phosphatase 2A and calcineurin in the activity-dependent alterations of *synaptic plasticity*.

3.4 Expert's Comment

The medical expert in our team confirmed that the generated relations draw attention to interesting connections between two well developed, but not sufficiently connected fields. In particular, she justified this statement by the following comment: Calcineurin is a calcium/calmodulin-dependent protein phosphatase [17]. Recent studies indicate that it participates in intracellular signaling pathways, which regulate synaptic plasticity and neuronal activities [16]. An impaired synaptic plasticity is thought to be also a consequence of the lack of FMR1 protein in fragile X syndrome which is one of the identified causes of autism [11].

4 Re-application of RaJoLink on a Restricted Domain

The four steps described in Section 4 resulted in uncovered relations that according to the expert evaluation could present a contribution towards better understanding of autism.

In addition, it should be mentioned that a full table with identified pairs of related articles proved to be very useful in our dialog with the domain expert since it guided the discussion very efficiently towards new ideas for further investigations. More concretely, the suggestion was to have a closer look at the significance of the fragile

X protein loss in autism as reported by Huber et al. [10]. This evaluation significantly helped us in reducing the hypothesis space. It encouraged us to further mine the data on autism in its particular relation to the fragile X. We did it by reapplying RaJoLink method, this time on a restricted set of articles that dealt with both, autism and fragile X, as follows.

With the goal to discover unsuspected associations between pieces of knowledge about autism and fragile X, we retrieved articles from PubMed that contain information about autism and that at the same time talk about the fragile X. We found 41 articles with their entire text published in the PubMed Central, which served as our input file of data on autism and fragile X. As in the case of our data mining on pure autism articles, we used them for ontology construction with OntoGen, and next, we utilized the OntoGen's statistical *.txt.stat file for the identification of those terms that rarely appeared in the PubMed documents collected in our input dataset.

When searching the statistical file we concentrated our attention on listed terms that appeared only in one document from the input dataset. In this way we chose some of these rare terms for the following text mining on the smallest pieces of autism and fragile X knowledge. The following three were chosen based on background knowledge: *BDNF (brain-derived neurotrophic factor), bicuculline*, and *c-Fos*. By searching in PubMed articles that treat each of the three selected terms domains, we constructed three separate ontologies. Afterwards, we searched the OntoGen's *.txt.stat files to find some interesting words that the listed domains have in common as joint terms. We found several promising terms belonging to three of the domains. One of such terms, which we found in the BDNF*.txt.stat file, in the bicuculline*.txt.stat file, as well as in the c-Fos*.txt.stat file, was the term *NF-kappaB*. Figure 2 illustrates how resulting joint terms were obtained for autism domain and for autism+fragile_X domain.

For a given hypotheses of NF-kappaB and autism relationship we found pairs of PubMed articles that could connect the knowledge about the transcription factor NF-kappaB with the domain of autism. We present three of such pairs in Table 2.

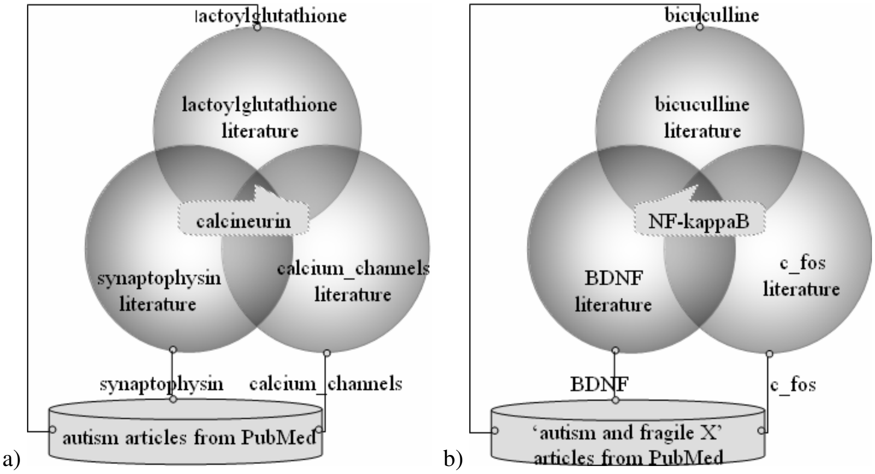


Fig. 2. Results obtained on (a) autism domain, and (b) autism+fragile_X domain

Table 2. Hypotheses for autism and NF-kappaB relationship

Autism literature	NF-kappaB literature
Araghi-Niknam and Fatemi [2] showed reduction of *Bcl-2*, an important marker of apoptosis, in frontal, parietal and cerebellar cortices of autistic individuals.	Mattson [12] reported in his review that activation of NF-kappaB in neurons can promote their survival by inducing the expression of genes encoding antiapoptotic proteins such as *Bcl-2* and the antioxidant enzyme Mn-superoxide dismutase.
Vargas et al. [21] reported altered *cytokine* expression profiles in brain tissues and cerebrospinal fluid of patients with autism.	Ahn and Aggarwal [1] reported that on activation NF-kappaB regulates the expression of almost 400 different genes, which include enzymes, *cytokine*s (such as TNF, IL-1, IL-6, IL-8, and chemokines), adhesion molecules, cell cycle regulatory molecules, viral proteins, and angiogenic factors.
Ming et al. [13] reported about the increased urinary excretion of an *oxidative stress* biomarker - 8-iso-PGF2alpha in autism.	Zou and Crews [24] reported about increase in NF-kappaB DNA binding following *oxidative stress* neurotoxicity.

The expert's comment to these finding was as follows. It is thought that autism could result from an interaction between genetic and environmental factors with an oxidative stress and immunological disorders as potential mechanisms linking the two [3], [13]. Both of the mechanisms are related to NF-kappaB as the result of our analysis. The activation of the transcriptional factor NF-kappaB was shown to prevent neuronal apoptosis in various cell cultures and in vivo models [12]. Oxidative stress and elevation of intracellular calcium levels are particularly important inducers of NF-kappaB activation. In addition, various other genes are responsive to the activation of the NF-kappaB, including those for cytokines. In this way the NF-kappaB can be involved in the complex linkage between the immune system and autism [3], [21]. So, according to our analysis one possible point of convergence between "oxidative stress" and "immunological disorder" paradigm in autism is NF-kappaB.

5 Conclusions

We present a literature mining method for searching pairs of papers in disjoint literatures that could, when linked together, contribute to a better understanding of complex pathological conditions, such as autism. We focus on rare terms to generate potentially new explanations for the impairments that are observed in the affected population. With this goal we further review the main aspects of the chosen rare terms with the ontology construction on each of these starting point domains. It is on these latter aspects that we focus furthermore, as we attempt to investigate whether any of chosen rare terms relate to each other. In fact, our assumption is that such known relations might lead us to discovering implicit knowledge about autism in previously unrelated biomedical literature. With the calcineurin and NF-kappaB examples we finally illustrate the potential of literature mining to detect links between unrelated biomedical articles.

By detecting published evidence of some autism findings on one hand that coincide with specific calcineurin and NF-kappaB observations on the other hand, we present possible relationships between autism and calcineurin literature, as well as between autism and NF-kappaB literature. Further research about timing, environmental conditions, maturational differences in brain development, and other determinants of calcineurin and NF-kappaB involvement in autism spectrum disorders is needed for stronger evidence, but in any case, the method has proved its potential in supporting experts on their way towards new discoveries in biomedical field.

Acknowledgments. This work was partially supported by the Slovenian Research Agency program Knowledge Technologies (2004-2008). We thank Nada Lavrač for her suggestion to use OntoGen and Blaž Fortuna for his discussions about OntoGen's performance.

References

1. Ahn, K.S., Aggarwal, B.B.: Transcription Factor NF-{kappa}B: A Sensor for Smoke and Stress Signals. Annals of the New York Academy of Sciences 1056, 218–233 (2005)
2. Araghi-Niknam, M., Fatemi, S.H.: Levels of Bcl-2 and P53 are altered in superior frontal and cerebellar cortices of autistic subjects. Cellular and Molecular Neurobiology 23(6), 945–952 (2003)
3. Belmonte, M.K., Allen, G., Beckel-Mitchener, A., Boulanger, L.M., Carper, R.A., Webb, S.J.: Autism and abnormal development of brain connectivity. The Journal of Neuroscience 27(42), 9228–9231 (2004)
4. Chen, H., Polo, S., Di Fiore, P.P., De Camilli, P.V.: Rapid Ca2+-dependent decrease of protein ubiquitination at synapses. Proceedings of the National Academy of Sciences of the United States of America 100(25), 14908–14913 (2003)
5. Erin, N., Bronson, S.K., Billingsley, M.L.: Calcium-dependent interaction of calcineurin with Bcl-2 in neuronal tissue. Neuroscience 117(3), 541–555 (2003)
6. Fatemi, S.H., Stary, J.M., Halt, A.R., Realmuto, G.R.: Dysregulation of Reelin and Bcl-2 proteins in autistic cerebellum. Journal of Autism and Developmental Disorders 31(6), 529–535 (2001)
7. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, pp. 82–88 (1996)
8. Fortuna, B., Grobelnik, M., Mladenić, D.: Semi-automatic Data-driven Ontology Construction System. In: Proceedings of the 9th International multi-conference Information Society IS-2006, Ljubljana, Slovenia, pp. 223–226 (2006)
9. Hristovski, D., Peterlin, B., Mitchell, J.A., Humphrey, S.M.: Using literature-based discovery to identify disease candidate genes. International Journal of Medical Informatics 74, 289–298 (2005)
10. Huber, K.M., Gallagher, S.M., Warren, S.T., Bear, M.F.: Altered synaptic plasticity in a mouse model of fragile X mental retardation. Proceedings of the National Academy of Sciences of the United States of America 99(11), 7746–7750 (2002)
11. Irwin, S., Galvez, R., Weiler, I.J., Beckel-Mitchener, A., Greenough, W.: Brain structure and the functions of FMR1 protein. In: Hagerman, R.J., Hagerman, P., J. Fragile X syndrome. The Johns Hopkins University Press, Baltimore, pp. 191–205 (2002)

12. Mattson, M.P.: NF-kappaB in the survival and plasticity of neurons. Neurochemical Research 30(6-7), 883–893 (2005)
13. Ming, X., Stein, T.P., Brimacombe, M., Johnson, W.G., Lambert, G.H., Wagner, G.C.: Increased excretion of a lipid peroxidation biomarker in autism. Prostaglandins, Leukotrienes, and Essential Fatty Acids 73(5), 379–384 (2005)
14. Pratt, W., Yetisgen-Yildiz, M.: LitLinker: Capturing Connections across the Biomedical Literature. In: Proceedings of the International Conference on Knowledge Capture (K-Cap'03), Florida, pp. 105–112 (2003)
15. PubMed: Overview (January 2007) http://www.ncbi.nlm.nih.gov/
16. Qiu, S., Korwek, K.M., Weeber, E.J.: A fresh look at an ancient receptor family: emerging roles for density lipoprotein receptors in synaptic plasticity and memory formation. Neurobiology of Learning and Memory 85(1), 16–29 (2006)
17. Rusnak, F., Mertz, P.: Calcineurin: Form and Function. Physiological Reviews 80(4), 1483–1521 (2000)
18. Smalheiser, N.R., Swanson, D.R.: Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine 57, 149–153 (1998)
19. Swanson, D.R.: Medical literature as a potential source of new knowledge. Bulletin of the Medical Library Association 78(1), 29–37 (1990)
20. Van Someren, M., Urbančič, T.: Applications of machine learning: matching problems to tasks and methods. The Knowledge Engineering Review 20(4), 363–402 (2006)
21. Vargas, D.L., Nascimbene, C., Krishnan, C., Zimmerman, A.W., Pardo, C.A.: Neuroglial activation and neuroinflammation in the brain of patients with autism. Annals of Neurology 57(1), 67–81 (2005)
22. Weeber, M., Vos, R., Klein, H., De Jong-van den Berg, L.T., Aronson, A.R., Molema, G.: Generating Hypotheses by Discovering Implicit Associations in the Literature: A case Report of a Search for New Potential Therapeutic Uses for Thalidomide. Journal of the American Medical Informatics Association 10(3), 252–259 (2003)
23. Winder, D.G., Sweatt, J.D.: Roles of serine/threonine phosphatases in hippocampal synaptic plasticity. Nature reviews Neuroscience 2(7), 461–474 (2001)
24. Zou, J., Crews, F.: CREB and NF-kappaB Transcription Factors Regulate Sensitivity to Excitotoxic and Oxidative Stress Induced Neuronal Cell Death. Cellular and Molecular Neurobiology 26(4-6), 383–403 (2006)

Automatic Generation of Textual Summaries from Neonatal Intensive Care Data

François Portet, Ehud Reiter, Jim Hunter, and Somayajulu Sripada

Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK
{fportet,ereiter,jhunter,ssripada}@csd.abdn.ac.uk

Intensive care is becoming increasingly complex. If mistakes are to be avoided, there is a need for the large amount of clinical data to be presented effectively to the medical staff. Although the most common approach is to present the data graphically, it has been shown that textual summarisation can lead to improved decision making. As the first step in the BabyTalk project, a prototype is being developed which will generate a textual summary of 45 minutes of continuous physiological signals and discrete events (e.g.: equipment settings and drug administration). Its architecture brings together techniques from the different areas of signal analysis, medical reasoning, and natural language generation. Although the current system is still being improved, it is powerful enough to generate meaningful texts containing the most relevant information. This prototype will be extended to summarize several hours of data and to include clinical interpretation.

1 Introduction

In the Intensive Care Unit (ICU), the interpretation of clinical data is an essential task. Generally, the data available consists of (i) continuously monitored physiological variables (e.g.: heart rate, blood pressure, etc.) and (ii) discrete events (e.g.: equipment settings, drug administration, etc.). However the volumes of data are so large (about 1 MB per patient per day), that attention overload (looking after several patients) and stress can lead to mistakes being be made. Hence, presenting these data in an efficient way is crucial for informed medical decision making.

Although the graphical presentation of data is becoming standard practice, an off-ward experiment conducted as part of the Neonate project (Hunter et al., 2003) showed that medical professionals, in some circumstances, are more likely to make better treatment decisions if they are given a textual summary of patient data, instead of a graphical one (Law et al., 2005). In this *GraphVsText* experiment, forty nurses and doctors with different levels of expertise were asked (individually) to say what actions(s) they would take for a baby whose recent history over a period of about 45 minutes was presented either graphically (as time series plots) or in the form of a textual summary; the text was written by senior clinicians based on the graphical presentation. Each participant was presented with 16 cases, eight graphical and eight textual. Although the clinicians said they preferred the more familiar graphical presentation, they chose more correct actions after reading the textual summaries. Recent experimental research comparing textual and graphical summaries of mobile phone manuals (Lagan-Fox et al., 2006) also showed a decision-making improvement with texts.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 227–236, 2007.

These results motivated the BabyTalk[1] project whose goal is the automatic generation of texts summarising baby's ICU data in neonatal units. Over the past decade, we have acquired considerable experience in Data Analysis in neonatology with the Cognate (Logie et al., 1997) and Neonate projects as well as in the generation of text from discrete and continuous data with the SunTime (Sripada et al., 2002, Yu et al. 2002) project. A number of techniques have been developed elsewhere for summarising clinical data. For example, the CLEF project (Hallett et al., 2005) aims at generating summaries of multiple text-based health reports. Perhaps the most successful applications have been tools that (partially) automate the process of writing routine documents, such as Hüske-Kraus's Suregen system (Hüske-Kraus, 2003a), which is regularly used by physicians to create surgical reports; see Hüske-Kraus, 2003b for a review of text generation in medicine. However, the complete summarisation of ICU data is more complex, involving the processing of time series, discrete events, and short free texts, which seems to have not been done before.

The BabyTalk project aims at providing summaries according to two different dimensions: duration and degree of abstraction. Four mains systems are planned:

1. **BT-45:** descriptive summary of 45 minutes data
2. BT-Nurse: summary of 12 hours of data to serve as a shift summary.
3. BT-Doc: similar to BT-nurse but with the intention of supporting decision making by junior doctors.
4. BT-Family: a daily summary for the baby's family, adapted to the emotional state of the recipient.

BT-45 is the simplest system as it covers a limit time period and is purely descriptive. It will be used as a stepping stone to the development of the other systems which involve longer durations and more interpretation. BT-45 will be evaluated by repeating the *GraphVsText* experiment with *three* types of presentation: graphical, text written by experts, text generated automatically. This paper describes the progress we have made towards the implementation of BT-45. The data that BT-45 must deal with and the target textual outputs are presented Section 2. The architecture is discussed in Section 3. The prototype has been informally tested by comparison with the manually generated summaries and the results of this comparison are presented in Section 4. The paper ends with a discussion about necessary improvements and other future activities within the project.

Although our work is carried out in the context of neonatal intensive care, we expect the principles to be applicable more widely to adult ICU and to other high-dependency units.

2 Inputs and Outputs

2.1 Input Data

The inputs to BT-45 are of two kinds: (i) continuous multi-channel time series data from the physiological monitors and (ii) discrete event data such as the entry of laboratory results, actions taken, etc.

[1] http://www.csd.abdn.ac.uk/research/babytalk/

Physiological time series data

A maximum of seven channels were recorded: the Heart Rate (HR), the pressures of oxygen and carbon dioxide in the blood (OX and CO), the oxygen saturation (SO), the peripheral and central temperatures of the baby (TP and TC) and the mean blood pressure (BM). We have over 400 hours of continuously recorded data from babies in the Neonatal Intensive Care Unit at the Edinburgh Royal Infirmary. As with all real ICU data, our data are sometimes incomplete (periods for which some probes are off) and contain periods of noise.

Discrete data

As part of the Neonate project (Hunter et al., 2003) we employed a research nurse to be present at the cot-side and to record *all* of the following types of event:

- the **equipment** used to monitor, ventilate, etc.;
- the **settings** on the various items of equipment (including the ventilator);
- the results of **blood gas** analysis and other **laboratory results**;
- the current **alarm limits** on the monitors;
- the **drugs** administered;
- the **actions** taken by the medical staff;
- occasional descriptions of the physical state of the baby (**observations**);

The unit we are working with is about to go ‘paperless’. This means that we can expect that the equipment used, settings, lab and blood gas results, alarm limits and medication will be automatically recorded. However, human activities such as actions and observations are more difficult to acquire electronically and it is not clear at present exactly what will be recorded and with what timing accuracy.

For the implementation of BT-45 we will use the data collected by the research nurse. However we realise that for systems to operate in the real world, they will only be able to access that information which is available to them electronically and it is part of our research agenda so see to what extent missing items can be derived from what *is* available.

2.2 Output Data

The summaries against which we will compare our automatically generated texts were created by a consultant neonatologist and an experienced neonatal nurse researcher. In order to obtain a comparison with the graphical presentation which was as valid as possible, we took steps to ensure that the texts were purely descriptive (i.e. did not contain higher level clinical interpretations) and contained information that came from the 45 minute data period only; 18 summaries were generated for time-periods varying between 30 minutes and 53 minutes (mean = 40.5). An example appears in Fig 5.

3 Architecture

The main architecture of the prototype is shown in Fig. 1. BT-45 creates a summary of the clinical data in four main stages. The physiological time series and the annotations are processed by **Signal Analysis (1)** to abstract the main features of the signals (artifacts, patterns, and trends). **Data Abstraction (2)** performs some basic reasoning to infer relations between events (i.e.: "A" causes "B"). From the large number of propositions generated, **Content Determination (3)** selects the most important**,** and aggregates them into a tree of linked events. Finally, **Micro Planning and Realization (4)** translates this tree into text. All of the terms used to describe the discrete events are **described by** an **Ontology (5)** of NICU concepts. These were mainly acquired during the Neonate project and are still being extended.

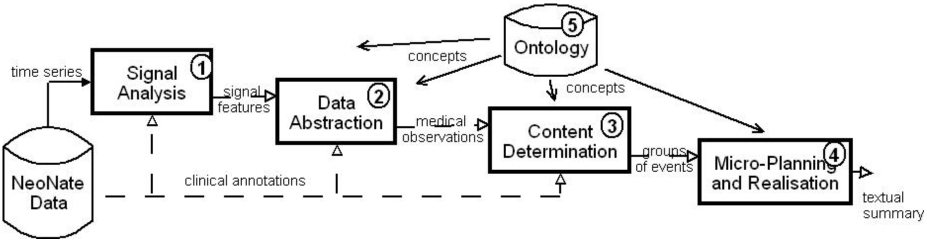


Fig. 1. Architecture of BT-45

3.1 Signal Analysis

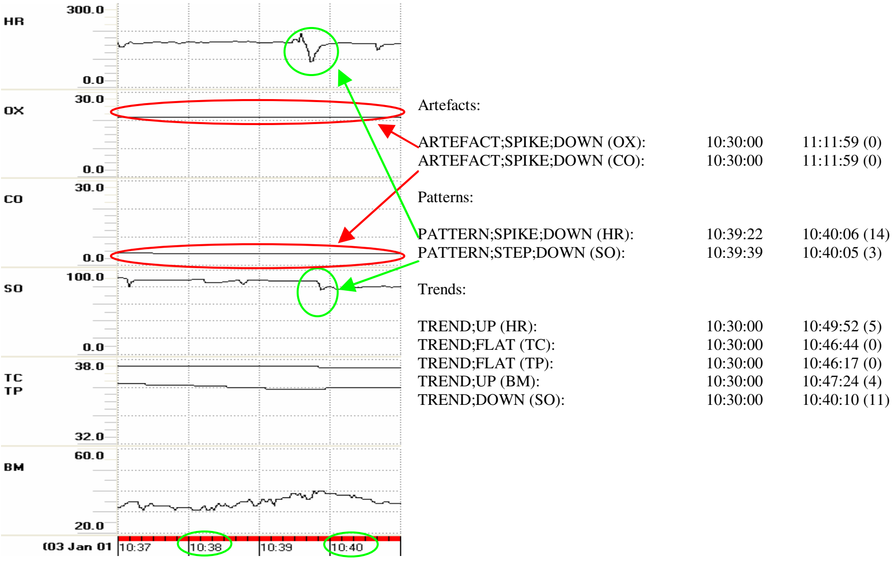
This module analyses the time series data to detect artifacts, patterns, and trends.

An artifact is defined as a sequence of signal sample values that do not reflect real physiological data. In BT-45, the first stage of detection consists of simple thresholding of the impossible values (mainly due to a probe falling off, being partially detached, being removed by a nurse, etc.). For example a heart rate cannot be 0 and a baby temperature cannot be physiologically below 30 degrees Celsius. The artifact time intervals are then merged if they are close to each other (within 10 sec). The second stage of the artifact detection is performed by an expert system which relates the artifacts between the different channels. For example, as the OX and CO channels are derived from the same probe (the transcutaneous probe), if an artifact appears on one channel, it should also appear on the other.

Pattern recognition is based on the rapid-change detector of the *SumTime-Turbine* project (Yu et al. 2002) and looks for cases where the signal data is changing rapidly. Pattern intervals are created by merging nearby rapid-change points, and these are then classified into two kinds of patterns, spike and step, using heuristics.

Trend detection uses bottom-up segmentation (Keogh et al., 2001). The code is a simplified version of the segmentation of the SumTime-Mousam project (Sripada et al., 2002). Bottom-up segmentation consists in merging neighbouring segments iteratively into larger ones. Before the merging, two neighbourhood segments are approximated by a line and if the error is less than a specific threshold then the segments are merged. The operation is repeated until the total error reaches a specific threshold or only one segment remains. In this implementation, every sample of the time series belonging to an artifact or a spike is ignored. This enables us to acquire the longer-term trends of a time series rather than rapidly changing features.

The output of Signal Analysis consists of events with a stated duration. For example, for scenario 1 during the period 10:38 to 10:40, the events presented Fig. 2 are generated. Each line consists of: **event type** (**channel**), **start time**, **end time** (**importance**)”, where importance is scored from 0 to 100. The first line shows that samples of the OX channel have been classified as artefact and the main shape of these samples corresponds to a downward spike. As OX and CO come from the same probe, the second stage of the artefact detection inferred the same period as artifact on the CO channel. Two rapid changes have been detected by the pattern recognizer on the HR and SO channel. Then trends have been established for other channels. Note that the computation of the upward trend on HR did not take into account the period during which a downward spike was detected.



```
Artefacts:

ARTEFACT;SPIKE;DOWN (OX):     10:30:00    11:11:59 (0)
ARTEFACT;SPIKE;DOWN (CO):     10:30:00    11:11:59 (0)

Patterns:

PATTERN;SPIKE;DOWN (HR):      10:39:22    10:40:06 (14)
PATTERN;STEP;DOWN (SO):       10:39:39    10:40:05 (3)

Trends:

TREND;UP (HR):                10:30:00    10:49:52 (5)
TREND;FLAT (TC):              10:30:00    10:46:44 (0)
TREND;FLAT (TP):              10:30:00    10:46:17 (0)
TREND;UP (BM):                10:30:00    10:47:24 (4)
TREND;DOWN (SO):              10:30:00    10:40:10 (11)
```

Fig. 2. Input and Output of Signal Analysis for scenario 1 from 10:38 to 10:40

3.2 Data Abstraction

Data abstraction consists in finding pairs of events that are related by heuristics. There are three kinds of link: **causes**, **includes** and **associates**. For examples, if a bradycardia is found during an intubation then this intubation is the likely **cause** of the bradycardia; **includes** is used for events that are always accompanied by other events (e.g.: hand-bagging is included in intubation), and **associates** is for obvious correlations (e.g.: overlapping spikes in OX and CO are associated as they come from the same probe).

The output of the Data Abstraction module consists of relations between events. Fig 3 shows the results of the processing of the events of Fig. 2. The first link found associates a downward step on SO with a downward spike on HR. The rule stated that if there are two downward patterns on HR and SO during a short period then they should be associated with the same external phenomenon. The last link states that the decrease in SO should have caused the increase of the FiO2 (fraction of inspired

```
Link (ASSOCIATED)
   PATTERN;STEP;DOWN (SO):                  10:39:39     10:40:05 (3)
   PATTERN;SPIKE;DOWN (HR):                 10:39:22     10:40:06 (14)
Link (ASSOCIATED)
   PATTERN;STEP;DOWN (SO):                  10:39:39     10:40:05 (3)
   PATTERN;SPIKE;DOWN (HR):                 10:40:27     10:41:31 (4)
Link (ASSOCIATED)
   PATTERN;STEP;DOWN (SO):                  10:39:39     10:40:05 (3)
   TREND;UP (SO):                           10:40:10     10:47:56 (10)
Link (CAUSES)
   SETTING;VENTILATOR;FiO2 (35.0):          10:38:38     10:38:38 (2)
  TREND;UP;SO (SO):                         10:40:10     10:47:56 (10)
Link (CAUSES)
   TREND;DOWN (SO):                         10:30:00     10:40:10 (11)
   SETTING;VENTILATOR;FiO2 (36.0):          10:30:10     10:30:10 (10)
```

Fig. 3. Output of Data Abstraction

oxygen). Indeed, the SO is known to be influenced by the nurse increasing the FiO2 in cases of desaturation.

3.3 Content Determination

Content Determination decides what information needs to be communicated in the text, and how this information should be structured. To do so, events are grouped according to relational links between them in a way similar to Hallet et al. (2005). The event groups are then used to compose the tree for the document. The decision as to which groups to mention in the text is based on heuristics which try to produce a document of a certain length. A pruning of events is also based on their importance and on a notion of not mentioning events which the reader will infer by herself. Content Determination also generates introductory information which describes the state of the baby at the start of the scenario. Finally, the components are temporally ordered, and linked with Temporal-Sequence relations.

Fig 4 shows the grouping derived from the linked events from Fig 2 and Fig 3. The left side of the figure shows a group of temporally ordered events and the links between events; dashed arrows represent associate links and full arrows causal links. The right side of the figure shows the tree composed from the group; the number of "+" signs gives the depth of the node) which is composed of phrases of different types. The reference to FiO2 forms the beginning of the paragraph and it is followed by a causal relation phrase and two other phrases. The pruning removes unimportant

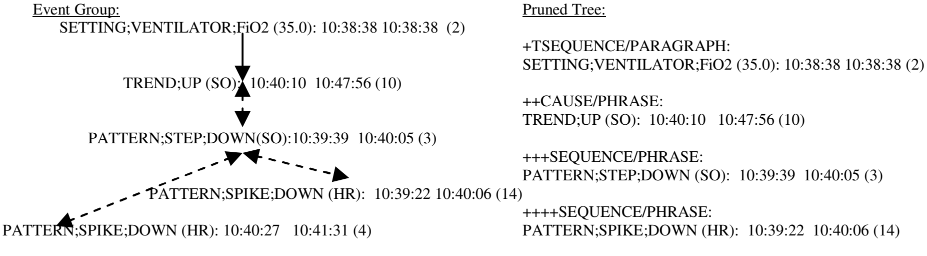

Event Group:
SETTING;VENTILATOR;FiO2 (35.0): 10:38:38 10:38:38 (2)
TREND;UP (SO): 10:40:10 10:47:56 (10)
PATTERN;STEP;DOWN(SO):10:39:39 10:40:05 (3)
PATTERN;SPIKE;DOWN (HR): 10:39:22 10:40:06 (14)
PATTERN;SPIKE;DOWN (HR): 10:40:27 10:41:31 (4)
Pruned Tree:
+TSEQUENCE/PARAGRAPH:
SETTING;VENTILATOR;FiO2 (35.0): 10:38:38 10:38:38 (2)
++CAUSE/PHRASE:
TREND;UP (SO): 10:40:10 10:47:56 (10)
+++SEQUENCE/PHRASE:
PATTERN;STEP;DOWN (SO): 10:39:39 10:40:05 (3)
++++SEQUENCE/PHRASE:
PATTERN;SPIKE;DOWN (HR): 10:39:22 10:40:06 (14)

Fig. 4. Group of liked events and the corresponding pruned tree

subtrees; that is why the leaf PATTERN;SPIKE;DOWN (HR): (4) has been remove but not the node PATTERN;STEP;DOWN(SO): (3) because it is the parent of the important leaf PATTERN;SPIKE;DOWN (HR): (14).

3.4 Microplanning and Realisation

The Microplanner converts the tree into text. Most events are converted into simple syntactic structures using mapping rules. For example, increasing the ventilator FiO2 level is converted into a phrase with verb "increase", object "FiO2", a prepositional phrase saying "to XX" (where XX is the new level), and possibly another prepositional phrase saying "at TT" (where TT is the time of the action). The tree (events related by discourse relations) is linearised top-down. A parent is always expressed before its children. The order that the children are expressed in depends on their discourse relation to their parent.

The realisation performs the actual translation into text from the output of the microplanner via a syntactic structure and completes the morphology and the layout of the document. The realisation of the tree of the Fig 4 gives:

"At 10:38 FiO2 is increased to 35 **so** saturations rise for 8 minutes to 96 **and** saturations drop to 78 **and** there is a bradycardia to 90".

Each event is converted by microplanning. The causal phrase is linked to the previous one by "**so**" and the sequence phrases are linked by an "**and**" to form a single sentence.

4 Preliminary Results

BT-45 is currently being developed and will be evaluated experimentally by comparing graphical presentation, human textual summaries and computerized textual summaries following the method described in Law et al. (2005). We have presented some outputs to a senior neonatologist who was very impressed by the generated texts even if they are not exploitable at present. As a general criterion, the length of the texts has been computed. This showed that our outputs were generally too long (median of 138 words) and with a greater standard deviation (56.4 words) than the human summaries (median of 110 words with STD = 41.1 words).

To demonstrate the performance of the prototype, we present an informal comparison of computerized and expert summaries for scenario 1, represented in Fig. 5. Scenario 1 concerns the re-intubation of a baby. The action we expected to be recommended at the end of the period was the decision to X-ray the baby in order to check the position of the endo-tracheal tube. Consequently, information about intubation, suction and the physiological variables are important.

- BT-45 introduced the start time and values of the physiological variables (as is normally the case) whereas, this time, the human writer didn't. Dealing with the multiplicity of styles in human writers is an important problem out of the scope of this paper (see Reiter et al., 2005). Moreover, BT-45 did not indicate that OX and CO are being re-sited (hence the values are wrong), but, at least the OX and CO channels have been detected as containing artifact and not included in the text.

BT-45

*** introduction ***

At 10:30 you see the baby.

HR = 148, mean BP = 28, central temperature = 37.5, peripheral temperature = 36.3 and sats = 96.

*** 1 ***

Saturations fall for 10 minutes to 77 so FiO2 is decreased to 36.

At 10:38 FiO2 is increased to 35 so saturations rise for 8 minutes to 96 and saturations drop to 78 and there is a bradycardia to 90.

*** 2 ***

At 10:46 peripheral temperature falls for 13 minutes to 33.5 and then central temperature falls for 5 minutes to 36.5.

At 10:51 the baby is intubated and so there is a significant bradycardia to 61, there is a desaturation to 77 and mean BP jumps to 47. As part of this procedure, at 10:47 the baby is hand-bagged so there is a bradycardia to 125.

At 10:52 toe/core temperature gap rises for 7 minutes to 2.4, at 11:02 FiO2 is decreased to 67 and FiO2 is increased to 79.

*** 3 ***

Saturations drop to 79 and then at 11:05 saturations rise for 6 minutes to 99 so FiO2 is decreased to 80.

The baby is sucked out and at 11:09 FiO2 is increased to 61.

EXPERT

*** introduction ***

You see the infant first at 1030.

The transcutaneous OX/CO electrode is being recalibrated.

*** 1 ***

In preparation for re-intubation, a bolus of 50ug of morphine is given at 1039 when the FiO2 = 35%. There is a momentary bradycardia and then the mean BP increases to 40. The sats go down to 79 and take 2 mins to come back up. The toe/core temperature gap increases to 1.6 degrees.

*** 2 ***

At 1046 the baby is turned for re-intubation and reintubation is complete by 1100 the baby being bagged with 60% oxygen between tubes. During the re-intubation there have been some significant bradycardias down to 60/min, but the sats have remained OK. The mean BP has varied between 23 and 56, but has now settled at 30. The central temperature has fallen to 36.1°C and the peripheral temperature to 33.7°C. The baby has needed up to 80% oxygen to keep the sats up.

*** 3 ***

Over the next 10 mins the HR decreases to 140 and the mean BP = 30-40. The sats fall with ETT suction so the FiO2 is increased to 80% but by 1112 the FiO2 is down to 49%.

Fig. 5. Examples of BT-45 and expert summaries for the same scenario

- The first part of the summary addresses the correspondence between desaturation and the FiO2 settings. This information is also present in the human summary but in a more condensed way and with more information about medication and intubation. This information is available to BT-45 and rules about the protocols for intubation should be added to the Data Abstraction module.
- The second part concerns the re-intubation. In this important period, BT-45 succeeded in tying the bradycardia and hand-bagging to the intubation. The temperature, even if not summarised enough is also described. The desaturation event seems to contradict the human text: "but the sats [SO] have remained OK". This is due to the expert's view that this desaturation (actually present in the data) is not relevant as far as the intubation is concerned.
- The third part is about the saturation problems following the intubation. BT-45 detected the falls in saturation and related them to the FiO2 setting and the suction but didn't mention the fall in heart rate and the variation in mean BP.

This informal comparison enabled us to identify three main problems:

- Crucial information about medication and medical activities such as intubation must be handled by the system by increasing significantly the number of expert rules in the Data Abstraction module.
- The lack of aggregation (e.g.: the FiO2 is increased, and the FiO2 is decreased, etc.) leads to texts which are too long and too far from the human style. The aggregation could be performed in the Data Abstraction module by a mechanism that groups the overlapping events of same type into a "sequence of events".
- Information is not highlighted in the text. Important event must be emphasized and less important events must be hidden. This is highly dependent on context e.g. if the baby is being intubated or is under specific medication.

Despite these drawbacks (which are being addressed in the next version) the BT-45 output contains the most important information in this scenario: intubation, hand bagging, suction, desaturation and bradycardia.

5 Discussion

BT-45 is an ongoing project and there is much to do in order to reach the quality of the experts' text (if indeed this is possible). However, our prototype has demonstrated that it is possible to perform simple data analysis and reasoning that are sufficient to generate a text where the most important information is presented.

Although the Signal Analysis module is composed of simple algorithms which perform in satisfactory way in our scenarios, they must be extended to deal with noisier data. Many artifact removal algorithms exist and an experiment to compare a number of them on noisy neonatal data is planned. Another issue is the detection of human activities from the signal (e.g.: re-siting of probes). For this, an approach based on trend and syntactic analysis will be investigated (Hunter and McIntosh, 1999).

The Data Abstraction is for the moment very basic. At this level we need to generate more high level abstractions such as the qualification of events (e.g.: "significant" bradycardia), aggregation of similar terms, more linking, etc. This is the weakest component of BT-45 at present and improvements will be based on the acquisition and formalisation of a greater range of expert knowledge.

The way the Content Determination module selects and organizes the information to present in the text is mainly based on the importance factor of the events. This clearly needs to be sensitive to the protocols/procedures being applied and to the characteristics of different babies. Thus, the decision as to which information to hide and to show is an important issue.

The microplanning and realization translates the tree of events into text. It should be more sensitive to aggregation and reference that can be controlled by stylistic parameters (such as desired sentence length). Then, a general lexicalisation engine must be set up to control the usage of technical vocabulary and vague modifiers (such as "small" spike). Moreover, a technique must be implemented to control the multiple time references in the text.

References

Hallett, C., Scott, D.: Structural variation in generated health reports. In: Proceedings of the 3rd International Workshop on Paraphrasing, Jeju Island, Korea (2005)

Hunter, J.R.W., McIntosh, N.: Knowledge-Based Event Detection in Complex Time Series Data. In: Horn, W., Shahar, Y., Lindberg, G., Andreassen, S., Wyatt, J.C. (eds.) AIMDM 1999. LNCS (LNAI), vol. 1620, pp. 271–280. Springer, Heidelberg (1999)

Hunter, J.R.W., Ferguson, L., Freer, Y., Ewing, G., Logie, R., McCue, P., McIntosh, N.: The NEONATE Database. In: Workshop on Intelligent Data Analysis in Medicine and Pharmacology and Knowledge-Based Information Management in Anaesthesia and Intensive Care, AIME-03, pp. 21–24 (2003)

Hüske-Kraus D.: Suregen-2: A Shell System for the Generation of Clinical Documents. In: Proceedings of EACL-2003 (demo session) (2003a)

Hüske-Kraus, D.: Text Generation in Clinical Medicine – a Review. Methods of Information in Medicine 42, 51–60 (2003b)

Keogh, E., Chu, S., Hart, D., Pazzani, M.: An Online Algorithm for Segmenting Time Series. In: Proceedings of IEEE International Conference on Data Mining. pp. 289–296 (2001)

Langan-Fox, J., Platania-Phung, C., Waycott, J.: Effects of Advance Organizers, Mental Models and Abilities on Task and Recall Performance Using a Mobile Phone Network. Applied cognitive psychology 20, 1143–1165 (2006)

Law, A.S., Freer, Y., Hunter, J.R.W., Logie, R.H., McIntosh, N., Quinn, J.: A Comparison of Graphical and Textual Presentations of Time Series Data to Support Medical Decision Making in the Neonatal Intensive Care Unit. Journal of Clinical Monitoring and Computing 19, 183–194 (2005)

Logie, R.H., Hunter, J.R.W., McIntosh, N., Gilhooly, K.J., Alberdi, E., Reiss, J.: Medical Cognition and Computer Support in the Intensive Care Unit: A Cognitive Engineering Approach. Engineering Psychology and Cognitive Ergonomics: Integration of Theory and Application, pp. 167–174 (1997)

Reiter, E., Sripada, S., Hunter, J., Yu, J., Davy, I.: Choosing Words in Computer-Generated Weather Forecasts. Artificial Intelligence 167, 137–169 (2005)

Sripada, S.G., Reiter, E., Hunter, J., Yu, J.: Segmenting Time Series for Weather Forecasting. Applications and Innovations in Intelligent Systems X, 193–206 (2002)

Yu, J., Hunter, J., Reiter, E., Sripada, S.G.: Recognising Visual Patterns to Communicate Time-Series Data in the gas turbine domain. Applications and Innovations in Intelligent Systems X, 105–118 (2002)

Anonymisation of Swedish Clinical Data

Dimitrios Kokkinakis[1] and Anders Thurin[2]

[1] Göteborg University, Department of Swedish Language, Språkdata Sweden
dimitrios.kokkinakis@svenska.gu.se
[2] Clinical Physiology, Sahlgrenska Univ. Hospital/Östra, Sweden
anders.thurin@vgregion.se

Abstract. There is a constantly growing demand for exchanging clinical and health-related information electronically. In the era of the *Electronic Health Record* the release of individual data for research, health care statistics, monitoring of new diagnostic tests and tracking disease outbreak alerts are some of the areas in which the protection of (patient) privacy has become an important concern. In this paper we present a system for automatic anonymisation of Swedish clinical free text, in the form of discharge letters, by applying generic named entity recognition technology.

Keywords: anonymisation, hospital discharge letters, entity recognition.

1 Introduction

An anonymisation system can provide a broad spectrum of services related to the growing demands for better forms of dissemination of confidential information about individuals (Personal Health Information – PHI) found in electronic health records (EHR) and other clinical free text. On a daily basis, hospitals store vast amounts of patient data, but due to confidentiality requirements these data – mostly texts – remain inaccessible for research and knowledge mining. In this paper we present an anonymisation system for Swedish, which re-uses components of a generic named entity recognition system (NER) ([1]; [2]). Generic NER is the process of identifying and marking all single or multi-word named persons, location and organizations, including time and measure expressions. NER is considered a mature technology that has numerous applications in a number of human language technologies, including information retrieval and extraction, topic categorization and machine translation.

2 Anonymisation

We define (*permanent*) *anonymisation* the process of recognizing and deliberately removing named entities and other identifying information about entities, including time expressions. Information about individuals, e.g. patients, may also include numerical, e.g. demographic or nominative information, such as age, sex and nationality, hence making the re-identification of those entities extremely difficult.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 237–241, 2007.

The related notion of *de-personalization* or *de-identification* is defined as the process of recognizing and deliberately changing, masking, replacing or concealing the names and/or other identifying information of relevance about entities. Identified information may be stored separately in an identification database. The linking between text and the identification database can be only made by a unique identifier, which makes the re-identification of the individuals extremely difficult without the use of an appropriate “key”.

3 Related Work

The “Scrub” system based on a set of detection algorithms utilizing word lists and templates that each detected a small number of name types in 275 pediatric records is presented by [3], which reports high rates on identified PHIs, 99-100%. In a similar system, [4] present comparable results. However, it is unclear in both studies what the recall figures or false positive rates were. A more elaborate de-identification system is presented by [5], using a variety of NLP tools. Each sentence found in a medical report was fed into a lexical analyzer which assigned syntactic and semantic information to each token. Rule-based pre-filters were then applied to eliminate non-name candidates. Over 99% precision and 93,9% recall figures are reported. In [6] a method based on lists of proper names and medical terms for finding and replacing those in pathology reports is presented. 98,7% correct identification on the narrative section and 92,7% on the entire report were reported. [7] describe the *k-anonymisation* approach, which de-associates attributes from the corresponding identifiers, each value of an attribute, such as date of birth, is suppressed, i.e. replacing entries with a “*”, or generalized, i.e. replacing all occurrences of, for instance, “070208”, “070209” etc. with “0702*. The interplay between anonymisation and evaluation within the framework of the *De-Id system* for surgical pathology reports is discussed by [8]. Three evaluations were conducted in turn and each time specific changes were suggested, improving the system’s performance. For a description of a number of methods for making data anonymous, see [9]. Finally, in the *Challenges in NLP for Clinical Data* workshop ([10]) there are details of systems participated in a shared task of automatic de-identification of medical summaries (e.g. age, phone, date, doctor).

4 Method

The NER system we use originates from the work conducted in the Nomen Nescio project between 2001-03, (*cf.* [2]). The system consists of five components, it is modular and scalable. The five components are:

- lists of multiword named entities
- a shallow parsing, rule-based component that uses finite-state grammars, one grammar for each type of entity recognized

- a module[1] that uses the annotations produced by the previous two components in order to make decisions regarding entities not covered by the previous two modules
- lists of single names (approx. 80 000)
- a revision/refinement module which makes a final control on an annotated document with entities in order to detect and resolve possible errors and assign new annotations based on existing ones, e.g. by combining annotation fragments.

Seven types of entities are recognized[2]: *persons*, *locations*, *organizations*, names of *drugs* and *diseases*, *time* and the *measure expressions* "age", "pressure" and "dosage". The annotation uses the XML identifiers ENAMEX, TIMEX and NUMEX. Each identifier contains an attribute that further specifies the entity; for details see [1].

The lack of annotated data in the domain prohibits us from using, and thus training, a statistically based system. Only minor parts of the generic NER system have been modified. These modifications dealt with: i) multiword place entities with the designators 'VC', 'VåC', 'Vårdc' and 'Vårdcentral' in attributive or predicative position, which all translate to *Health Care Center*, e.g. 'Åsa VC' or 'VåC Åsa' – which were inserted into the system; ii) insertion of the designators 'MAVA' *acute medical ward*, 'SS', 'SS/SU' and 'SS/Ö', where 'SS' stands as an acronym for the organization 'Sahlgrenska Sjukhuset' *Sahlgrenska Hospital* and iii) the use of medical terminology, particularly names of drugs (<www.fass.se>) and diseases, eponyms, (<mesh.kib.ki.se/>), in order to cover for a variety of names that conflict with regular person names. For instance, the drug name 'Lanzo' *lansoprazol* was part of the person's name database, while 'Sjögrens' in the context 'Sjögrens syndrom' *Sjogren's syndrome* could also be confused with frequent Swedish names. Therefore, the drug and disease modules are applied before the person/location in order to prohibit erroneous readings of such NEs. Examples of various NE types are for instance: *DSG[dosage]: 20 mg 1 x 1*; *LOC[country]: Somalia* and *MDD[disease]: Tourettes.* An example of annotated data (before [a] and after [b1-2] anonymisation) is given in below. The content of anonymised NEs can be translated into XML identifiers (b1) or to dummy characters (b2), e.g. "X" for capital letters, "x" for lower case characters, and "N" for numbers, while punctuation remains unchanged. The number of the dummy characters in each anonymised NE corresponds to the length of the original.

(a) Pat från <ENAMEX TYPE="LOC">Somalia</ENAMEX> op <TIMEX TYPE="TME">-91</TIMEX> med [...] får <ENAMEX TYPE="MDC">Waran</ENAMEX> [...] <ENAMEX TYPE="PRS">dr Steffan A. Janson</ENAMEX> rekommenderar biopsi […]
(b1) Pat från <COUNTRY> op <YEAR> med [...] får Waran [...] <PERSON{dr}> rekommenderar biopsi [...]
(b2) Pat från <COUNTRY>Xxxxxxx</COUNTRY> op <TIME>-NN</TIME> med [...] får Waran [...] <PERSON>dr Xxxxxxx X. Xxxxxx</PERSON> rekommenderar biopsi [...]

[1] The module is inspired by the *document centred approach* by [11]. This is a form of on-line learning from documents under processing which looks at unambiguous usages for assigning annotations in ambiguous words. A similar method has been also used by [12], called *labelled consistency* for de-identification of PHIs. This module has not used in the current work, since we applied bulk annotation on a very large sample, while this module has best performance in single, coherent articles.

[2] These name categories are a subset of the original system which also covers three more entities, namely *artifacts*, *work&art* and *events* (e.g. names of conferences).

5 A Corpus of Clinical Data and Evaluation

In this study we used as a medical free text a large corpus (~1GB) of discharged letters extracted from the EHR system MELIOR©. The corpus consists of database posts taken from tables of special interest for research such as “clinical history” and “final diagnoses”. The subcorpus we used for the evaluation consists of 200 randomly extracted passages from this corpus, which we believe gives a good indication of the performance of the NER system. A passage may consist of one or more sentences. The size of the evaluation material was 14,000 tokens. The only pre-processing of the texts has been the tokenization, basic separation of punctuation from the surrounding words, while the evaluation work was conducted on a locally installed version of the system at the department of Clinical Physiology, at the Sahlgrenska/Östra University Hospital in Gothenburg, behind a firewall in a secure environment.

For the evaluation we manually examined the selected sample. We calculated precision, recall and f-score using the formulas: *P = (Total Corr. + Partially Corr.) / All Produced* and *R = (Total Corr. + Partially Corr.) / All Possible*. Partially correct means that an annotation gets partial credit, e.g., if the system produces an annotation for ‘Alzheimers sjukdom’ *Alzheimer's disease* as *<ENAMEX TYPE=”MDD”> Alzheimers</ENAMEX> sjukdom*, instead of *<ENAMEX TYPE=”MDD”> Alzheimers sjukdom</ENAMEX>*, then such annotations are given a half point, instead of a perfect score. F-score is calculated as: *F=2*P*R/P+R*. The error analysis conducted indicated that the performance of the generic NER system is influenced by the features of the texts used for the evaluation. We emphasize the word “generic”, since simple means can increase the P&R figures, for instance, the majority of unmarked temporals were of the form “Number/Number –Number” (1/7 -00), very characteristic of the data and not part of the Swedish standard for time expressions, which the system leaves unannotated. The analysis of the results, particularly for the cases that the system failed to produce an annotation (insufficient coverage) or when the annotation was erroneous, revealed that errors where due: i) spelling and ungrammatical constructions (e.g. ‘ischaemi’ – instead of ‘ischemi’), ii) insufficient context/short sentences (e.g. ‘ACB-op -94’ – by-pass operation 1994) and iii) abbreviations (e.g. ‘epitel-ca’ – squamous cell cancer – instead of ‘epitel-cancer’).

Table 1. Evaluation results

Entity	Correct	Partial	Missed	Wrong	Precision	Recall	F-score
Person	44	-	2	2	95,65%	95,65%	95,65%
Location	15	2	10	-	94,11%	59,25%	72,71%
Organization	5	2	-	3	60%	85,71%	70,59%
Time	146	3	45	-	98,99%	76,03%	86%
Measure	491	8	37	-	99,19%	93,75%	96,39%
Diseases	187	8	25	-	97,94%	86,81%	92,03%
Drugs	562	58	18	1	95,16%	92,63%	93,82%
Total	1450	81	137	6	96,97%	89,35%	93%

6 Conclusions

In this paper we have described a system for anonymising hospital discharge letters using a generic NER system slightly modified in order to cope with some frequent characteristic features of the domain. The coverage and results of our approach (avg. f-score of 93%) provides a way for accessing the content of clinical free text in a manner that enables one to draw inferences without violating the privacy of individuals, although some work still remains. For the near future, we intend to evaluate a larger sample and propose adjustments to the system for even increased performance and also get the appropriate approval from the university hospital's ethical committee, for releasing some of the data for further research.

Acknowledgments. This work has been partially supported by the "Semantic Interoperability and Data Mining in Biomedicine", a Network of Excellence (NoE) funded by the European Commission under Framework 6.

References

1. Kokkinakis, D.: Reducing the Effect of Name Explosion. LREC Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP tasks. Portugal (2004)
2. Bondi Johannessen, J., et al.: Named Entity Recognition for the Mainland Scandinavian Languages. Literary and Linguistic Computing 20, 1 (2005)
3. Sweeney, L.: Replacing Personally-Identifying Information in Medical Records, the Scrub System. J. of the Am Med Informatics Assoc. Washington, DC, 333–337 (1996)
4. Ruch, P., et al.: Medical Document Anonymisation with a Semantic Lexicon. J Am Med Inform Assoc (Symposium Suppl), 729–733 (2000)
5. Taira, R.K, Bui, A.A., Kangarloo, H.: Identification of Patient Name References within Medical Documents Using Semantic Selectional Restrictions. In: AMIA Symposium. pp. 757–61 (2002)
6. Thomas, S.M., Mamlin, B., Schadow, G., McDonald, C.: A Successful Technique for Removing Names in Pathology Reports Using an Augmented Search and Replace Method. In: AMIA Symposium, pp. 777–781 (2002)
7. Sweeney, L.: k-anonymity: a Model for Protecting Privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10(5), 557–570 (2002)
8. Gupta, D., Saul, M., Gilbertson, J.: Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research. Am J of Clin Pathology 121(6), 176–186 (2004)
9. Hsinchun, C., Fuller, S.S., Friedman, C., Hersh, W.: Medical Informatics – Knowledge Management and Data Mining in Biomedicine, pp. 109–121. Springer, Heidelberg (2005)
10. Uzuner, O., Kohane, I., Szolovits, P.: Challenges in Natural Language Processing for Clinical Data Workshop (2006) www.i2b2.org/NLP/Schedule-final.pdf
11. Mikheev, A., Moens, M., Grover, C.: Named Entity Recognition without Gazeteers. In: Proc. of the 9th European Assoc. of Computational Linguistics (EACL), Norway, pp. 1–8 (1999)
12. Aramaki, E., Imai, T., Miyo, K., Ohe, K.: Automatic Deidentification by using Sentence Features and Label Consistency. Challenges in NLP for Clinical Data. Washing. DC (2006)

MetaCoDe: A Lightweight UMLS Mapping Tool

Thierry Delbecque[1] and Pierre Zweigenbaum[1,2]

[1] LIMSI-CNRS, Orsay, France
[2] INALCO, CRIM, Paris, France
{thd,pz}@limsi.fr

Abstract. In the course of our current research on automatic information extraction from medical electronic literature, we have been facing the need to map big corpora onto the concepts of the UMLS Metathesaurus, both in French and in English. In order to meet our specific needs in terms of processing speed, we have developed a lightweight UMLS tagger, MetaCoDe, that processes large text collections at an acceptable speed, but at the cost of the sophistication of the treatments. In this paper, we describe MetaCoDe and evaluate its quality, allowing potential users to balance the gain in speed against the loss in quality.

Keywords: Information Extraction, Information Retrieval, Natural Language Processing, Medical Terminologies, UMLS, Concept Mapping, MetaMap, MMTx.

1 Introduction

Recognizing and extracting medical concepts from free text is an essential task in a number of activities such as article indexing, health record coding, text mining on medical corpora, automatic knowledge discovery from medical literature, and so on [1,2,3]. In the context of some of our current research related to Question Answering (QA) systems, we need to quickly extract concepts of the extensive UMLS Metathesaurus [4] from large text corpora gathered from various medical Web sites. Preliminary experiments using the automated UMLS indexer MetaMap Transfert (MMTx) [5] led to too long computing times and too much memory usage, and motivated us to write a light UMLS tagger (MetaCoDe), that could run using less memory and proceed in a shorter time. The saved time could then be devoted to specific QA-oriented computations. This short paper is devoted to the description of the MetaCoDe algorithm (section 2). It also gives some evaluation measures of the quality of the concept mappings it produces, taking MetaMap's as a reference (section 3).

2 MetaCoDe Algorithm

2.1 Overview

In the following, CUI stands for the identifier of a concept in the Metathesaurus, and SUI stands for the identifier of a string in the Metathesaurus. We refer the

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 242–246, 2007.

reader to [6,4,5] for more information on the Metathesaurus architecture, and on the linguistic tools offered by the NLM including MetaMap.

MetaCoDe involves the chaining of a set of distinct operations, as illustrated in figure 1. The first bunch of operations is to prepare both the corpus and the necessary terminological resources (preliminary tasks); the last operation is the mapping itself. The principle of first tailoring general resources to a given corpus is used, *e.g.*, by the Unitex corpus processor [7].

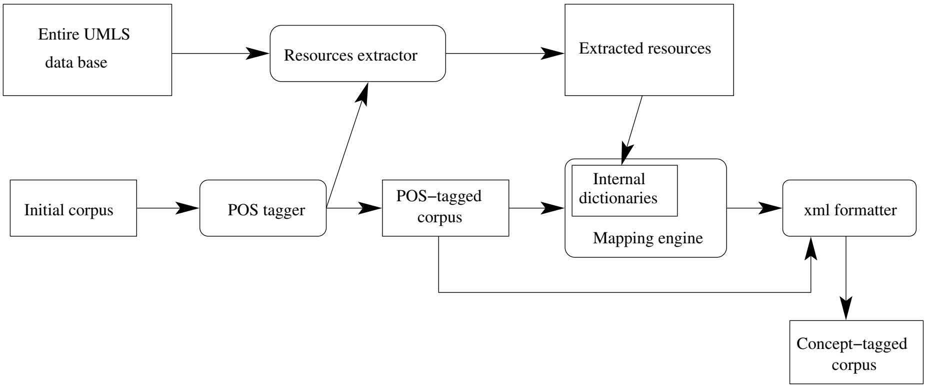


Fig. 1. MetaCoDe overall process

2.2 Preliminary Tasks

The preliminary operations consist of the following tasks:

- The input text file is tokenized, then analyzed with the TreeTagger to determine the part of speech and lemma of each token. In the remainder of the treatment, only this categorial information (no inflectional information) is used;
- A list of the 'content' terms (original form + lemmatized form) encountered in the POS-tagged file is built: 'content' term is defined as having its category in {NOUN, PROPER NAME, ADJECTIVE, ADVERB, GERUNDIVE};
- Using both this content term list and the English Metathesaurus word index (or the French one, if the mapping is applied to a French text), four resource files are created:
  WDSUI: mapping from individual word to SUI strings using this word;
  CUISUI: mapping from SUI to related CUI;
  SUILENGTH: gives the length (token number) of each SUI;
  CUISTY: gives the semantic class of a CUI.

Each of these files is made as small as possible, as they are to be loaded into memory during the tagging process.

2.3 The Mapping Algorithm

During its initialization, the mapping algorithm loads the content of the previously built resource files into internal hash tables. Then the following is done:

- The text is parsed in order to extract noun phrases. This parsing step relies on part-of-speech patterns. As different parametrisations of TreeTagger may yield different tags, part-of-speech patterns are described in an external parameter file; this also allows us to keep language-specific parameters outside of the program;
- For each noun phrase $\mathcal{F}$ the following is done:
  - for each word $\mathcal{W}$ of $\mathcal{F}$, the algorithm finds all SUIs in which $\mathcal{W}$ or its lemma occurs. This gives rise to a set of SUIs, $\{s_1, \cdots, s_n\}$, each being associated with a subset W_i of words from $\mathcal{F}$. This set of SUIs thus receives a lattice structure: the one organizing the subsets;
  - this lattice is pruned, by removing from it all the SUIs s for which the string length is greater than $|W_s|$ plus a given amount; this prevents the algorithm from selecting UMLS strings that would be specializations of what is actually expressed in the input noun phrase;
  - the maximal elements of the pruned lattice are selected, giving a first candidate set C_I; From C_I the algorithm builds a set $\{e_1, \cdots, e_n\} = C'_I$ such that each W_{e_i} and W_{e_j} are disjoint, and $|W_{e_i}|$ are maximal.
  - using the internal dictionaries, each CUI, then each semantic type, is obtained for each SUI of C'_I.

At this point, it is clear that the main simplification of MetaCoDe, compared with MetaMap, lies in the fact that MetaCoDe lacks any variant generator [5]. As a matter of fact, MetaCoDe relies on the terminological variability naturally occurring in the Metathesaurus (since several SUIs are frequently associated with a common CUI) to compensate for this lack, a small extra level of variation being offered thanks to the use of lemmas. The evaluation aims at measuring the consequences of this approach.

3 Evaluation

3.1 Method

Mapping time was evaluated on a corpus of 7,260 medical abstracts extracted from MEDLINE, accounting for 2,160,613 words.

Mapping quality evaluation was measured on a random sample of 30 independent medical abstracts issued from MEDLINE, accounting for a total of about 9,200 words. To perform this evaluation, both MetaCode and MetaMap were applied to this corpus[1]. We differentiated the following cases, using the output of MetaMap as a gold standard:

MATCH: given a noun phrase, or part of a noun phrase, a correct mapping onto an UMLS concept was proposed by MetaCoDe;

[1] Parameter settings: –ms300m –mx900m, all other parameters at default values.

Table 1. (a) Raw counts of decision categories and (b) derived measures

(a)

MATCH	AMB	BAD	MISS
1,300	397	93	307

(b)

$\tilde{P}$	$\tilde{R}$	$\tilde{F}$	Amb.Rate = AMB/MATCH
0.93	0.76	0.83	0.31

AMBIGUOUS: given a noun phrase, or part of a noun phrase, a correct mapping was proposed, but at the same time incorrect propositions were also generated. For instance, the term *study* can be mapped both to C0008972 (Research Activity) and C0557651 (Manufactured Object). It is clear that $|AMBIGUOUS| \leq |MATCH|$ (by definition, each AMBIGUOUS case is also counted as MATCH);

BAD: given a noun phrase or a part of a noun phrase, MetaMap proposed a mapping, but MetaCode only proposed different mappings;

MISS: a concept was identified by MetaMap, but no concept was proposed by MetaCoDe.

According to these definitions, we have built the following measures:

$$\tilde{P} = \frac{|MATCH|}{|MATCH|+|BAD|};\ \tilde{R} = \frac{|MATCH|}{|MATCH|+|MISS|+|BAD|};\ \tilde{F} = \frac{2\tilde{P}\tilde{R}}{\tilde{P}+\tilde{R}}$$

the notations $\tilde{P}$, $\tilde{R}$ and $\tilde{F}$ emphasizing that they are respectively pseudo-precision, pseudo-recall, and pseudo F-measure, as they are dependent on the behaviour of MetaMap. The rationale behind these definitions lies in the difficulty to assess anything else than a true positive by simply querying the UMLS database.

3.2 Results

MetaCoDe has been developed in PERL and C++ using Microsoft tools, but it could easily be ported to UNIX systems. For efficiency reasons, the mapping procedure has been written in C++ using the Standard Template Library (STL) for representing internal data structures. The algorithm was kept simple: the total number of C++ lines is only about 750, making it easy for anyone to adapt it to their own needs. Version 2006AA of the UMLS was used. It contains 4 million English strings (SUIs) associated to 1.3 million CUIs.

The total mapping time spent by MetaCoDe on the larger corpus was 26 minutes, including initialisation, where MMTx needed 50 hours[2]. Initialisation restricted resources to 1 million CUIs and 2.6 million SUIs.

Quality was measured on the smaller, 9,200 word corpus, with results shown on table 1. These figures do not measure the absolute mapping quality of MetaCoDe, but the decrease in mapping quality relative to a reference tool, MetaMap, considered in our study as a gold standard. More precisely, our work can be considered as a study about the consequences of not using variant generation

[2] Both on a Pentium 4 biprocessor, 3GHz, 1.5 Go; no other task running.

algorithms, relying instead only on the capacity of the Metathesaurus by itself to propose variants.

The low decrease in 'precision' ($\tilde{P} = 0.93$) is inherent to our algorithm, which is rather conservative. The true consequence of not using variant generation is mostly reflected by the pseudo-recall indicator ($\tilde{R} = 0.76$). Unsurprisingly, the decrease in 'recall' is significant: MetaMap, but not MetaCode, could detect some concepts under variant forms that were not listed in the UMLS. Nevertheless, it is not as dramatic as one might have expected. Furthermore, easy improvements of the algorithm can be achieved by simply modifiying the pattern set used to build noun phrases, *e.g.*, by including isolated verbal forms as MetaMap does.

4 Conclusion

MetaCoDe is a rather rudimentary tool, that may be improved easily:

- to be made more parameterisable, *e.g.*, to choose the target vocabulary;
- to deliver a score for the mapping propositions it produces, as MetaMap.

In spite of these drawbacks, it has some positive features that one could consider:

- it runs quickly, saving time for additional treatments;
- it is a small C++ program, still rather simple to modify and improve.

Therefore MetaCoDe should have value in a context where very large corpora must be processed in a reasonable time and the target task is not hampered by a small decrease in precision and recall. Such an approach was exemplified in [8], where better results were obtained by less complex processes run on larger corpora. The source files of MetaCoDe and of extra tools such as a PERL/Tk corpus browser are freely available upon request to the authors.

References

1. Rindflesch, T.C., Aronson, A.R.: Semantic processing in information retrieval. In: Safran, C. (ed.) Proc Annu Symp Comput Appl Med Care, pp. 611–615 (1993)
2. Bodenreider, O., Nelson, S.J., Hole, W.T., Chang, F.H.: Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. In: Proc AMIA Symp., pp. 815–819 (1998)
3. Aronson, A.R., Mork, J.G., Gay, C.W., Humphreys, S.M., Rogers, W.J.: The NLM Indexing Initiative's Medical Text Indexer. Medinfo. 2004 11(Pt. 1), 268–272 (2004)
4. Humphreys, B.L., Lindberg, D.A., Schoolman, H.M., Barnett, G.: The Unified Medical Language System: An informatic research collaboration. J Am Med Inform Assoc 5(1), 1–11 (1998)
5. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proc AMIA Symp., pp. 17–21 (2001)
6. Lindberg, D.A., Humphreys, B.L., McCray, A.T.: The Unified Medical Language System. Methods Inf Med 32(2), 81–291 (1993)
7. Paumier, S.: Unitex 1.2 User Manual (2006) http://www-igm.univ-mlv.fr/~unitex/UnitexManual.pdf
8. Curran, J.R., Moens, M.: Scaling context space. In: Proc 38th ACL, pp. 231–238 (2002)

Unsupervised Documents Categorization Using New Threshold-Sensitive Weighting Technique

Frederic Ehrler[1,2] and Patrick Ruch[2]

[1] Artificial Intelligence laboratory, University of Geneva, Geneva Switzerland
[2] Medical Informatics Services, University Hospital of Geneva, Geneva Switzerland
Frederic.ehrler@cui.unige.ch

Abstract. As the number of published documents increase quickly, there is a crucial need for fast and sensitive categorization methods to manage the produced information. In this paper, we focused on the categorization of biomedical documents with concepts of the Gene Ontology, an ontology dedicated to gene description. Our approach discovers associations between the predefined concepts and the documents using string matching techniques. The assignations are ranked according to a score computed given several strategies. The effects of these different scoring strategies on the categorization effectiveness are evaluated. More especially a new weighting technique based on term frequency is presented. This new weighting technique improves the categorization effectiveness on most of the experiment performed. This paper shows that a cleaver use of the frequency can bring substantial benefits when performing automatic categorization on large collection of documents.

Keywords: Unsupervised categorization, Gene Ontology.

1 Introduction

The emergence of high-throughput methods for acquiring information about the sequences, expressions and functions of genes has provided an abundance of valuable new data. The unclassified documents have little chance to be retrieved when needed. Automatic classification methods attempt to reproduce human judgments in order to assign categories to documents. Tasks like browsing and research of documents are then simplified for the users of documents collections.

The method presented in this paper consists in using string matching technique to extract assignations between concepts extracted from the Gene Ontology (GO) and the sentences that compose the documents. The properties of the generated assignations are used to compute a score. This score allows ranking the assignations to select only the most interesting ones. In our experiment, different more or less conservative matching techniques are compared. Moreover, the variation of effectiveness brought by the modification of the ranking consequently to the use of different scoring strategies is analyzed.

A similar problem of categorization has been tackle during the "Bio Creative[1]" workshop in 2004 [1,2,3]. The best system hasn't outperformed a recall of 12%.

[1] http://biocreative.sourceforge.net/

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 247–251, 2007.

Among the diverse approaches adopted, three main strategies can be characterized. Those based on pattern matching [3], those based on machine learning technique [1] and those based on template extraction.

The plan of the paper is the following: The datasets and the evaluation metric are presented in section 2. The section 3 describes the different matching and scoring strategies employed to generate the ranked list of categories. The section 4 covers the experiments followed by the obtained results and their explanation. Finally the conclusion is cover in section 5.

2 Data and Metrics

Gene Ontology merges three structured vocabularies, organized as ontology describing gene products in terms of their associated biological process cellular component and molecular function in a species-independent manner [4]. The used version of the GO (December 2003) is composed by 16685 terms [5,6].

One way to evaluate the effectiveness is to use the precision and recall metric. The former metric measure the proportion of relevant categories retrieved and the latter measure the proportion of retrieved categories that are relevant. Usually to decide whether a category is relevant, we look if the score related to this category is above a threshold. However, in our case, instead of using the threshold on the score, we fix a number of requested categories and we keep the best ones up to reach this limit.

3 Methods

The categorization procedure is performed in several steps. First, the documents are split into sentences. Then, each time an occurrence of a word belonging to one of the concepts is found in a sentence, an assignation is created. For every assignation, a score is computed in order to rank them and select the most promising ones. The different steps of the procedure are explained with more details in the following subsections.

3.1 The Matching Technique

Every word belonging to a concept occurring in a sentence identifies a possible assignation. The matching technique presented below concerns the different methods, which have been used to detect the matches between the concept and the sentence.

- **Exact matching:** This technique searches in the sentences for exact occurrences of the words that compose the concepts. This kind of search support no spelling variation between the searched terms and the terms found.
- **Sub-string matching:** Contrarily to exact matching, with sub-string matching there is no need to find the exact words of the concept in the sentence to create a match, an ordered characters sequence of these words is sufficient.
- **Stemmed matching:** This matching procedure preliminary stem the words that compose the concepts and the sentences. Then it looks for the similar word to apply a match. The stemming method used is the Porter stemmer [7,8].

3.2 Score Computation

The score of an association is composed of two subparts. First, a matching score proportional to the number of words shared between the concept and the sentence that compose the association. Second, a score refinement, that gives a variable importance to the discovered association depending of the chosen strategy. The overall score is used to rank the assignations and identify which ones are the most relevant.

The matching score of an association between a concept *C* composed of |*C*| terms and a sentence S is dependant of two values, the proportion of words of the concept found in the associated sentence, and the length of the target concept. Taking the length of the concept into account allows, given a same proportion of match, to advantages the association containing the longest concepts.

The score refinement denotes the use of additional information to perform a re-ranking of the assignations in order to improve the retrieval effectiveness at top ranks. There are several options presented below.

- **The position refinement:** Favor the assignations whose the words of the concept are found in a sequential order in the sentence.
- **The frequency based weighting refinement:** Modify the importance of the words given their frequency in the concepts. When applied this formula gives a low weight to words having a high frequency in the controlled vocabulary and a high weight to rare words.
- **The trigger refinement:** In specific situation, using word frequency to differentiate between relevant and irrelevant assignations can have undesirable outcome. Indeed, a concept mostly composed by frequent words is not likely to obtain a good score even if all its words are found exactly in a sentence. As every word of the concept possesses a low weight, the total score will be also low. As we consider that such a concept must have a higher rank, we tried to modify the ranking by introducing the trigger refinement. *The trigger refinement is a way to avoid to be polluted by associations based only on infrequent word, but without losing the importance of associations where infrequent words co-occur with more important words.* We define an un-triggered word as a word which is insufficient to trigger an assignation between a concept and a sentence, therefore no assignations are created if only on un-triggered words are found in the sentence. However, if any un-trigger word is attended with a trigger words, the un-trigger word is considerate as a word of weight 1. The difference with traditional weight is that the triggered strategy is sometime more restrictive and sometime more generous. The trigger property of a word is decided given the document frequency of the word in the collection. The words that have a number of occurrences above an experimentally chosen threshold are considered as insufficient to trigger a match if they are found alone.

4 Experiments

The test set is composed 1642 associations between a GO concept and one of the 640 documents, which arise from the “Journal of Biological Chemistry”. We haven’t consider the full-text but only the abstract as we assume that the abstract contains most of the important information and allow to be focused on the key concepts.

One present below the recall obtained at the breakeven point (precision equals recall) and the recall obtained when the 10 best categories are returned for each axis of the ontology. The recall at breakeven point allows comparing our results with those obtained at BioCreative and tests the capacity of the tool to obtain a good recall at high rank. The recall obtained with the ten best results for each axis of the ontology allows identifying a reasonable upper bound.

Table 1. Recall obtained at breakeven point (recall equal precision) and recall obtained when the 10 best categories are returned for each axes

Match/ computer	Exact		Sub-string		Stemmed	
	BRK	10 Best	BRK	10 Best	BRK	10 Best
Basic	20.4%	44.1%	20.6%	47.3%	20.8%	47.3%
Position	17.4%	43.4%	18.3%	43.1%	18.2%	45.1%
Triggered	22.4%	45.7%	21.4%	47.7%	21.8%	47.5%
Weighted	11.9%	36.4%	11.6%	33.9%	12.2%	36.3%

At breakeven point, the best recall (22.4%, with 367 correct assignations returned) is obtained by the combination of the perfect match method with the triggered refinement. The results obtained using perfect match are the best and those using the sub-string matching technique are the worst. This is not surprising, as discussed in the presentation of the matching techniques, "perfect match" is the matching method that gives the best precision, and "sub-string matching" is the one that gives the worst one.

In the second experiment conjointly to the global increase of recall there is a dramatic loss of precision (4% in the best case). The use of the sub-string matching method in association with the trigger refinement allows reaching the best recall (48.3% with 793 correct assignations returned). The best results are given by the use of "sub-string matching" method and the worst are obtained by the use of "perfect matching" method. This is once more expected as the fuzzy method is able to generate many possibly relevant associations even with weak clues and therefore gives a high recall. On the other hand, the "precise matching" method misses some relevant assignations dues to its incapacity to deal with small differences of spelling between the vocabulary used in the concepts and the one used in the sentences.

The results of this second experiment demonstrate that even with a very liberal method, 50% of the concepts remain undetected. Additional resources could perhaps be used to extend the words related to every concept in order to increase the number of matches and in the same manner the number of possible candidates.

In the two experiments, the use of the trigger strategy leads always to the best performance. Whereas with simple frequency based weighting the associations composed of concepts containing frequent words are not likely to obtain a good score, the trigger strategy is crafty enough to give when needed a significant importance to the high frequent words in order to increase the total score of the association.

5 Conclusion

A fully modifiable tool has been built in order to evaluate different strategies of categorization based on matching techniques. Several matching techniques and several scoring strategies have been combined and their effects on the performance have been evaluated. One of the main discoveries is that the use of trigger words over performs the other tested strategies.

The good results obtained by trigger strategy confirm that words with high document frequency are not apposite to discover relevant associations. As frequent words are shared by different concepts, they don't bring enough information to discriminate between the correct and the wrong associations. However, when high frequency words are found conjointly with other words they must not be underestimated as it is usually done with the more traditional frequency weighting strategy.

Using such method is an easy way to improve the performance of categorization without increasing the complexity of the process.

References

1. Rice, S., Nenadic, G., Stapley, B.: Protein function assignment using term-based support vector machines – bioCreative Task Two 203. BioCreative NoteBook Papers (2004)
2. Krallinger, M., Padron, M.: Prediction of GO annotation by combining entity specific sentence sliding window profiles. BioCreative NoteBook Papers (2004)
3. Ruch, P., Chichester, C., Cohen, G., Ehrler, F., Fabry, P.J.M., Muller, H., Geissbuhler, A.: A Report on the TREC 2004 Experiment: Genomics Track. The Thirteenth Text Retrieval Conference, TREC-2004, Gaithersburg, MD (2004)
4. Goertzel, B., Goertzel, I., Pennachin, C., Looks, M., Queiroz, M., Prosdocimi, F., Lobo, F.: Inferring Gene Ontology Category Membership via Cross-Experiment Gene Expression Data Analysis
5. The Gene Ontology Consortium: Creating the Gene Ontology Resource: Design and Implementation. Genome Res 11, 1425–1433 (2001)
6. MCCray, A., Brown, A., Bodenreider, O.: The lexical Properties of the Gene Ontology. In: AMIA Annual Symposium, pp. 504–508 (2002)
7. Porter, M.F.: An algorithm for suffix stripping. Program 14, 130–137 (1997)
8. Kraaij, W., Pohlmann, R.: Viewing Stemming as Recall Enhancement. In: 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 40–48 (1996)

Application of Cross-Language Criteria for the Automatic Distinction of Expert and Non Expert Online Health Documents

Natalia Grabar[1,2] and Sonia Krivine[3]

[1] INSERM, UMR_S 729, Eq. 20, Paris, F-75006 France
[2] Health on the Net Foundation, SIM/HUG, Geneva, Switzerland
[3] FircoSoft, 37 rue de Lyon, 75012 Paris, France
natalia.grabar@spim.jussieu.fr, sonia.krivine@free.fr

Abstract. Distinction between expert and non expert documents is an important issue in the medical area, for instance in the context of information retrieval. In our work we address this issue through stylistic corpus analysis and application of machine learning algorithms. Our hypothesis is that this distinction can be observed on the basis of a little number of criteria and that such criteria can be language and domain independent. The used criteria have been acquired in source corpus (Russian) and then tested on source and target (French) corpora. The method shows up to 90% precision and 93% recall, and 85% precision and 74% recall in source and target corpora.

1 Introduction

Medical information searchable online presents various technical and scientific content but this situation is not clear for non expert users. As a matter of fact, when reading documents with high technical content non expert users can have some understanding problems, because they are anxious, pressed or unfamiliar with the health topic. This situation can have a direct impact on users' well-being, their healthcare or communication with medical professionals. For this reason, search engines should distinguish documents according to whether they are written for medical experts or non expert users. Distinction between expert and non expert documents is closely related to the health literacy [1], and the causal effect it can have on healthcare [2]. For the definition of the readability level, several formulae have been proposed (*i.e.*, Flesch [3], Fog [4]), which rely on criteria like average length of words, sentences and number of difficult words. Distinction between expert and non expert documents can also be addressed through algorithms proposed by the area of text categorisation and applied to various features: Decision Tree and Naive Baayes applied to manually weighted MeSH terms [5]; TextCat[1] tool applied to *n-grams* of characters [6]; SVM applied to a combination of various features [7].

[1] *www.let.rug.nl/∼vannoord/TextCat*

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 252–256, 2007.

In our work, we aim at applying machine learning algorithms to corpora which gather documents from different languages and domains. To ease this process, we propose to use a little set of features, which would be easy to define and to apply to a new language or domain. Aimed features should be shared by different languages and domains. Assuming that documents represent the context of their creation and usage through both their content and style, we propose to set features at the stylistic level. Features are thus defined on the basis of the source corpus and then applied to the target corpus. Languages and domains of these two corpora are different. The cross-domain and especially cross-language aspect of features seems to be a new issue in the text categorisation area.

2 Material and Method

Working languages are Russian (source language) and French (target language). Corpora are collected online: through general engines in Russian and the specific medical search engine CISMeF in French. In Russian, the keywords used are related to *diabetis and diet*, and the distinction between expert and non expert dociments is performed manually. The French search engine already proposes this distinction and we exploit it in our work. We used keyword *pneumologie* (*pneumology*) when querying CISMeF. Table 1 indicates size and composition of corpora in both studied languages. The French corpus contains more documents, which is certainly due to the current Internet situation. Moreover, we can observe difference between sizes of expert and non expert corpora: the non expert corpus is bigger in Russian, while the expert corpus is bigger in French.

Table 1. Expert and non expert corpora in Russian and French languages

	Russian		French	
	nbDoc	occ	nbDoc	occ
Expert documents	35	116’000	186	371’045
Non expert documents	133	190’000	80	87’177
Total	168	306’000	266	458’222

The objective of our work is to develop tools for categorising health documents according to whether they are expert or non expert oriented. We use several machine learning algorithms (`Naive Bayes`, `J48`, `RandomForest`, `OneR` and `KStar`) within Weka[2] tool in order to compare their performances and to check the consistency of the feature set. The main challenge of the method relies on the universality of the proposed features defined on the basis of source language (Russian) and domain (diabetis) and then applied to target language (French) and domain (pneumonology).

Stylistic features have emerged from a previous contrastive study of expert and non expert corpora in Russian [8] realised with lexicometric tools. For the current

[2] Weka (*Waikato Environment for knowledge analysis*), developped at University Waikato, New-Zeland, is freely available on *www.cs.waikato.ac.nz/∼ml/index.html*

Table 2. Evaluation of algorithms on source and target corpora

Method	Expert		Non expert		Method	Expert		Non expert	
	Prec.	Recall	Prec.	Recall		Prec.	Recall	Prec.	Recall
NaveBayes	43	83	94	72	NaveBayes	93	36	31	91
J48	**83**	**42**	86	98	J48	81	83	43	41
RandomForest	**83**	**42**	86	98	RandomForest	**87**	**81**	**52**	**64**
OneR	43	25	82	91	OneR	83	87	53	45
KStar	70	58	**90**	**93**	KStar	85	74	42	59

work, we selected a set of 14 features related to the document structure, and marks of persons, punctuation and uncertainty. Learning and test corpora are composed of respectively 66% and 33% of the whole corpora collected. Evaluation is done on independent corpus through classical measures: precision, recall, F-measure and error rate.

3 Results and Discussion

Results obtained on the Russian corpus are presented in the left part of table 2. For each method (first column), we indicate figures of precision and recall. KStar shows the best results with *non expert* documents: 90% precision and 93% recall, and nearly the best results for the *scientific* category: 70% precision and 58% recall. J48 and RandomForest, both using decision trees, present identical results for two studied categories: 83% precision and 42% recall for *scientific* documents and 86% precision and 98% recall for *non expert* documents. From the point of view of precision, these two algorithms are suitable for the categorisation of documents as *scientific*. The right part of table 2 indicates evaluation results of the same algorithms applied to the French corpus (175 documents for learning and 91 for test). RandomForest has generated the most competitive results for both categories (*expert* and *non expert*). Surprisingly, OneR, based on the selection of only one rule, produced results which are close to those of RandomForest. As general remark, scientific documents are better categorised in French and non expert documents in Russian, which is certainly due to a larger size of corresponding data in each language. Low performances of NaveBayes in both languages seem to indicate that the Bayes model, and specifically its underlying hypothesis on independance of criteria, is too naive for the task of classification of documents as expert and non expert oriented. Whereas, we assume that stylistic and discourse criteria equally participate in the encoding of stylistic specificities of medical documents [9,10].

Language model. We could analyse two language models, generated by OneR and J48 algorithms. OneR selects one (best) rule in each corpus. In our experiment, this algorithm selected hypertext link <a> tag in Russian and 2^{nd} plural pronoun in French. These features allow to produce nearly the best results in the target corpus (French), while in Russian this algorithm is the less competitive. The model produced by J48 in Russian selects hypertext link <a> tags together with 1st singular pronoun я (*I*), italic characters (tag <i>), lists (<ol>) and table (<table>) tags. On French corpora, J48 selects the following five criteria: 2nd plural pronoun, <table> tag, 2nd singular pronoun, <ol> tag and exclamation mark. J48 is one of the most suitable algorithms in Russian but it shows average performances in French. Surprisingly, a majority of the most relevant criteria are related to the HTML tagging of documents but not to linguistic information. This observation seems to indicate that categorisation of web documents should be based also on non textual criteria. According to the theory of genres [11], this observation emphasizes the importance of the layout of documents, their typography and intertextuality.

Analysis of errors common to various classifiers. Within Russian corpus, six documents are wrongly categorised by several algorithms. Their analysis indicates that these documents are ambiguous as for their categorisation, both manual and automatic, and that discourse distinction between expert and non expert document is set on a continuum axis. Thus, there is no dichotomy between these two categories and borderline documents are difficult to categorise.

Suitability of proposed features. The proposed reduced set of features contains 14 criteria related to the document structure, marks of persons, punctuation and uncertainty. The obtained results seem to indicate that these stylistic features are suitable for the categorisation of documents according to their discourse (*expert* and *non expert*). Indeed, their application to the target corpus shows promising performances, although the target corpus is composed of documents in a different language and describing different medical topics. Moreover, these features are easy to adapt to a new language. But their application to other corpora has to be verified. One of their limitations is that several of these features remain specific to HTML documents.

4 Conclusion and Perspectives

We have presented an experiment on automatic distinction of expert and non expert Web documents. For this, learning algorithms and a set of 14 stylistic criteria have been used. Criteria have been acquired on a source corpus (Russian language, diabetis related topic) and then applied to target (French language, pneumonology related topic) and source corpora. Evaluation results show that decision tree algorithms J48 and RandomForest are the most suitable for the categorisation of documents as expert and non expert. They generate the best results in the target corpus and for the *expert* category in the source corpus. As we have noticed, results depend on the size of learning corpora. It would be interesting to apply the system to a larger collection of documents and to confirm the stability of the acquired language models. But we can consider these results as promising, especially as documents are extracted from various websites and learning and test steps are performed on independent datasets.

The obtained results seem to indicate that the proposed stylistic features are suitable for the categorisation of documents according to their discourse: their application to the target corpus shows promising performances, although the target corpus is composed of documents in different languages and describing different medical topics. Nevertheless, it could be interesting to apply other criteria, for instance argumentation structures [12], for the distinction between expert and non expert documents.

We assume that categorisation results can be more precise. For instance, within the category *scientific* we can distinguish scientific articles and didactical material; and within *non expert* category we can distinguish cook recipes, articles for large audience and food recommandations. In French, we built an intermediate category composed of documents written for medical students: *courses*, *teaching material*. It could be interesting to categorise this material through the proposed language model. It could be interesting to apply our method to other medical areas and other types of documents (clinical), and to compare it with results produced by other approaches.

References

1. McCray, A.: Promoting health literacy. Journal of American Medical Informatics Association 12, 152–163 (2005)
2. AMA: Health literacy: report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. JAMA 281(6), 552–557 (1999)
3. Flesch, R.: A new readability yardstick. Journal of Applied Psychology 23, 221–233 (1948)
4. Gunning, R.: The art of clear writing. McGraw Hill, New York (1973)
5. Zheng, W., Milios, E., Watters, C.: Filtering for medical news items using a machine learning approach. In: AMIA, pp. 949–53 (2002)
6. Poprat, M., Markó, K., Hahn, U.: A language classifier that automatically divides medical documents for experts and health care consumers. In: MIE 2006 - Proceedings of the XX International Congress of the European Federation for Medical Informatics, Maastricht, pp. 503–508 (2006)
7. Wang, Y.: Automatic recognition of text difficulty from consumers health information. In: IEEE. (ed.) Computer-Based Medical Systems (2006)
8. Krivine, S., Tomimitsu, M., Grabar, N., Slodzian, M.: Relever des critères pour la distinction automatique entre les documents médicaux scientifiques et vulgarisés en russe et en japonais. In: TALN (2006)
9. Benveniste, E.: La nature des pronoms. Problémes de linguistique gnérale 1, 251–257 (1966)
10. Malrieu, D., Rastier, F.: Genres et variations morphosyntaxiques. Traitement automatique des langues 42, 548–577 (2001)
11. Genette, G.: Théorie des genres. Seuil, Paris (1986)
12. Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D., Lovis, C., Veuthey, A.: Using argumentation to extract key sentences from biomedical abstracts. Int J Med Inform (2006)

Extracting Specific Medical Data Using Semantic Structures

Kerstin Denecke[1,2] and Jochen Bernauer[3]

[1] Technical University Braunschweig, Mühlenpfordtstr. 23, D-38106 Braunschweig
[2] University of Hannover, Research Center L3S, Appelstr. 9a, D-30167 Hannover
[3] University of Applied Science, Prittwitzstr. 10, D-89075 Ulm
kdenecke@web.de, bernauer@fh-ulm.de

Abstract. In this paper, we discuss the architecture, functionality and performance of a medical information extraction system. The system is based on an approach to automatic generation of semantic structures for free-text. Using a multiaxial nomenclature (Wingert Nomenclature) and existing language-engineering technologies, a conceptual graph-like representation is produced for each sentence of a text. These semantic structures are then exploited to extract information. The components that might be adopted for processing texts in another language than German are identified. Results of first evaluations of the system's performance in an information extraction (IE) subtask in the medical domain are presented: The filling of selected template slots obtained values of 81- 95% precision and 83-97% recall.

Keywords: Information Extraction, Natural Language Understanding, Natural Language Processing.

1 Introduction

Supporting and describing the process of medical treatment is one of the most important roles of medical documentation. Different kinds of medical documents (e.g., discharge summaries, radiology reports etc.) are created in daily routine and require further interpretation and processing. However, the mostly unstructured textual format of those documents, their extent and their immense number makes it difficult and time consuming to process them manually. Medical natural language processing (NLP) addresses these issues: By transforming natural language text into standardised and normalised semantic structures, data, that is concealed in written text, is made accessible. It can be re-used by different applications such as quality assessment or decision support. Repeated documentation is avoided. Hence, the expenditure of time and money for documentation and retrieval purposes can be reduced.

Current approaches to medical NLP are often limited to a certain medical domain and their construction is fairly complicated. For most of the systems, a domain model or a knowledge base has been built up only for this purpose. E.g., LifeCode NLP System [5] is limited to the domain of radiology because its knowledge base has been created only for this purpose. In MedLEE [6], the Medical Entries Dictionary (MED) is re-used indeed but it was initially limited to the domain of radiology. In the

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 257–264, 2007.

meantime, MED has been extended continuously and MedLEE has been amplified to process documents of different domains. Other NLP systems are limited to a certain language: MedIE [9] and LifeCode NLP System can only process medical documents in English. Most implemented medical NLP systems achieve recall values between 80-85% and precision values between 95-99%.

The system introduced in this article aims to avoid the limitations of existing NLP systems. It is based on a method for automatically mapping natural language text to semantic structures. This method uses a multi-axial, multilingual terminology of medical terms. The built-up semantic structures are subsequently used for information extraction, i.e. to gather specific information in natural language text.

2 Architecture and Functionality

The system consists of a variety of components (illustrated in Fig. 1). These form three major modules: a Textpreprocessor, a Semantic Interpreter and an Extractor.

2.1 Textpreprocessor

The *Textpreprocessor* consists of a paragraph classifier, a preparser and a simple parser. First, a paragraph classifier dissects a document into several subsections and assigns each section to one of the classes that are defined for the different document types by means of regular expressions. Next, a preparser identifies the sentences of a text. On the basis of a shallow syntactic analysis provided by a simple parser, each sentence is then decomposed into segments using prepositions [7]. A segment is a smaller part of a sentence that reaches from the beginning of a sentence to the first preposition (head segment) or it starts with a preposition and reaches to the next preposition or to the end of the sentence (prepositional segment). E.g., the nominal phrase *Aufnahme des Patienten unter dem Verdacht einer Ureterkolik*

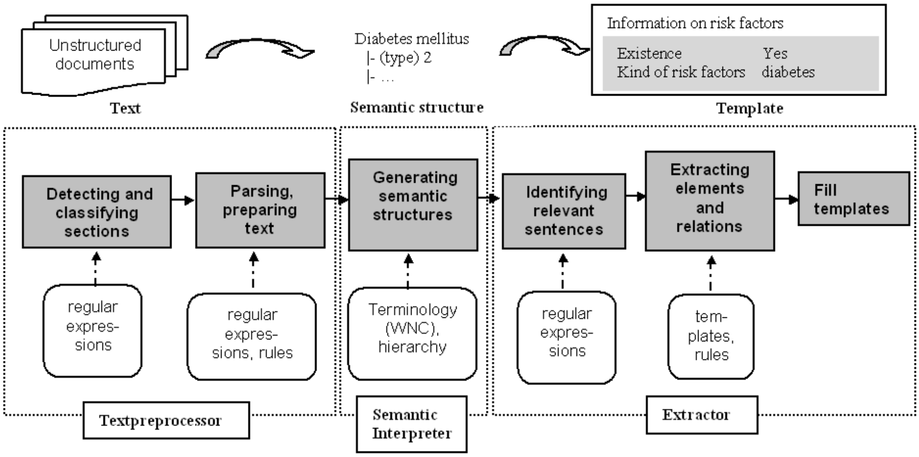

Unstructured documents
Text
Diabetes mellitus
|- (type) 2
|- ...
Semantic structure
Information on risk factors
Existence Yes
Kind of risk factors diabetes
Template
Detecting and classifying sections
Parsing, preparing text
Generating semantic structures
Identifying relevant sentences
Extracting elements and relations
Fill templates
regular expres-sions
regular expres-sions, rules
Terminology (WNC), hierarchy
regular expres-sions
tem-plates, rules
Textpreprocessor
Semantic Interpreter
Extractor

Fig. 1. Information Extraction Processing Pipeline with necessary resources

([Hospitalisation of the patient] [on suspicion of an urethra colic]) is decomposed into the head segment *[Aufnahme des Patienten]* and the prepositional segment *[unter dem Verdacht einer Ureterkolik]* using the preposition *'unter'*. The syntactic analysis also determines morphemes for each token and assigns a word class to each word. After parsing, each sentence is checked for special expressions by means of regular expressions. They are tagged before the indexing process starts and are interpreted separately. Special expressions include dates, quantities and dosage specifications, expressions with special meanings, like *Ausschluss von (exclusion of), Hinweis auf (evidence of).*

2.2 Semantic Interpreter

The *Semantic Interpreter* produces a semantic analysis for each sentence in the input text. The method makes use of a terminology of medical terms (Wingert Nomenclature[1] (WNC)) and an indexing algorithm for noun phrases expressing medical terms. First, each sentence is mapped to a set of one or more WNC indices by comparing morphemes of an input phrase and of WNC entries [8]. For each index, a semantic concept is created that consists of the index and a corresponding free-textual description (they form the referent of the concept), a semantic type (which is the concept type) and optionally, a semantic role (e.g., *evidence of*).

The semantic analysis identifies in the set of concepts for each sentence a central information unit and its modifying information by means of a hierarchy of semantic types of the WNC (each index belongs to one semantic category of the WNC). The hierarchy reflects the assumption that each sentence or each segment deals with one semantic entity, which in clinical narratives is normally a diagnosis/morphological change or a procedure/treatment. Concepts of one of these semantic types are considered as central information units which are specified by all others. First, the main information of each segment is identified. The corresponding concept is linked to the modifying concepts by relations. For the head segment in the sentence in Fig. 2 the index belonging to the category 'treatment' (*V000230*) is selected as main information of this segment. It is linked to the concept of the index *J000E57* which is considered as modifying information. A relation links two concepts and bears a relation type (in Fig. 2 the type is *job*). Second, the main information of the segments are joined: The main information of the head segment is considered as main concept of the whole sentence and is modified by the main information of the prepositional segments (here: modified by the concept *urethra colic*).

Using the described procedure, each sentence is mapped to a semantic representation consisting of concepts and relations (see Fig. 2). For more detail on the method see [2]. Terminology, indexing algorithm, preparser and parser are in routine use in several products of ID GmbH (Information and Documentation in Medicine, http://www.id-berlin.de).

[1] The WNC is a nomenclature based on the work of F. Wingert [9], who revised SNOMED II and translated it to German in 1984. It comprises a comprehensive multi-axial terminology of medical terms which allows for the encoding of different aspects of medical events. Ten different axes are available, e.g., topology, morphology, function, procedure, diagnosis. ID GmbH (http://www.id-berlin.de) continuously extends and enriches the WNC since several years.

2.3 Extractor

The *Extractor* uses the semantic structures generated for a text to extract specific information and fill in templates as in Fig. 3. For this, two modules must be provided: Templates, the system has to complete and extraction rules specifying how to identify and extract specific bits of information. A template description consists of a set of slots to be filled and a set of relevant paragraph classes which are used to restrict searching to relevant subsections. The template on risk factors consists, for example, of the two slots *Existence* and *Kind of risk factors* (see Fig.3). The required information is presumed to be in the paragraph *diagnoses.*

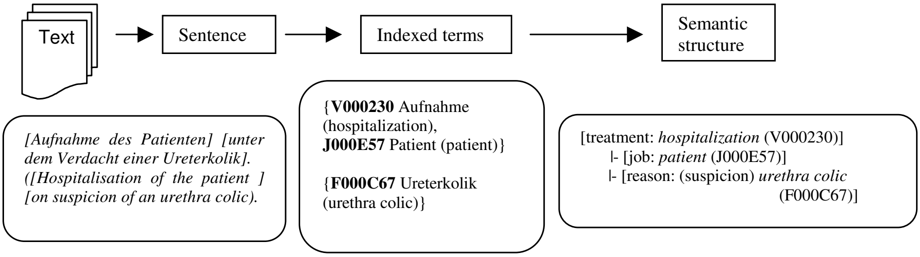

Text
Sentence
Indexed terms
Semantic structure
[Aufnahme des Patienten] [unter dem Verdacht einer Ureterkolik]. ([Hospitalisation of the patient] [on suspicion of an urethra colic).
{**V000230** Aufnahme (hospitalization), **J000E57** Patient (patient)}
{**F000C67** Ureterkolik (urethra colic)}
[treatment: *hospitalization* (V000230)]
|- [job: *patient* (J000E57)]
|- [reason: (suspicion) *urethra colic* (F000C67)]

Fig. 2. Generation of a semantic representation for *"Aufnahme des Patienten unter dem Verdacht einer Ureterkolik" (Hospitalization of the patient on suspicion of an urethra colic)*

For each slot of a template an extraction rule is defined which characterizes the desired information. An extraction rule can (by means of regular expressions) check input data for trigger words one or more of which a sentence must match to be considered relevant. In addition, rules can include conditions which extracted information has to meet. Furthermore, possible slot values can be fixed in extraction rules.

Subject	**Slot**	**Examples for values**	**No. of possible values**
Information on risk factors	Existence	Yes, No	2
	Kind of risk factors	M0009E1 (obesity)	7
Information on hospitalisation	Kind of admission	first admission	2
	Reason of admission	emergency	7
	Admission diagnosis	appendicitis	any WNC concept of the category morphology, diagnosis or function
Information on patient's state at discharge	State at discharge	comfortable	10
	General condition	good	6

Fig. 3. Templates on risk factors, state at discharge and hospitalization of a patient with examples for values

In the extraction process, information is collected from

- the written text itself (using a pattern-based method for filling templates) and
- from its semantic representation.

Using regular expressions, the system looks for specific words or strings in an input document to extract information from it. In case of a match, the corresponding slot is filled with a pertinent value. Some information can be derived from the section class (e.g., the slot *date of administration* of a medication that is extracted from the section *discharge* will be filled with the value *"at discharge").* Sometimes, the system uses regular expressions to determine whether a sentence is or is not relevant for further processing. E.g., for extracting relevant sentences containing information on the *hospitalization* of a patient, the text is skimmed for sentences matching the regular expression *(.*Aufnahme.*zur.*|.*wurde.*aufgenommen.*) (admission because of...|had been admitted).* Matching sentences are considered to be relevant for further processing.

To extract information from semantic structures, the latter are searched for different kinds of information (or combinations thereof):

- specific concepts (e.g., the concept for *mutual*),
- concepts belonging to specific semantic categories (e.g., concepts of category *diagnosis* or *date*),
- concepts with specific semantic roles (e.g., a concept with semantic role "*suspicion of*").

During processing each relevant semantic structure is searched for the information specified by the extraction rule. For extracting the *reason of admission*, for example, out the sentence '*Aufnahme des Patienten unter dem Verdacht einer Ureterkolik'* (*Hospitalization of the patient on suspicion of an urethra colic.)* its semantic structure (see Fig. 2) is searched for a concept that is a *diagnosis* (i.e. its semantic category is *diagnosis*, *function* or *morphological chang)e* and which has the semantic role *reason.* These conditions are given by the extraction rule. If a concept meets all the conditions specified in an extraction rule, a corresponding value is filled into the slot. For some slots, this process is completed at the first match. With other slots, the process is iterated until all the text or every semantic structure has been searched.

3 Results

In this section, evaluation results concerning the system's performance are presented. The evaluation is performed for the template filling module (step 6 in Fig.1). The quality is measured in precision and recall scores. Recall is a measure of 'completeness'; precision is a measure of 'cleanness'.

The evaluation of the system's performance regarding the template filling considered the mining of discharge summaries for information on the hospitalization of a patient, on risk factors and on the patient's state at discharge (see templates in Fig. 3). These tasks have been selected because in order to extract this information, all the different mentioned methods (explained in 2.3) are used. To evaluate the system's performance, a gold standard comprising 50 discharge summaries from a surgical department in a hospital was created. For each of these documents, relevant sentences were selected. The templates were filled manually by a physician. The system was run over the documents, and the results were compared with the values defined manually. The results of the first tests are given in Table 1.

Table 1. Results of the template filling evaluation

Template	Precision	Recall
Hospitalisation	81%	83%
State at discharge	95%	97%
Risk factors	96%	98%

The highest precision and recall scores were obtained in the extraction of described *risk factors*. This can be traced to the extraction method: To extract data on *risk factors*, the system looks for specific, predefined indices (e.g., indices for *obesity, arterial hypertension, diabetes*). The template filling quality for information on the *patient's state at discharge* is slightly lower. This extraction is realised by means of regular expressions. If a regular expression was missing, the corresponding information could not be extracted.

The extraction of *information on hospitalisation* has the lowest precision and recall values. The desired information comprises the kind and the reason for hospitalisation as well as the admission diagnosis (Fig. 3). These bits of information are directly extracted from the semantic structure of a sentence. For this reason, the extraction quality depends among other things on the indexing algorithm, which is the main source of error in the process of generating semantic structures (30% of the errors result from this). Apart from indexing errors, wrong paragraph detection and paragraphs that were not considered because of exclusion by the extraction rule as well as missing or not specific trigger words are responsible for errors in term extraction (Tab. 2). For some cases, the system requires additional knowledge and understanding to be able to extract relevant data.

Table 2. Main reasons of error and their quota of errors in template filling

Reasons of error	Error quota
Unknown words / problems during indexing	30%
Wrong paragraph detection	38%
Missing trigger words	27%
Processing needs additional knowledge	20%

With average values of 85% precision and 87% recall, the results for template filling are nevertheless promising. But the evaluation process itself has only begun so far and more subtasks will have to be tested with larger sets of documents.

4 Discussion

The first results show that the system is able to yield good results. The quality increases, if the information the system has to detect is given in more detail (like the different indices on risk factors). The system described here achieves recall and precision measures comparable to other reputable natural language processors with similar functionality in the medical domain [5, 6, 9]. It uses an existing medical terminology which covers all medical domains and is therefore not limited to a certain medical domain. To ameliorate the system's performance, minor extensions of the vocabulary will be appropriate.

Extraction rules as well as the database of special expressions can easily be added to the system, but they have to be defined manually. Other systems use statistical methods and learning algorithms to determine extraction rules automatically (e.g., Tessi® Extraction Engine [4] and MedIE [9]).

To process "new" document types, the system's modules will have to be "taught" the structural information provided by documents of the "new" type, i.e. their trigger words, the paragraph classes and their characteristic.

Another benefit of the system is its adaptability to other languages: The system is implemented for medical documents written in German, but with some extensions, it can process documents in other languages. The indexing algorithm is multilingual and the terminology is available in different languages (e.g. in English). To adapt the system completely to another language, its language-dependent components have to be modified; these are: the parser for structuring sentences using prepositions or rather its list of prepositions, the different databases with special expressions, semantic roles, stop words and the regular expressions in the extraction rules. The priority sequences and conditions for determining the main information in a sentence as well as the rules for extracting information from conceptual structures are language independent. In an evaluation of the system with regard to its multilinguality, the system identifies sentences containing diagnoses in 20 English radiology reports with precision and recall measures of 86% and 79%. The single accommodation performed was the use of a WNC with English and Latin terms instead of the lexicon with German terms.

5 Conclusion

In this paper, a system that extracts and mines a variety of information from free-text clinical records is described. Combining a variety of processing modules of different kinds with existing technologies and a medical terminology, the system achieved promising results in a first evaluation. It will be possible to improve the system's performance significantly by extending data bases and adjusting the extraction algorithms.

The next conceptual extension of the system will be a module for the determination of relationships beyond sentence boundaries including the unification of templates as well as the resolution of anaphors. In addition, the transferability of the general methods to other existing terminologies is going to be tested.

Currently, it is analysed, whether the methods are suitable to fill automatically slots of standardised electronic health records like the Continuity of Care Record (CCR, [10]). The CCR is a standard specification for exchanging patient data. At the moment, the dataset has to be created manually by a physician. IE methods like the one introduced in this paper can extract the required data from existing unstructured documents and fill it in the corresponding CCR slots. Then, a physician only has to check the generated CCR and add missing information where appropriate.

Acknowledgements. This project is supported by the software company ID Information and Documentation in Health Service who kindly approves access to its software products and knowledge resources.

References

[1] Cohen, A., Hersh, W.R.: A survey of current work in biomedical text mining. Briefings in Bioinformatics 6(1), 57–71 (2005)

[2] Denecke, K., Kohlhof, I., Bernauer, J.: Information Extraction Based On Multiaxial Indexing And Phrase Structure Analysis. In: Proc 20th Intern Congress of the Europ Fed for Med Inform (MIE 2006), August 2006, Maastricht (2006)

[3] Sowa, J.F.: Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA (2000)

[4] Language & Computing, Tessi® Extraction Engine (last access: 04/01/2007) http://www.landcglobal.com

[5] Mamlin, B.W., Heinze, D.T., McDonald, C.: Automated Extraction and Normalization of Findings from Free-Text Radiology Reports. In: JAMIA Proc Am Med Inform Assoc Annual Sympos, pp. 420–424 (2003)

[6] Mendonca, E.A., Haas, J., Shagina, L., Larson, E., Friedman, C.: Extracting information on pneumonia in infants using natural language processing of radiology reports. J of Biomed Inform 38(4), 314–321 (2005)

[7] Romacker, M., Hahn, U.: Empirical Data for the Semantic Interpretation of Prepositional Phrases in Medical Documents. In: Proc of the 2001 AMIA Annual Symposium, Washington, 2001, pp. 563–567 (2001)

[8] Wingert, F.: SNOMED Manual. Springer, Heidelberg (1984)

[9] Zhou, X., Han, H., Chankai, I., Prestrud, A., Brooks, A.: Approaches to text mining for clinical medical records. In: Proc ACM Sympos Applied Computing (SAC '06), April 2006, pp. 235–239. ACM Press, NY (2006)

[10] Waegemann, C.P., Engelbrecht, R., Klein, F.: CCR: Eine neue Lösung für Kontinuität der Information im Gesundheitswesen (CCR: A new solution for continuity of information in Health Care). In: mdi. Forum der Medizin Dokumentation und Medizin Informatik, September 2006, vol. 8(3), pp. 112–114 (2006)

Part V

Ontologies

Using Semantic Web Technologies for Knowledge-Driven Querying of Biomedical Data

Martin O'Connor[1], Ravi Shankar[1], Samson Tu[1], Csongor Nyulas[1], Dave Parrish[2], Mark Musen[1], and Amar Das[1]

[1] Stanford Medical Informatics, Stanford University School of Medicine,
Stanford, CA 94305
martin.oconnor@stanford.edu
[2] The Immune Tolerance Network, Pittsburgh, PA, USA

Abstract. Software applications that work with biomedical data have significant knowledge-management requirements. Formal knowledge models and knowledge-based methods can be very useful in meeting these requirements. However, most biomedical data are stored in relational databases, a practice that will continue for the foreseeable future. Using these data in knowledge-driven applications requires approaches that can form a bridge between relational models and knowledge models. Accomplishing this task efficiently is a research challenge. To address this problem, we have developed an end-to-end knowledge-based system based on Semantic Web technologies. It permits formal design-time specification of the data requirements of a system and uses those requirements to drive knowledge-driven queries on operational relational data in a deployed system. We have implemented a dynamic OWL-to-relational mapping method and used SWRL, the Semantic Web Rule Language, as a high-level query language that uses these mappings. We have used these methods to support the development of a participant tracking application for clinical trials and in the development of a test bed for evaluating biosurveillance methods.

1 Introduction

Biomedical applications often have significant knowledge and information management requirements. These requirements are felt at all system development stages, from initial system planning through final data analysis. Because of their complexity, these applications can benefit considerably from an ontology-driven systems development approach. In the past few years, there has been a surge in the development of ontology-based modeling projects to support biomedical application development [2]. However, few of these systems emphasize the knowledge requirements for day-to-day activities in deployed biomedical systems. The operational requirements of these systems are heavily influenced by current deployment technologies. For example, most data in these applications are stored in relational databases, a practice that will continue for the foreseeable future. The schema design of these databases often reflects the operational requirements of the

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 267–276, 2007.

system implementers and often maps poorly to the original domain level concepts used in system design.

As a result, knowledge used in system design can be difficult to use during monitoring and in during final data analysis, leading to *ad hoc* approaches that obtain little benefit from the knowledge-level technologies used in the rest of the system. A serious consequence is that data collected during system operation may be an imprecise and partial reflection of the intentions of the system designers. These shortcomings are often not noticed until system deployment, when they may not be correctible. The consequences can be serious, and may require lengthy and expensive data cleanup processes, or even the discarding of data.

Thus, there is a need for principled approaches to overcome inconsistencies between knowledge-level concepts in system design and corresponding operational data collected in a deployed system. To address this problem, we have developed an ontology-database mapping method that allows knowledge-driven applications to work directly with relational data. We developed this method to work with the Semantic Web ontology language OWL [1]. We chose OWL because of its expressive power and the fact there is expanding array of tools being developed for it. In particular, there are significant current efforts to develop OWL-based tools and methodologies in the biomedical field, such as cBio [2]. We also make extensive use of OWL's rule language SWRL [3], which dramatically increases OWL's problem solving range. In particular, SWRL supports the development of querying tools that allow dynamic knowledge-driven access to relational data. Our current work builds on ontology authoring technologies that we have been developing over the past decade. In particular, we have used Protégé-OWL [4], an open source framework that provides tools for constructing OWL ontologies and knowledge-based applications.

2 Background

The Semantic Web project is a shared research plan that aims to provide explicit semantic meaning to data and knowledge on the World Wide Web [5]. Currently, the state-of-the art in Web information retrieval does not lend itself to automated information processing. Future Semantic Web applications will be able to integrate data and knowledge automatically, through the use of a standardized language that describes the content of Web-accessible resources.

2.1 OWL and SWRL

OWL was developed as an ontology language for constructing ontologies that provide high-level descriptions of Web content. These ontologies are created by building hierarchies of classes describing concepts in a domain and relating the classes to each other using properties. OWL can also represent data as instances of OWL classes—referred to as individuals—and it provides mechanisms for reasoning with the data and manipulating it. OWL also provides a powerful constraint language for precisely defining how to interpret concepts in an ontology.

OWL provides limited deductive reasoning capabilities, however, and recent work has concentrated on adding rules to it. The Semantic Web Rule Language (SWRL;

[6]) allows users to write Horn-like rules that can be expressed in terms of OWL concepts and that can reason about OWL individuals. SWRL provides deductive reasoning capabilities that can infer new knowledge from an existing OWL ontology. For example, a SWRL rule expressing that a person with a male sibling has a brother would require capturing the concepts of 'person', 'male', 'sibling' and 'brother' in OWL. Intuitively, the concept of person and male can be captured using an OWL class called `Person` with a subclass `Man`; the sibling and brother relationships can be expressed using OWL properties `hasSibling` and `hasBrother`, which are attached to `Person`. The rule in SWRL would be:

```
hasParent(?x, ?x) ^ hasSibling(?y,?z) -> hasBrother(?x,?z)
```

Executing this rule would have the effect of setting the `hasBrother` property of `x` to `z`. Similarly, a rule that asserts that all persons who own a car should be classified as drivers can be written as follows:

```
Person(?p) ^ hasCar(?p, true) -> Driver(?p)
```

Again, this rule would require that the property `hasCar` and the class `Driver` exist in an OWL ontology. Executing this rule would have the effect of classifying all car-owner individuals of type `Person` to also be members of the class `Driver`.

One of SWRL's most powerful features is its ability to support user-defined methods or *built-ins* [7]. A number of core built-ins for common mathematical and string operations are defined in the SWRL proposal. For example, the built-in `greaterThan` can be used to determine if one number is greater than another. A sample SWRL rule using this built-in to help classify persons aged greater than 17 as adults can then be written as:

```
Person(?p)^ hasAge(?p,?age) ^ swrlb:greaterThan(?age,17) -> Adult(?p)
```

When executed, this rule would classify individuals of class `Person` with a `hasAge` property value greater than 17 as members of the class `Adult`.

Named OWL individuals in an ontology can also be referred to directly in SWRL rules. For example, one could rewrite the above rule to classify an individual named "Fred" as a driver as follows:

```
Person(Fred) ^ hasCar(Fred, true) -> Driver(Fred)
```

SWRL allows new libraries of built-ins to be defined and used in rules. Users can define built-in libraries to perform a wide range of tasks, which could include currency conversion libraries, and libraries including statistical, temporal or spatial operations.

2.2 OWL-Relational Mapping

As ontology development tools have been increasingly used to address real-world problems, scalability has become an important topic [8]. Initially, ontologies were used only for system specification and were stored in flat files and fully loaded into application memory when in use. This approach works well for small ontologies, but it does not scale well.

Systems that use ontology-based tools to work with operational data require tools that scale to deal with large amounts of data. Recent approaches to tackle this problem have focused on triple stores [9], which use native representation of RDF triples to store ontologies. Triple stores are analogous to relational database management systems and provide efficient storage and retrieval of ontology information. RDF query languages like SPARQL [10] can be used to provide SQL-like query functionality on triple stores. OWL is built using RDF, and OWL ontologies can be stored in triple-store back ends without loss of semantics.

However, using triple store back ends to store operational data in biomedical systems is not currently practical. These technologies are still in their infancy, and relational databases will continue to be used in these systems for the foreseeable future. As discussed earlier, this solution leads to a separation of knowledge and data, which creates a semantic gap. For example, a clinical trial monitoring application would track patients during a trial. The application may be required to answer the question "How many patients are in the intervention phase of the trial?", which would require access to the trial's operational database. Ideally, this question would be asked using the domain-relevant concepts in trial knowledge bases and these data would be automatically retrieved from the appropriate relational database. However, the relational model and the triple-based model that underlies OWL are incompatible with each other.

A number of developmental research systems aim to reconcile the two models [11]. The most straightforward approach is to statically map a relational database to a triple-store and to write queries against the store. This approach suffers from several shortcomings. First, there is an issue of data duplication. There are also questions about how frequently to update triple stores to reflect changes in associated relational database. Knowledge-driven applications requiring up-to-date information require frequent synchronization, which may be cumbersome. And, of course, if knowledge-driven updates are to be supported, the synchronization issue arises in the reverse direction.

Ideally, knowledge-driven data requests would retrieve data from live relational databases. This approach would require automatic or semi-automatic dynamic mapping between relational databases and triple-based formats. A software layer would translate knowledge-level queries into SQL-queries and retrieve the required data from a database. Further reasoning with the retrieved knowledge could be performed in memory. If updates were to be allowed, the reverse transformation would also be supported.

3 Knowledge-Level Querying with SWRL

There are no standard OWL-based query languages. Several RDF-based query languages exist but they do not capture the full semantic richness of OWL. To tackle this problem, we have developed a set of built-in libraries for SWRL that allow it to be used as a query language.

This work was performed with the SWRLTab, a Protégé-OWL development environment for working with SWRL rules [12, 13]. The SWRLTab has several software components, including (1) an editor that supports interactive creating, editing, reading, and writing of SWRL rules; (2) a rule engine bridge that provides the infrastructure necessary to incorporate third-party rule engines into the SWRLTab to execute SWRL rules; (3) a built-in bridge that provides a mechanism for defining Java implementations of SWRL built-ins, and (4) a set of built-in libraries containing mathematical, string, and temporal operators, in addition to a query library than can be used to turn SWRL into a query language.

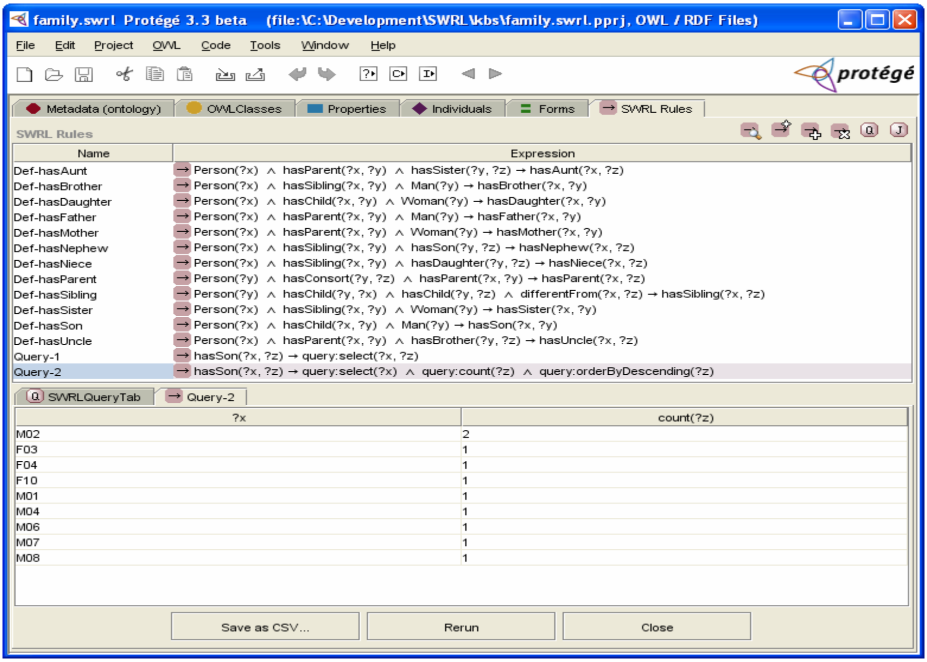
This query library allows SWRL rules to be used to query OWL knowledge bases. It contains SQL-like built-ins that can be used in a rule to construct retrieval specifications. For example, the following rule, written with these built-ins, retrieves all patients in an ontology whose age is less than 25, together with their ages:

```
Patient(?p) ^ hasAge(?p,?a) ^ swrlb:lessThan(?a,25) ->
query:select(?p,?a)
```

This query will return pairs of patients and ages. Query built-ins can be used with other built-in libraries provided by the SWRLTab. The ability to use built-ins freely in a query provides a means of continuously expanding the power of the query language.

The following query lists all patients together with their ICD9 codes:

```
Patient(?p) ^ hasICD9(?p, ?icd) -> query:select(?p, ?icd)
```



Name	Expression
Def-hasAunt	Person(?x) ∧ hasParent(?x, ?y) ∧ hasSister(?y, ?z) → hasAunt(?x, ?z)
Def-hasBrother	Person(?x) ∧ hasSibling(?x, ?y) ∧ Man(?y) → hasBrother(?x, ?y)
Def-hasDaughter	Person(?x) ∧ hasChild(?x, ?y) ∧ Woman(?y) → hasDaughter(?x, ?y)
Def-hasFather	Person(?x) ∧ hasParent(?x, ?y) ∧ Man(?y) → hasFather(?x, ?y)
Def-hasMother	Person(?x) ∧ hasParent(?x, ?y) ∧ Woman(?y) → hasMother(?x, ?y)
Def-hasNephew	Person(?x) ∧ hasSibling(?x, ?y) ∧ hasSon(?y, ?z) → hasNephew(?x, ?z)
Def-hasNiece	Person(?x) ∧ hasSibling(?x, ?y) ∧ hasDaughter(?y, ?z) → hasNiece(?x, ?z)
Def-hasParent	Person(?y) ∧ hasConsort(?y, ?z) ∧ hasParent(?x, ?y) → hasParent(?x, ?z)
Def-hasSibling	Person(?y) ∧ hasChild(?y, ?x) ∧ hasChild(?y, ?z) ∧ differentFrom(?x, ?z) → hasSibling(?x, ?z)
Def-hasSister	Person(?x) ∧ hasSibling(?x, ?y) ∧ Woman(?y) → hasSister(?x, ?y)
Def-hasSon	Person(?x) ∧ hasChild(?x, ?y) ∧ Man(?y) → hasSon(?x, ?y)
Def-hasUncle	Person(?x) ∧ hasParent(?x, ?y) ∧ hasBrother(?y, ?z) → hasUncle(?x, ?z)
Query-1	hasSon(?x, ?z) → query:select(?x, ?z)
Query-2	hasSon(?x, ?z) → query:select(?x) ∧ query:count(?z) ∧ query:orderByDescending(?z)

?x	count(?z)
M02	2
F03	1
F04	1
F10	1
M01	1
M04	1
M06	1
M07	1
M08	1

Fig. 1. The SWRLQueryTab, a subcomponent of the Protégé-OWL SWRLTab that graphically displays results of SWRL rules containing query built-ins

This query will return pairs of patients and their ICD9 codes. Assuming a patient can have more than one ICD9 code, multiple pairs would be displayed for each patient—one pair for each code.

The query built-in library also supports basic counting with a built-in called `count`. The following is a query to count the number of ICD9 codes for each patient:

```
Patient(?p) ^ hasICD9(?p, ?c) -> query:select(?p) ^ query:count(?c)
```

The query library also provides basic aggregation, ordering, and duplicate elimination operators. A JDBC-like Java API executes queries and processes results in Java applications. A graphical interface allows interactive execution of rules and examination of results (Figure 1). This software is part of the current Protégé-OWL distribution.

4 OWL-Relational Mapping

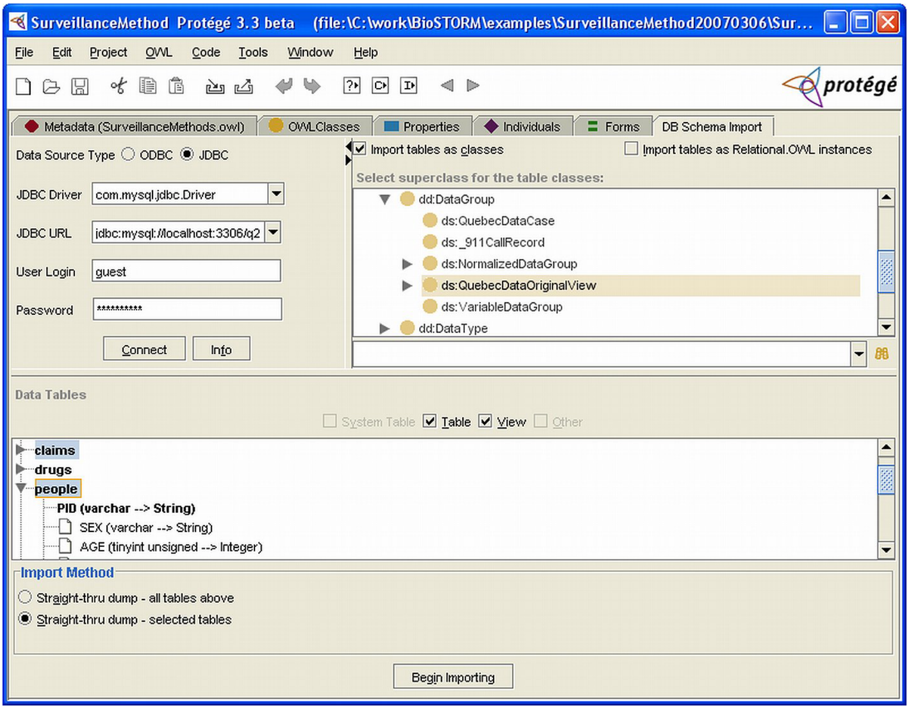
We are using SWRL to provide a rich high level language to specify the data requirements of OWL-based biomedical applications [14, 15]. Our goal is to use SWRL to help unify the domain-level specification of system data with the run-time operational data needs of system components. In conjunction with OWL, SWRL is provides both a formal domain-level description of data in the biomedical systems that we are developing and a run-time query mechanism to execute domain-level queries on operational data in these systems.

We have developed a relational-to-OWL mapping technology to serve as a relational data access mechanism for these applications. As mentioned in Section 2, to support knowledge-driven querying of relational databases, tool are required to dynamically map data from relational databases to concepts described in OWL. Two primary components are required to build this technology: (1) schema and mapping ontologies to describe both the schema of an arbitrary relational database and the mapping of data stored in these schemas to triples in an OWL ontology. Tools to produce these ontologies, either automatically or semi-automatically with user guidance, are also required; and (2) mapping tools and a query engine that use the schema source and mapping ontologies to dynamically map data and to translate knowledge-driven data requests to queries on a relational database.

4.1 Schema and Mapping Ontologies

We have developed a schema ontology in OWL that provides a knowledge-level description of a relational schema. The ontology describes schemas in a database and associated tables together with the columns and column data types contained in each table. It also describes primary and foreign key relationships for tables in a schema. We have also developed a mapping ontology that uses this schema ontology to describe how relational tables are to be mapped to OWL concepts. The fundamental goal is to specify the mapping of rows in a relational table to triples in an RDF model, which will then be mapped to OWL classes, properties and individuals. This process cannot normally be performed automatically, and additional user markup is usually required.

We have written a tool in Protégé-OWL that generates a schema ontology automatically from any RDBMS supporting the JDBC interface. The tool allows user specification of relational-to-OWL mappings. Users can interactively specify the source database to be mapped and the target OWL concepts to which it must be mapped. Figure 2 shows a screenshot of this application.


SurveillanceMethod Protégé 3.3 beta (file:\C:\work\BioSTORM\examples\SurveillanceMethod20070306\Sur...
File Edit Project OWL Code Tools Window Help
protégé
Metadata (SurveillanceMethods.owl) OWLClasses Properties Individuals Forms DB Schema Import

Data Source Type ○ ODBC ● JDBC

JDBC Driver com.mysql.jdbc.Driver

JDBC URL jdbc:mysql://localhost:3306/q2

User Login guest

Password **********

Connect Info

☑ Import tables as classes ☐ Import tables as Relational.OWL instances

Select superclass for the table classes:

- dd:DataGroup
  - ds:QuebecDataCase
  - ds:_911CallRecord
  - ds:NormalizedDataGroup
  - ds:QuebecDataOriginalView
  - ds:VariableDataGroup
- dd:DataType

Data Tables

☐ System Table ☑ Table ☑ View ☐ Other

- claims
- drugs
- people
  - PID (varchar --> String)
  - SEX (varchar --> String)
  - AGE (tinyint unsigned --> Integer)

Import Method

○ Straight-thru dump - all tables above
● Straight-thru dump - selected tables

Begin Importing

Fig. 2. Screenshot of the schema import and mapping tool. Users can specify links between the source relational information and OWL entities.

4.2 Mapping Software and Query Engine

We also developed mapping software that works with a query engine to allow queries written in SWRL to use data retrieved from a relational database. At run-time, the software uses the source and mapping ontologies to transform the data in a relational database to OWL entities. We extended our existing query engine to interact with this software to retrieve these mapped OWL entities. The engine takes SWRL queries written terms of OWL classes, properties, and individuals and generates requests to the mapping software for the OWL entities in a relational database. The mapping software then generates SQL queries to retrieve that appropriate data from the database identified by the schema and mapping ontologies.

We have developed an array of optimization techniques to improve the performance of the mapping task. Consider, for example, a query that retrieves all patients older than 17 years and orders results by age:

```
Patient(?p) ^ hasAge(?p,?age) ^ swrlb:greaterThan(?age,17)
-> swrlq:select(?p) ^ swrlq:orderBy(?age)
```

Assuming that patients and their age properties are mapped from a database, a naïve mapping implementation would retrieve all patient individuals and their ages, and then apply the knowledge-level `greaterThan` built-in to the retrieved data. For common mathematical built-ins, optimizations can be made by generating SQL queries that exclude unnecessary data. In this case, an additional clause can exclude patients not aged over 17. Given the greater performance of modern relational database systems, these optimizations can lead to significant performance improvements for some queries.

We have optimized the underlying SQL queries generated to retrieve data for built-ins by adding additional information to built-ins. This information is defined by a *built-in annotation ontology* and describes the nature of the operations they perform. Since SWRL built-ins are stored as standard OWL individuals, we added this information with each built-in by associating annotations with it. The mapping software then reads this information when generating SQL requests for data.

Optimizations are also possible by rewriting SWRL queries. For example, a SWRL query's meaning is independent of the order of its constituent atoms. Atoms can thus be reordered without changing meaning. This reordering can reap considerable benefits because normal atom ordering can conspire to reduce performance. Typically, more general atoms are used near the start of a SWRL rule or query and become progressively more specific. By reordering atoms to place more specific—and thus less likely to be satisfied—atoms near the beginning of a SWRL query, many unnecessary evaluations can be avoided. Clearly, atoms that use mapped relational data should be reordered as late as possible in a query.

Rule base level optimizations are also possible. A rule may have a major 'axis of evaluation' that can be exploited for optimization. For example, a temporal rule may reason with data over a particular temporal interval and ignore all data outside it. Identifying the interval automatically by processing the rule is usually not possible. However, by using the markup strategy used for built-ins, rules for identifying them can be annotated by a developer. The markup information can be used to identify the variables in a rule that specify the start and end of the temporal interval it is dealing with. This information can be used at run time to ensure the retrieval of data specifically in that range. We have implemented this strategy for temporal rules that, with the temporal ontology we developed earlier, restricts data access for rules using temporal built-ins. We adopted a similar approach in an earlier knowledge-driven system that dealt with pattern detection within temporal intervals [7]. Similar strategies for other domains with discrete data access patterns are also conceivable.

Standard database optimization techniques can also be used. For example, an artifact of the triple-to-relational mapping approach is that some queries can generate expensive multi-way relational joins. Adding intermediate views in addition to auto-generated primary key columns to tables and then creating indexes with them can improve performance dramatically. We have seen 100x improvements to some queries using this technique. This approach has been in use for many years in object-relational mapping software and can be used without breaking legacy database applications that worked with the unmodified schemas.

5 Results and Discussion

We are using these technologies in the development of a knowledge-based framework called Epoch [14,15] and in the creation of a visit and specimen tracking application for the Immune Tolerance Network (ITN; [16]). The ultimate goal of the ITN is the discovery of tolerance mechanisms common to multiple immune disorders. To support this goal, we have used knowledge-based techniques to support a consistent approach to data specification, planning, and execution of clinical trials throughout the entire organization. We are applying our mapping method to two areas that are vital in trials: (1) tracking study participants as they advance through the trials, and (2) tracking specimens as they are processed in laboratories. We used SWRL in our tracking application to express knowledge-level queries than were then mapped to ITN's data repositories to determine the number of patients in various stages of a clinical trial. The work outlined in this paper is part of the first phase of this process. It serves as a starting point to address the issue of consist data representation across trials in the ITN and of using these data in the monitoring of trials. We are actively working on methodologies to integrate additional trial data, such as participant's clinical data, and assay results. Using these data, we plan to develop further knowledge-driven applications to support a range of tasks, including site management, and assay analysis.

We are also using these tools in the development of an ontology-driven test bed for evaluating biosurveillance methods [17]. This test bed will to provide a scalable architecture for configuring biosurveillance methods. It allows users to draw on real-world surveillance data sources held in relational database and to configure, run, and evaluate alternative surveillance methods using the data. Since biosurveillance methods typically work with large quantities of data, this test bed places significant performance demands on the mapping software. The ability to meet these demands is crucial to enable this technology to meet the scalability requirements of the Semantic Web.

While the work described here is targeted to clinical trials and surveillance applications, the underlying technologies have general applicability. The technologies and tools are reusable and are distributed for general use with Protégé-OWL system, one of the most widely used ontology development environments.

Acknowledgements. This work was supported in part by the Immune Tolerance Network, which is funded by the National Institutes of Health under Grant NO1-AI-15416, and also by the Centers for Disease Control and Prevention under grant number SPO-34603. We thank Valerie Natale for her editorial comments

References

1. OWL Overview: http://www.w3.org/TR/owl-features/
2. cBio: http://www.bioontology.org/
3. SWRL Submission: http://www.w3.org/Submission/SWRL/

4. Knublauch, H., Fergerson, R.W., Noy, N.F., Musen, M.A.: The Protégé OWL Plugin: An Open Development Environment for Semantic Web applications. In: Proc Third ISWC (ISWC 2004), Hiroshima, Japan, pp. 229–243 (2004)
5. Berners-Lee, T.: The Semantic Web, Scientic American (May 2001)
6. SWRL Submission: http://www.daml.org/2003/11/swrl
7. SWRL Built-ins: http://www.daml.org/2004/04/swrl/builtins
8. Lopez, V., Sabou, M., Motta, E.: PowerMap: Mapping the Real Semantic Web on the Fly. In: 5th International Semantic Web Conference, November 5-9, 2006 (2006)
9. http://simile.mit.edu/reports/stores/
10. http://www.w3.org/TR/rdf-sparql-query/
11. Chen, H., Wang, Y., Wang, H., Mao, Y., Tang, J., Zhou, C., Yin, A., Wu, Z.: Towards a Semantic Web of Relational Databases: a Practical Semantic Toolkit and an In-Use Case from Traditional Chinese Medicine. In: Fifth International Semantic Web Conference, Georgia, USA (2006)
12. SWRLTab: http://protege.cim3.net/cgi-bin/wiki.pl?SWRLTabBuiltInLibraries
13. O'Connor, M.J., Knublauch, H., Tu, S.W., Grossof, B., Dean, M., Grosso, W.E., Musen, M.A.: Supporting Rule System Interoperability on the Semantic Web with SWRL. In: Fourth International Semantic Web Conference, Galway (2005)
14. O'Connor, M.J., Shankar, R.D., Das, A.K.: An Ontology-Driven Mediator for Querying Time-Oriented Biomedical Data. In: 19th IEEE International Symposium on Computer-Based Medical Systems, Salt Lake City, Utah (2006)
15. Shankar, R.D., Martins, S.B., O'Connor, M.J., Parrish, D.B., Das, A.K.: A Knowledge-Based System for Managing Clinical Trials. In: 19th IEEE International Symposium on Computer-Based Medical Systems, Salt Lake City, Utah (2006)
16. Rotrosen, D., Matthews, J.B., Bluestone, J.A.: The Immune Tolerance Network: a New Paradigm for Developing Tolerance-Inducing Therapies. J Allergy Clinical Immunology 110(1), 17–23 (2002)
17. Crubezy, M., O'Connor, M.J., Buckeridge, D.L., Pincus, Z.S., Musen, M.A.: Ontology-Centered Syndromic Surveillance for Bioterrorism. IEEE Intelligent Systems 20(5), 26–35 (2005)

Categorical Representation of Evolving Structure of an Ontology for Clinical Fungus

Arash Shaban-Nejad and Volker Haarslev

Department of Computer Science and Software Engineering, Concordia University,
H3G1M8 Montreal, Quebec, Canada
{arash_sh,haarslev}@cs.concordia.ca

Abstract. With increasing popularity of using ontologies, many industrial and clinical applications have employed ontologies as their conceptual backbone. Ontologies try to capture knowledge from a domain of interest and when the knowledge changes, the definitions will be altered. We study change management in the FungalWeb Ontology, which is the result of integrating numerous biological databases and web accessible textual resources. The fungal taxonomy is currently unstable and evolves over time. This evolution can be seen in both nomenclature and the taxonomic structure. In an experiment we have focused on changes in medical species of fungus which can potentially alter the related disease name and description in an integrated clinical system. In order to address certain aspects of representation of changes in an ontology driven clinical application we propose a methodology based on category theory as a mathematical notation, which is independent of a specific choice of ontology language and any particular implementation.

Keywords: Bio-Ontologies, Category Theory, Change Management, Fungal Genomics.

1 Introduction

Ontologies provide an underlying discipline of modeling medical applications by defining concepts, properties and axioms. They are useful in current medical applications for: sharing common vocabularies, describing semantics of programming interfaces, providing a structure to organize knowledge, reducing development effort for generic tools and systems, improving the data and the tool integration, reusing organizational knowledge [2] and capturing behavioral knowledge. We have implemented the FungalWeb Ontology [1] which is a formal bio-ontology in the domain of in the domain of fungal enzymology with a large number of instances implemented in OWL-DL. We are now trying to develop a change management mechanism to update ontological knowledge representations. Ontologies such as living organisms are evolving over the time in order to fix the errors, reclassifying the taxonomy, adding/removing concepts, attributes, relations and instances. Modifying and adjusting ontologies in response to changing data or requirements is not a trivial task. One of the most fundamental questions in our research is: how to represent changes? In order to address certain aspects of representation of changes in an ontology driven

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 277–286, 2007.

application in the biomedical domain, in this paper we propose a method based on category theory. In our research, we have focused on ontologies not in isolation but as artifacts that are part of an integrated healthcare system. As an experiment we have focused on changes in medical species of fungus which can potentially alter the related disease name and description in an integrated clinical system.

2 Fungi Phylogeny and Evolution

Fungi are widely used in industrial, medical, food and biotechnological applications. They are also related to many human, animal and plant diseases, food spoilage and toxigenesis [4]. Fungi are also interesting because their cells are surprisingly similar to human cells [5]. The reason is that fungi split from animals about 1.538 billion years ago - 9 million years after plants did – therefore fungi are more closely related to animals than to plants [6]. It is estimated that there are about 1.5 million fungal species [7] on the earth, but only about 10% of those are known and only a few of the known fungus have an identified usage such as yeast for making bread, beer, wine, cheese and a few antibiotics [5]. A small percentage of discovered fungi have been linked to human diseases, including dangerous infections. Treating these diseases can be risky because as mentioned above human and fungal cells are very similar. Any medicine that kills the fungus can also damage the human cells. Thus knowing more about fungi and correct identification of each fungi species is crucial and can improve the quality of fungal-based products and also helps to identify new and better ways to treat serious fungal infections in humans. Fungus are also the main source of agricultural and plant diseases, so identifying them will help for tracking and controlling these diseases [5]. Typically, fungal evolution studies have been based on comparative morphology, cell wall composition [8], ultrastructure [9], cellular metabolism [10], and the fossil records [11]. Recently, by advances in cladistic and molecular approaches new insight is provided [12]. Some other new identification methods are based on Immuno-taxonomy and polysaccharides [12], which are highly suited antigens for the identification of fungi at the genus and species levels [13]. The following fungal chemical substances are also used as complementary characters to the classical morphological taxonomy of fungi: proteins, DNA, antigens, carbohydrates, fatty acids and secondary metabolites. One can find a review of the methods for employing the substances in [14]. These substances are very valuable at many taxonomic levels and they play an increasing role in the clarification of the phylogeny (a classification or relationship based on the closeness of evolutionary descent) of fungi [13].

2.1 Name Changes in Fungal Taxonomy

Most fungal names are not stable and change with time. Fungal names reflect the data about organisms and as our understanding of the relationships among taxa increases, names will be forced to change so that they do not implicitly contradict the data [15]. Most names are currently based on the phenotype (visible characteristics of organism). As more data become available, however, we run into various problematic issues, such as convergent evolution, seen as the evolution of the same form in different families and even orders, so that similar anamorphs (the imperfect (asexual) state of a fungus)) may have completely different, unrelated teleomorphs (the sexual stage in the life cycle of a fungus; considered the perfect stage). These names then have to change, as they no longer convey the correct information to the user [15]. These name changes may cause confusion and affect the validity of different queries. An example about eyespot disease of cereals and issues related to naming its associated fungi is actually represented at [16]. The morphological conceptualization is not sufficient, and will no longer work because all names based only on morphology have to be re-evaluated. In addition, the phylogenetic based conceptualization also has its own limitations, as sometimes the decision of where to draw the line between different species is not easy to make [15]. Another issue in fungal taxonomies is dual nomenclature (two names for one organism) due to the anamorph/teleomorph debate [15]. This is caused by the fact that it is frequently impossible to say when an asexual state belongs to a specific sexual state without the backup of molecular data. A study on revision of the fungi names [17] shows that between 1960 and 1975, 212 names of foliicolous lichenized fungi were described or used by A.C. Batista and co-workers.

2.2 Managing Name Changes

We are currently in the middle of a revolution in fungal taxonomy [15]. Names are linked to data. Older names, are mostly classified based on small data sets (mostly phenotypic), and therefore they are subject to change. How biologists can deal with this process of continuous change? To answer to this question one needs to refer to the nature of ontological structure, where names in taxonomy are only meaningful and valuable once linked to descriptive datasets which were extracted and managed from various databases and literatures in an integrated environment. The incorporation of DNA data is also needed to ensure stability in names and reliable species recognition. By advances in the technology in the future, biologists hope to preserve the fungal taxonomy from change by using unique DNA signatures and species identifier numbers to recognize the species rather than using their name [19]. Currently only around 16% of 100000 known fungal species are represented by DNA sequence data [15], which is approximately 1.1% of the estimated 1.5 million species on Earth, thus it seems that a very low percentage of the already discovered fungal species are in fact being preserved from the change [20]. The changing nomenclature of fungi medical importance is often very confusing. Currently some of the pathogenic fungi have a very unstable taxonomy. For instance, the name of the fungi, Allescheria boydii which can cause various infections in humans, was changed to Petriellidium boydii and then to Pseudallescheria boydii within a short time [23]. Consequently, the infections caused by this organism were referred to as allescheriasis, allescheriosis, petriellidosis, and pseudallescheriosis in the medical literature [24]. In order to manage the changes in fungal names and clarify the ambiguities, the Nomenclature Sub-Committee of the International Society for Human and Animal Mycology (ISHAM) published its regulations for mycosis nomenclature [23, 24]. Based on these regulations a disease should be named, with a meaningful name describing the disease, while in the traditional disease taxonomies the names "fungus+sis" indicate only a causative fungal genus which could be highly influenced by the taxonomic changes. In addition, in the new regulation the value of names of the "pathology A due to fungus B" construction was emphasized [23], e.g., "subcutaneous infection due to *Alternaria longipes*" [12].

2.3 Changes and Revisions in Taxonomic Structure

By advancing in molecular biology and changing the fungal nomenclature, one can expect changes in taxonomical structure and relationships. Here are some examples:

Example 1: *Glomeromycota* was discovered in 2001 [25] as a new fungal phylum. The arbuscular mycorrhizal (AM) fungi and the endocytobiotic fungus, *Geosiphon pyriformis*, are analyzed phylogenetically by their small subunit rRNA gene sequences. By studying their molecular, morphological and ecological characteristics, it is discovered that they can be separated from all other major fungal groups in a monophyletic clade [25]. Consequently they are removed from the polyphyletic *Zygomycota*, and located into a new monophyletic phylum, the *Glomeromycota* with four new orders *Archaeosporales*, *Paraglomerales*, *Diversisporales* and *Glomerales* [25].

Example 2: The sedge parasite *Kriegeria eriophori* has never been satisfactorily classified, because a number of its characters at the gross micromorphological and ultrastructural levels appeared to be autapomorphic [26]. Recently by using the nucleotide sequence data approach which provides more information than standard morphological approaches, some of the ultrastructural characters were discovered to be synapomorphies for a group containing *K. eriophori* and *Microbotryum violaceum.* These characters serve to define the new subclass Microbotryomycetidae [26].

3 Category Theory and Ontologies

Category theory is a new domain of mathematics, introduced and formulated in 1945 [27]. A formal model of objects based on "category theory" is introduced in [28]. Employing formalisms based on logics and mathematics in order to move the Web from being only human understandable, to being both human and machine understandable is the known goal of Semantic Web defined by W3C [30]. Category theory is closely connected with computation and logic [31] which allows an ontology engineer to implement different states of design models to represent the reality. Using categories one can recognize certain regularities to distinguish a variety of objects, capture and compose their interactions and differentiate equivalent interactions, identify patterns of interacting objects and extract some invariants in their action, or decompose a complex object in basic components [32]. Categorical notations consist of diagrams with arrows. Each arrow $f: X \rightarrow Y$ represents a function. A Category C includes:

- A class of objects and a class of morphisms ("arrows") and for each morphism f there exists one object such as A as the domain of f and one object such as B as the codomain. (Figure 7.1 (a))
- For each object, A, an identity morphism which has domain A and codomain A. ("ID$_A$ ") (Figure 7.1 (b))
- For each pair of morphisms f:A→B and g:B→C, (i.e. cod(f) = dom(g)), a *composite morphism*, $g \circ f$: A→C exists (Figure 7.1 (c)).

Representation of a category can be formalized using the notion of a diagram.

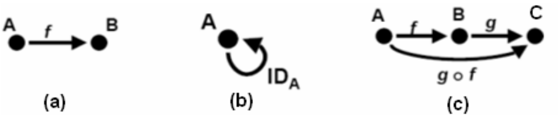


Fig. 1. Categorical concepts representation

The concept of ontology is based on the categorization of things in the real world. Category theory with its logical and analytical features has the potential to be considered as a vehicle for representation of ontologies. An ontology can be viewed in an interconnected hierarchy of theories as a sub-category of a category of theories expressed in a formal logic [29]. In fact we can use category theory to represent ontologies as a modular hierarchy of domain knowledge. Ontological relationships represented using category theories are considered to be directed [18] to show the direction of information. These “relationships” are known as “morphism”.

3.1 The Category Class

Classes can be defined as a set of properties (attributes and methods) shared by a set of individuals within an equivalence class. Whitmire [31] was one of the few who identified a model based on category theories for object oriented applications measurement. Here we follow his approach for demonstration of ontological elements. We can define category Class with attribute domains as objects and set-theoretic functions as arrows. In category theory, the cross product of two objects is an object. We can also define some operations for a class. In ontology, a concept or an instance can transit from one state to another based on its behavior in response to a change. An event can be formally modeled as an ordered pair *E* = <*St1, St2*> [32]. St1 is the start state and St2 is the end state. St1 and St2 are not necessarily distinct and they might refer to the same state [22] (when an even does not change the state). Category *Class* is defined with 3 types of objects and 3 types of arrows. The 3 types of objects are [31]:

1- The state space for the class, labeled with the name of the class.
2- The domain sets for the attributes in the class, labeled with the name of the domain.
3- The steady states (a situation in which the relevant variables are constant over time) for objects of the class, labeled with the name for the state used in the domain.

Three types of arrows are: projection, selection and operation arrows.

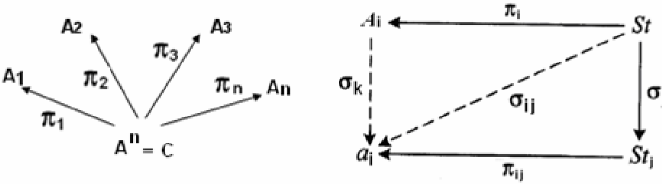
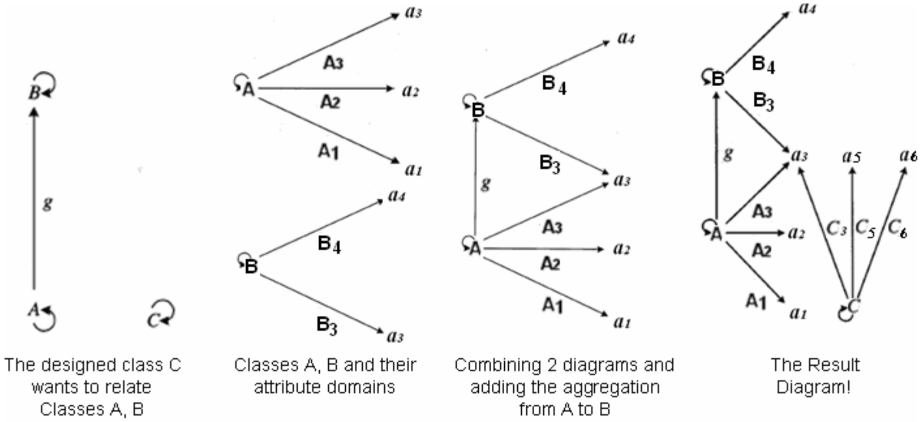


Fig. 2. Representation of the n attribute domains, and the state space of class C (adapted from [31])

The projection arrow for each attribute is drawn from the state space to the attribute domain and labeled with the name of the attribute. The value of the *i*th attribute is provided by π_i. A selection arrow for each state is drawn from the state space to the state and labeled as σ_x where *x* is the name of the state. An operation arrow for each event *E* = <St_1, St_2> drawn from S_1 to S_2 and labeled with the name of the method to which the operation corresponds [31]. One can select a state using the selection function σ_i which gives the *i*th state.

3.2 Operations on a Class

Most common operations during ontology evolution are: add a class, delete a class, combine two classes into one, add a generalization relationship, add an association relationship, add/delete a property and add/delete a relationship. Figure 3 represents adding a class to our available structure. Figure 4 (a) and (b) demonstrate adding and dropping a relationship respectively.



The designed class C wants to relate Classes A, B

Classes A, B and their attribute domains

Combining 2 diagrams and adding the aggregation from A to B

The Result Diagram!

Fig. 3. Adding a class to the available structure, based on categorical operation (adapted from [31])

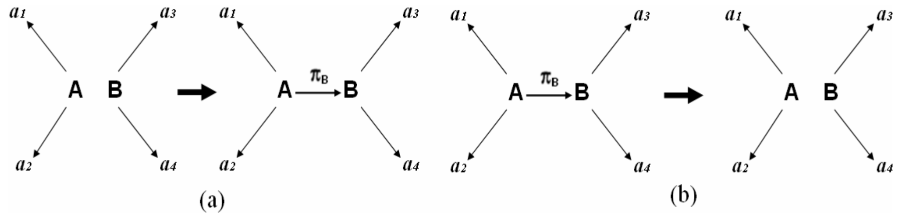


Fig. 4. (a) ADD an Aggregation Relationship (**b**) Drop a Relationship [31]

4 Managing Changes Using Category Theory

The categorical representation enables the progressive analysis of ontologies. After describing the ontological concepts within categories representing a modular hierarchy of domain knowledge, we employ category theory to analyze ontological changes in the following ways:

I. **By comparing a previous state of a class with a later state:** A categorical model [31] is able to describe the state space (set of all possible states for a given state variable set) for a class as a cross product of attribute domains and the operations of a class as transitions between states. It also allows the definition of message passing and method binding mechanisms. Category theory has a special type of mapping between categories called *functor*. Functors are defined as morphisms in the category of all small categories (where classes are defined as categories) [21]. The role of time is not usually taken into account in current ontology evolution studies. Considering time in ontologies can increase the complexity and needs a very expressive ontology language to represent it. In our approach, we represent conceptualization of things indexed by times, for example from the FungalWeb Ontology: "*enzyme* has_pH_optimum at *t*" is rendered as "*enzyme*-at-*t* has_ pH_optimum". Then we use a set of categories indexed by time using functors to capture different state of ontological structure at different time points. The category *O* at time t that is represented as O_t models the state of the ontologies and all the related interactions at this time. Using a functor allows us to represent the transition from O_t to $O_{t'}$ (Figure 5) where the time changes from *t* to *t'*. In addition, each sub ontology *A* can be modeled by the series of its successive states A_t from its '*Creation*' to '*Destruction*' [32].

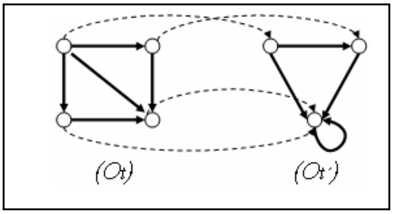


Fig. 5. Using Functor

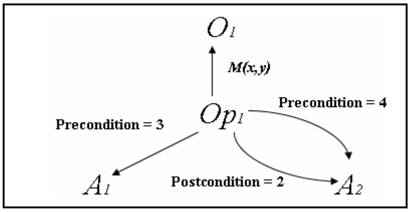


Fig. 6. Measuring Coupling

II. **By measuring coupling:** Coupling specifies the extent of the connections between elements of a system and it can identify the complexity of an evolving structure. Measuring coupling is useful for predicting and controlling the scope of changes to an ontological application. Often a change in one class can cause some changes to the dependent classes. When the coupling is high, it indicates existence of a large number of dependencies in an ontological structure which must be checked to analyze and control the chain of changes. Coupling for ontological elements can be described by a number of connections and links between them. So, we focus on arrows in category theory to study these connections. For analyzing a conditional change we followed the formal model described in [31] by identifying three types of arrows in our category: precondition, post-condition and message-send arrows for an existing category [31]. The type of message is determined by the types of changes caused by a method. In the category shown in

Figure 6, the coupling for the operation *Op1* is a nonnegative number which can be calculated by the count of the three types of arrows (post-conditions, preconditions and M(x,y)).

5 Application Scenario

Bioinformatics is a challenging domain in knowledge management. Biological data are highly dynamic and bioinformatics applications are large and have complex interrelationships between their elements. In addition, they usually have various levels of interpretations for one particular concept. In 1958 Rosen [3] proposed to use category theory in biology, in the frame of a ''relational biology''. At this time, we are applying the proposed methods for managing changes in the FungalWeb Ontology which is the result of integrating numerous biological databases, web accessible textual resources and interviews with domain experts and reusing some existing bio-ontologies. Figure 7 demonstrates a portion of the FungalWeb application in categorical representation.

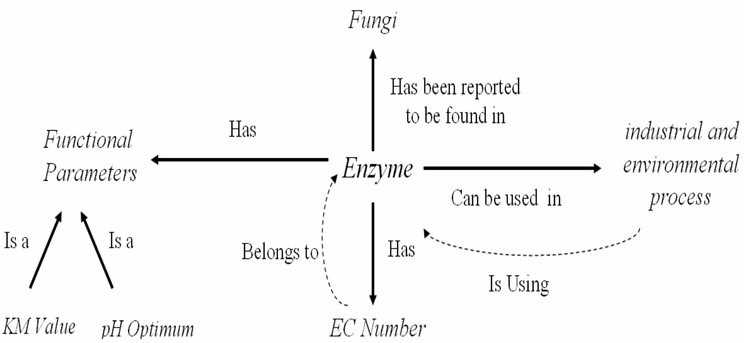


Fig. 7. A portion of the FungalWeb application

Based on our application we designed our class diagrams following the method described in [31] (Figure 8). The Opi arrows in this figure represent the operations for the class. In this class, the operation or event op1 causes an object in state St1 to transition to state St2. The operation Op1 has no effect upon the object if it is in any other

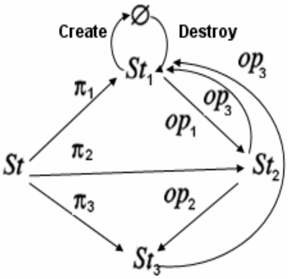


Fig. 8. A Class diagram for part of a class structure

state, since there is no arrow labeled Op1 which originates in any other state. The object ∅ in the diagram is the null state. The create arrow represents the creation of the object by assigning an identifier to the object and setting its state to the initial defined state, and destroy arrow represents its destruction.

6 Conclusions

As the knowledge about fungi species grows and new methods become available one can anticipate a fundamental change in the current fungal taxonomy structure. We believe category theory has a significant potential to be considered as a supplementary tool to capture and represent the full semantics of ontology driven applications and it can provide a formal basis for analyzing complex evolving biomedical ontologies. For the future research we plan to generalize our usage of category theory along with other formalisms such as Petri nets, Named graphs and Description Logics in order to improve ontological conceptualization change management. For ontology versioning we also plan to use category theory to determine the degree of semantic similarity between different ontology versions. In addition the work on employing other categorical constructors such as *pushoust* and *pullbacks* for analyzing changes in taxonomical structures is still in progress.

References

1. Baker, C.J.O., Shaban-Nejad, A., Su, X., Haarslev, V., Butler, G.: Semantic Web Infrastructure for Fungal Enzyme Biotechnologists. Journal of Web Semantics 4(3), 168–180 (2006)
2. Santos, G., Villela, K., Schnaider, L., Rocha, A., Travassos, G.: Building Ontology Based Tools for a Software Development Environment. In: Melnik, G., Holz, H. (eds.) LSO 2004. LNCS, vol. 3096, Springer, Heidelberg (2004)
3. Rosen, R.: The Representation of Biological Systems from the Standpoint of the Theory of Categories. Bulletin of Mathematical Biophysics 20, 245–260 (1958)
4. Bernabé, M., Ahrazem, O., Prieto, A., Leal, J.A.: Evolution of Fungal Polysaccharides F1SS and Proposal of Their Utilisation as Antigenes for Rapid Detection of Fungal Contaminants. E. Journal of Env., Agr. & Food Chem. 1, 30–45 (2002)
5. McLaughlin, D., Rinard, P., Cassutt, M.: Discovery about evolution of fungi has implications for humans. University of Minnesota (20 October, 2006)
6. Nikoh, N., Hayase, N., Iwabe, N., Kuma, K., Miyata, T.: Phylogenetic relationships of the kingdoms Animalia, Plantae and Fungi, inferred from 23 different protein species. Mol. Biol. Evol. 11, 762–768 (1994)
7. Heywood, V.H. (ed.): Global Biodiversity Assessment. Cambridge University Press, Cambridge (1995)
8. Bartnicki-Garcia, S.: The cell wall in fungal evolution. In: Evolutionary biology of the fungi, pp. 389–403. Cambridge University Press, New York (1987)
9. Heath, I.B.: Nuclear division: a marker for protist phylogeny. Prog. Protis. 1, 115–162 (1986)
10. LéJohn, H.B.: Biochemical parameters of fungal phylogenetics. Evol. Biol. 7, 79–125 (1974)

11. Hawksworth, D.L., Kirk, P.M., Sutton, B.C., Pegler, D.N.: Ainsworth and Bisby's dictionary of the fungi, 8th edn. Intern. Myco. Institute, Egham, United Kingdom (1995)
12. Guarro, J., Gene, J., Stchigel, A.M.: Developments in fungal taxonomy. Clinical Microbiology Reviews 12(3), 454–500 (1999)
13. Notermans, S., Dufrenne, J., Wijnands, L.M., Engel, H.: H.J. Med.Vet. Mycol.26, 41–48 (1988)
14. Frisvad, J.C., Bridge, P.D., Arora, D.K.: Fungal chemical taxonomy. Marcel Dekker, Inc., New York-Basel-Hong Kong (1998)
15. Crous, P.W.: Plant pathology is lost without taxonomy. Outlooks on Pest Management 16, 119–123 (2005)
16. Crous, P.W., Groenewald, J.Z., Gams, W.: Eyespot of cereals revisited: ITS phylogeny reveals new species relationships. European J. Plant Pathol. 109, 841–850 (2003)
17. Lucking, R., Serusiaux, E., Maia, L.C., Pereira, E.C.G.: A Revision of the Names of Foliicolous Lichenized Fungi Published by Batista and Co-workers Between 1960 and 1975. The Lichenologist 30(2), 121–191(71) (1998)
18. Krötzsch, M., Hitzler, P., Ehrig, M., Sure, Y.: Category Theory in Ontology Research: Concrete Gain from an Abstract Approach. Technical Report, AIFB, U of Karlsruhe (March 2005)
19. Crous, P.W., Groenewald, J.Z.: Hosts, species and genotypes: opinions versus data. Australasian Plant Pathology 34(4), 463–470 (2005)
20. Hawksworth, D.L.: Fungal diversity and its implications for genetic resource collections. Studies in Mycology 50, 9–17 (2004)
21. Awodey, S.: Category Theory. Oxford University Press, Oxford (2006)
22. Wand, Y.A.: A Proposal for a Formal Model of Objects. In: Kim, W., Lochovsky, F. (eds.) Object-Oriented Concepts, Databases, and Applications, pp. 537–559. ACM Press, New York (1989)
23. Odds, F.C., Arai, T., Di Salvo, A.F., Evans, E.G.V., Hay, R.J., Randhawa, H.S., Rinaldi, M.G., Walsh, T.J.: Nomenclature of fungal diseases, A report from a Sub-Committee of the Intl' Society for Human and Animal Mycology (ISHAM) (1992)
24. Odds, F.C., Rinaldi, M.G.: Nomenclature of fungal diseases. Curr. Top. Med. Mycol. 6, 33–46 (1995)
25. Schüßler, A., Schwarzott, D., Walker, C.: A new fungal phylum, the Glomeromycota: phylog eny and evolution. Mycol. Res. 105(12), 1413–1421 (2001)
26. Swann, E.C., Frieders, E.M., McLaughlin, D.J.: Microbotryum, Kriegeria, and the changing paradigm in basidiomycete classification. Mycologia 91, 51–66 (1999)
27. Eilenberg, S., Mac Lane, S.: General Theory of Natural Equivalences. Transac tions of the American Mathematical Society 58, 231–294 (1945)
28. Mac Lane, S.: Categories for the Working Mathematician (corrected 1994). Springer, Heidelberg (1971)
29. Healy, M.J., Caudell, T.P.: Ontologies and Worlds in Category Theory: Implications for Neural Systems. Axiomathes Journal 16, 165–214 (2006)
30. Caldwell, B., Chisholm, W., Vanderheiden, G., White, J.: Web Content Accessibility Guidelines 2.0. W3C Working Draft 11 March 2004 (2004)
31. Whitmire, S.A.: Object Oriented Design Measurement. John Wiley & Sons, Chichester (1997)
32. Ehresmann, A.EC., Vanbremeersch, J.P.: The Memory Evolutive Systems as a Model of Rosen's Organism-(Metabolic, Replication) Systems, vol. 16, pp. 137–154. Springer, Heidelberg (2006)

Replacing SEP-Triplets in SNOMED CT Using Tractable Description Logic Operators

Boontawee Suntisrivaraporn[1,⋆], Franz Baader[1],
Stefan Schulz[2], and Kent Spackman[3]

[1] TU Dresden, Germany
{meng,baader}@tcs.inf.tu-dresden.de
[2] Freiburg University Hospital, Germany
stschulz@uni-freiburg.de
[3] Oregon Health & Science University, USA
spackman@ohsu.edu

Abstract. Reification of parthood relations according to the SEP-triplet encoding pattern has been employed in the clinical terminology SNOMED CT to simulate transitivity of the part-of relation via transitivity of the is-a relation and to inherit properties along part-of links. In this paper we argue that using a more expressive representation language, which allows for a direct representation of the relevant properties of the part-of relation, makes modelling less error prone while having no adverse effect on the efficiency of reasoning.

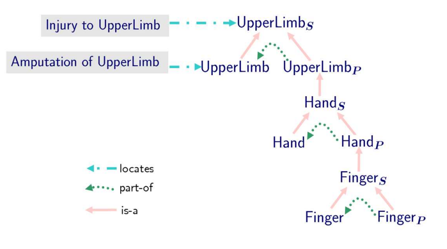
1 Introduction

Description logics (DLs) [1] are a successful family of knowledge representation formalisms, which can be used to represent and reason about ontologies in a logically well-founded way. The Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) [2] is a clinical terminology with a broad coverage of health care, which has been developed with the help of a rather inexpressive description logic dialect known as $\mathcal{EL}$ [3]. In $\mathcal{EL}$, one can build class descriptions using the operators *conjunction* ($C \sqcap D$) and *existential restriction* ($\exists r.C$). For example, the $\mathcal{EL}$ class description $\mathsf{Inflammation} \sqcap \exists \mathsf{has\text{-}location.Appendix}$ describes a kind of inflammation characterized by its location being in some appendix. This description can be used as a definition (expressed by the DL symbol $\equiv$) for appendicitis: it constitutes both necessary and sufficient conditions for classifying a real world entity as being an instance of appendicitis. Classes defined this way are said to be *fully defined*. If only necessary conditions are given for a class, it is called *primitively defined* (expressed by the DL symbol $\sqsubseteq$). For instance, $\mathsf{LeftHand} \sqsubseteq \mathsf{BodyPart} \sqcap \mathsf{LeftLateral}$ is such a primitive definition.

DL systems provide their users with automated reasoning services, which can be used to infer implicit knowledge from the explicitly represented knowledge. In particular, they can *classify* an ontology, i.e., compute all the implied is-a

[⋆] Funded by the German Research Foundation (DFG) under grant BA 1122/11-1.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 287–291, 2007.



$$\begin{array}{lll}
\mathsf{UpperLimb} & \sqsubseteq & \mathsf{UpperLimb}_S \\
\mathsf{UpperLimb}_P & \sqsubseteq & \mathsf{UpperLimb}_S \sqcap \\
 & & \exists \mathsf{part\text{-}of}.\mathsf{UpperLimb} \\
\mathsf{Hand}_S & \sqsubseteq & \mathsf{UpperLimb}_P \\
\mathsf{Hand} & \sqsubseteq & \mathsf{Hand}_S \\
\mathsf{Hand}_P & \sqsubseteq & \mathsf{Hand}_S \sqcap \exists \mathsf{part\text{-}of}.\mathsf{Hand} \\
\mathsf{Finger}_S & \sqsubseteq & \mathsf{Hand}_P \\
\mathsf{Finger} & \sqsubseteq & \mathsf{Finger}_S \\
\mathsf{Finger}_P & \sqsubseteq & \mathsf{Finger}_S \sqcap \exists \mathsf{part\text{-}of}.\mathsf{Finger}
\end{array}$$

Fig. 1. Complete SEP-triplets in SNOMED CT

relationships (i.e., subclass/superclass relationships, expressed by the *subsumption* symbol $\sqsubseteq$) between (names of) fully or primitively defined classes. The advantage of using the inexpressive DL $\mathcal{EL}$ for developing SNOMED CT is that classification is *tractable* (i.e., the is-a hierarchy can be computed in polynomial time). Efficiency and scalability of reasoning are very important for an ontology of the size of SNOMED, with about 370,000 classes. The disadvantage is that not all relevant properties can be explicitly expressed. In particular, $\mathcal{EL}$ does not allow to state that relations such as part-of are transitive, and consequently the reasoner does not take transitivity into account during classification. For example, even if the finger is defined to be part of the hand, and the hand to be part of the upper limb, an $\mathcal{EL}$ reasoner cannot deduce that the finger is part of the upper limb since it does not "know" that part-of is supposed to be transitive.

In order to overcome such limitations in DLs without transitive relations, the SEP-triplet encoding was proposed in [4]. In the next section, we will briefly sketch this approach, and also show that, in addition to transitivity reasoning, it can encode inheritance of properties along part-of links. For example, injury to finger can thus be classified as subclass of injury to hand. We will then point out some disadvantages of the SEP-triplet encoding, and propose to replace SEP-triplets by the direct representation of transitive relations in the DL $\mathcal{EL}^+$ [5]. In addition to transitive relations, $\mathcal{EL}^+$ can also express so-called right-identity rules [6], which can be used to explicitly represent inheritance of properties along part-of and other relations. In spite of its higher expressive power compared to $\mathcal{EL}$, reasoning in $\mathcal{EL}^+$ is still tractable [3,5]. In fact, we will see that not only does the replacement make the classification reasoning faster, but it also helps simplify the ontology structure and thus ease the modelling and maintenance.

2 SEP-Triplets in SNOMED CT

SEP-triplets are extensively employed in the anatomical part of SNOMED CT. Figure 1 illustrates the encoding technique with an example. The left-hand side of the figure provides a graphical representation, whereas the right-hand side shows the formal representation in $\mathcal{EL}$. For every proper SNOMED class, called *entity* class (E-class) in the following, there are two auxiliary classes, the *structure* class (S-class) and the *part* class (P-class). In the example, we have three entity classes: Finger, Hand and $\mathsf{UpperLimb}$, and thus three triplets. Intuitively, the E-class is supposed to be instantiated by entire anatomical objects (such as my hand), and the P-class by the proper parts of the referred objects (such as any part of my hand). The S-class, finally, is instantiated by both entire objects and their parts. This intuition explains the is-a links from the E-class and the P-class to the S-class, as well as the part-of link from the P-class to the E-class. The main idea underlying the SEP-triplet approach is to represent a part-whole relationship between two entity classes not by a part-of link between the E-classes, but rather by an is-a link between the S-class of the "part" and the P-class of the "whole". It should be noted, however, that the formal representation of the intuition underlying the three classes of the SEP-triplet approach is in fact limited to these links, and thus only consequences that follow from the presence of these links can be drawn. This is, however, sufficient to simulate transitivity of part-of through the inherently transitive relation is-a: $\mathsf{Finger} \sqsubseteq \mathsf{Finger}_S \sqsubseteq \mathsf{Hand}_P \sqsubseteq \mathsf{Hand}_S \sqsubseteq \mathsf{UpperLimb}_P \sqsubseteq \exists \mathsf{part\text{-}of}.\mathsf{UpperLimb}$ allows us to conclude that every finger is part of some upper limb.

Since characteristics are inherited along the is-a hierarchy, the SEP-triplet encoding also allows us to simulate inheritance of characteristics along the part-of hierarchy. In our example, by connecting an injury via a location link to the *S-class*, we can ensure that 'injury to finger' is classified as 'injury to hand' and 'injury to upper limb'. To suppress such inheritance along the part-of hierarchy (viz., 'amputation of finger' should not be classified as 'amputation of hand' or 'amputation of upper limb'), one needs to connect via location to the *E-class*.

There are, however, several problems with the SEP-triplet encoding. First, from a formal ontological point of view, it partially conflates the is-a hierarchy with the part-of hierarchy, which is dangerous since the two relationships are completely different by nature [7]. In SNOMED, it has indeed turned out that is-a links can be ambiguous, i.e., it is not always clear whether they are introduced as part of the SEP-triplet approach, or are supposed to represent a genuine generalization relationship. Second, the SEP-triplet approach is error prone since it works correctly only if it is employed with a very strict modelling discipline. In SNOMED, triplets are often modelled in an incomplete way, in particular, the P-class and the part-of link to it from the E-class are missing in most cases. In addition, the auxiliary S-class is often used as if it were a proper entity class; for instance, incorrect links to this class rather than the E-class may result in unintended consequences like the classification of 'amputation of finger' as a subclass of 'amputation of upper limb'. Third, the approach introduces for every proper class in the ontology two auxiliary classes, which results in a drastic increase in the ontology size.

3 Replacing SEP-Triplets by Using the DL $\mathcal{EL}^+$

The DL $\mathcal{EL}^+$ extends $\mathcal{EL}$ with relation inclusions of the form $r_1 \circ \ldots \circ r_n \sqsubseteq s$, which express that the composition of the relations $r_1, \ldots, r_n$ must be interpreted as a subset of the relation s. These inclusions generalize several expressive means

$$\begin{aligned}
\mathsf{Finger} &\sqsubseteq \mathsf{BodyPart} \sqcap \exists\mathsf{proper\text{-}part\text{-}of}.\mathsf{Hand} && (1)\\
\mathsf{Hand} &\sqsubseteq \mathsf{BodyPart} \sqcap \exists\mathsf{proper\text{-}part\text{-}of}.\mathsf{UpperLimb} && (2)\\
\mathsf{UpperLimb} &\sqsubseteq \mathsf{BodyPart} && (3)\\
\mathsf{AmputationOfFinger} &\equiv \mathsf{Amputation} \sqcap \exists\mathsf{has\text{-}exact\text{-}location}.\mathsf{Finger} && (4)\\
\mathsf{AmputationOfHand} &\equiv \mathsf{Amputation} \sqcap \exists\mathsf{has\text{-}exact\text{-}location}.\mathsf{Hand} && (5)\\
\mathsf{AmputationOfUpperLimb} &\equiv \mathsf{Amputation} \sqcap \exists\mathsf{has\text{-}exact\text{-}location}.\mathsf{UpperLimb} && (6)\\
\mathsf{InjuryToFinger} &\equiv \mathsf{Injury} \sqcap \exists\mathsf{has\text{-}location}.\mathsf{Finger} && (7)\\
\mathsf{InjuryToHand} &\equiv \mathsf{Injury} \sqcap \exists\mathsf{has\text{-}location}.\mathsf{Hand} && (8)\\
\mathsf{InjuryToUpperLimb} &\equiv \mathsf{Injury} \sqcap \exists\mathsf{has\text{-}location}.\mathsf{UpperLimb} && (9)\\
\mathsf{proper\text{-}part\text{-}of} \circ \mathsf{proper\text{-}part\text{-}of} &\sqsubseteq \mathsf{proper\text{-}part\text{-}of} && (10)\\
\mathsf{proper\text{-}part\text{-}of} &\sqsubseteq \mathsf{part\text{-}of} && (11)\\
\mathsf{part\text{-}of} \circ \mathsf{part\text{-}of} &\sqsubseteq \mathsf{part\text{-}of} && (12)\\
\epsilon &\sqsubseteq \mathsf{part\text{-}of} && (13)\\
\mathsf{has\text{-}exact\text{-}location} &\sqsubseteq \mathsf{has\text{-}location} && (14)\\
\mathsf{has\text{-}location} \circ \mathsf{proper\text{-}part\text{-}of} &\sqsubseteq \mathsf{has\text{-}location} && (15)
\end{aligned}$$

Fig. 2. A re-engineered extract of SNOMED CT without SEP-triplets

useful in bio-medical ontologies: *(i)* transitivity of r as $r \circ r \sqsubseteq r$, *(ii)* reflexivity of r as $\epsilon \sqsubseteq r$ (where ϵ stands for the empty composition), *(iii)* relation hierarchies as $r \sqsubseteq s$, and *(iv)* right-identity rules as $r \circ s \sqsubseteq r$. It has been shown in [3] that the presence of such axioms does not increase the complexity of reasoning—classification in $\mathcal{EL}^+$ is still tractable.

When replacing the SEP-triplet encoding by the direct representation of transitivity of the part-of relation, we must be careful not to disrupt the rest of the ontology. Especially since the proper classes representing entire anatomical objects as well as the auxiliary S- and P-classes are used by definitions in other parts of the ontology, we must still be able to describe them if needed. Most importantly, we must be able to deduce the same consequences from the direct representation that could be drawn from the SEP-triplet encoding.

Figure 2 shows the part of the re-engineered ontology that corresponds to our example. First, note that we now distinguish between the part-of relation (which is reflexive and transitive) and the proper-part-of relation (which is transitive and a sub-relation of part-of).[1] The direct representation of transitivity allows us to draw the same consequences as in the SEP-triplet approach (e.g., that the finger is part of the upper limb), but dispenses with the auxiliary classes. Whenever any of the P- and S-classes are needed (e.g., since they occur in other parts of the ontology) they can be pre-coordinated as fully defined classes, as illustrated here for the class hand: $\mathsf{Hand}_P \equiv \exists\mathsf{proper\text{-}part\text{-}of}.\mathsf{Hand}$ and $\mathsf{Hand}_S \equiv \exists\mathsf{part\text{-}of}.\mathsf{Hand}$. Note that we need no explicit is-a relationships among the three nodes in a triplet. Because part-of is reflexive, it is inferred that $\mathsf{Hand} \sqsubseteq \exists\mathsf{part\text{-}of}.\mathsf{Hand} \sqsubseteq \mathsf{Hand}_S$. Analogously, $\mathsf{Hand}_P \sqsubseteq \exists\mathsf{proper\text{-}part\text{-}of}.\mathsf{Hand} \sqsubseteq \exists\mathsf{part\text{-}of}.\mathsf{Hand} \sqsubseteq \mathsf{Hand}_S$, since part-of is a super-relation of proper-part-of.

[1] A more precise modelling, which expresses that part-of has to be interpreted as reflexive closure of proper-part-of is not possible since it would cause intractability.

In order to allow for inheritance of characteristics along the proper-part-of hierarchy, we must explicitly state this inheritance property by a right-identity rule (see (15) in Fig. 2). To avoid unintended inheritance of characteristics (e.g., in the case of amputation), we use two distinct relations: has-location, which is inherited from a part to its whole, and has-exact-location, a sub-relation of has-location, which is not inherited that way. Intuitively, has-exact-location associates an event with a location in which it happens as a whole, for instance, 'amputation of upper limb' happens exactly to the upper limb as a whole and not just any part of it. In contrast, has-location relates an event to any containing spatial location it occurs in, i.e., either part or whole of the specified location. For instance, 'injury to upper limb' happens to the upper limb as a whole or any of its parts.

The proposed re-engineering has been put into practice by experimenting with the anatomy fragment of SNOMED CT. Although the SEP model has been adopted in SNOMED CT, it is incomplete in the sense that many SEP-triplets consist of only one or two nodes, and the correct is-a and part-of links are not always present. For this reason, it required a considerable effort to locate and complete all triplets, in order to enable a correct replacement. However, the obtained results are quite promising: by our re-engineering, the number of anatomical classes dropped from 54,380 to 18,125, and the time needed by our CEL reasoner (version 0.94) [5] from 900.15 seconds to 18.99 seconds. An empirical analysis of our proposed re-engineering of the entire SNOMED CT ontology still needs to be done, however. In particular, this will show how the introduction of right-identity rules to enable inheritance of characteristics along the aggregation hierarchy and the introduction of two different relations for location influence classification time.

References

1. 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, U.K (2003)
2. SNOMED Clinical Terms. College of American Pathologists, Northfield, IL (2006)
3. Baader, F., Brandt, S., Lutz, C.: Pushing the EL envelope. In: Proc. of the Nineteenth Int. Joint Conf. on Artificial Intelligence (IJCAI-05), Edinburgh, UK, 2005, Morgan-Kaufmann Publishers, San Francisco (2005)
4. Schulz, S., Romacker, M., Hahn, U.: Part-whole reasoning in medical ontologies revisited: Introducing SEP triplets into classification-based description logics. In: Chute, C.G. (ed.) Proc. of the 1998 AMIA Annual Fall Symposium, Hanley & Belfus, pp. 830–834 (1998)
5. Baader, F., Lutz, C., Suntisrivaraporn, B.: CEL—a polynomial-time reasoner for life science ontologies. In: Furbach, U., Shankar, N. (eds.) IJCAR 2006. LNCS (LNAI), vol. 4130, pp. 287–291. Springer, Heidelberg (2006)
6. Spackman, K.A.: Managing clinical terminology hierarchies using algorithmic calculation of subsumption: Experience with SNOMED-RT, OHSU Technical Report (2000)
7. Patrick, J.: Aggregation and generalisation in SNOMED CT. In: Proc. of the 1st Semantic Mining Conf. on SNOMED CT, Copenhagen, Denmark (2006)

Building an Ontology of Hypertension Management

Olivier Steichen, Christel Daniel-Le Bozec, Marie-Christine Jaulent, and Jean Charlet

INSERM, UMR_S 872, Eq. 20, F-75006 Paris, France
ost@club-internet.fr

Abstract. The analysis of customized decisions during hypertension management in a specialized unit requires a detailed representation of clinical cases. We are building a specific ontology to code medical records and process them with computerized tools. Relevant concepts to describe and justify medical decisions are extracted from three sources: (i) Clinical guidelines; (ii) Items of the semi-structured medical record form used in the clinical unit; (iii) Free-text answers from 5,000 completed record forms. Combining terminological sources is mandatory to cover the whole spectrum of possible justifications for clinical decisions, including contextual specificities and patients' particulars.

Keywords: Ontologies, Computerized Medical Records, Guidelines.

1 Introduction

1.1 Qualitative Practice Assessment

Clinical guidelines are often taken as references for practice assessment. However, clinical decisions sometimes have to be customized, beyond or against guidelines, with regard to the characteristics of each patient and specific clinical circumstances. However, even customized decisions must remain justified. They can be assessed with qualitative methods that rely on detailed and comparative analysis of individual cases [1]. Computerized tools are mandatory to achieve such studies on a large scale. They can help clustering similar cases according to clinical characteristics and they can ease the comparative study of medical decisions, provided that medical records are formally represented with enough details for the intended level of analysis.

1.2 Medical Ontologies Modeling

Ontologies are considered as the richest and most formalized terminological resources to share and process the semantic content of electronic health records. As such, they can meet our requirements for medical records representation. Because each ontology has a specific conceptual scope, granularity level and structure, it usually can't be reused for purposes it wasn't primarily intended

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 292–296, 2007.

for [2]. In order to assess customized clinical decisions in a hypertension unit, we are building a specific ontology of hypertension management to represent the content of medical records. This paper describes our practical ontological modeling process.

Two broad approaches compete to elicit concepts to be included in an ontology. Top-down approaches rely on prior conceptualizations made by field experts (in clinical medicine: structured data entry forms, clinical guidelines, academic books, etc.). In bottom-up approaches, concepts are extracted from the experts' textual production during actual practice (in clinical medicine: medical records, discharge summaries, mails to colleagues, etc.). We followed a custom composite approach, combining top-down and bottom-up steps.

2 Material and Methods

We used three sources to find the concepts used to describe and justify medical decisions during hypertension management: (i) Items of the semi-structured medical record form used in the hypertension unit, representing the generic determinants of practice in specialized settings; (ii) A corpus of clinical guidelines, representing the generic determinants of practice in family medicine; (iii) A corpus of free-text comments from filled record forms, allowing physicians to consign the unforeseen determinants of specialized practice and the justifications of difficult decisions.

A computerized semi-structured medical record is used for 30 years in the hypertension unit and counts 176 questions [3]. The list of clinical items found in the form (headings, questions and predefined answers of checklist questions) has been generated by a SQL query on the hospital database. After anonymization, the free-text answers from the 5,109 clinical record forms filled in year 2005 resulted in a 350,000 words corpus in French. Eight hypertension guidelines, published from 1999 to 2005, were assembled to constitute a 56,000 words corpus in English.

The clinical record form and the guidelines were the starting points of our first two ontological modeling steps, both top-down. They provided core concepts for hypertension management. The free-text answers from completed forms fed a bottom-up modeling step, intended to enrich this ontological core.

The natural language processing tools SYNTEX and UPÉRY were used to extract term candidates (terms likely to represent pertinent concepts in the field) from guidelines and free text answers [4]. Ontological modeling is currently carried out, concept by concept, in the Differential Ontology Editor (DOE) [5]. Concepts are manually linked with SNOMED-CT concepts (January 2006 version), accessed through the Virginia Tech SNOMED-CT Browser [6]. This mapping of our domain ontology with a reference ontology gives an opportunity to evaluate the domain coverage of SNOMED-CT and the adequacy of its organization with regard to our intended use.

3 Results

Concepts from the record form items and from guidelines have been completely extracted. By now, only the most frequent concepts in the free-text comments – occurring at least 50 times in the corpus (about 1% of the records) – have been extracted. These concepts mostly relate to some generic aspects of hypertension management, as do concepts from the record form items and from guidelines. Therefore, we presumably have identified most of the core concepts used for routine hypertension management. We are currently structuring this ontological core. Only concepts coming from the record form items have been linked to SNOMED-CT at the present time.

3.1 Top-Down Steps: Considering Experts' Conceptualizations of Their Field

The 176 questions and 177 predefined answers of checklist questions in the record form provided 243 clinical concepts. The same concept can underlie several items, for example a date question ("date of the last myocardial infarction?") and a predefined answer of a checklist question ("heart history?" ☒ myocardial infarction). The relevance of all these concepts is assumed, given the clinical appropriateness of the form as a result of step-by-step improvements over 30 years. We found a matching SNOMED-CT term for 212 of these concepts. Guidelines analysis uncovered 258 clinical concepts: 163 concepts already found in the form items and 95 additional concepts.

The concepts found in the form items but not in the guidelines were either too specialized (like renal infarction as a cause of hypertension), too fine-grained (like glomerulopathy as a specific type of kidney disease) or related to other cardiovascular risk factors management (like diabetes or dyslipidemia). Two examples of additional concepts found in guidelines but not in the form items are "sleep apnea" as a cause of hypertension and "dementia" as a consequence of hypertension.

3.2 Bottom-Up Steps: Considering Free-Text Answers in Medical Records

The analysis of term candidates in free-text comments revealed 233 frequent concepts (more than 50 occurrences). Among them, 162 had already been found in the record items or in the guidelines and 71 were completely new.

Concepts found in guidelines and actually used by physicians in free-text answers are undoubtedly valuable for the management of patients in the specialized unit. For example, free-text comments count 89 occurrences of "sleep apnea" or of semantically related terms, like "snoring". On the other hand, no occurrence of "dementia" was found in free-text comments. According to the physicians, dementia is not an issue faced in the hypertension unit. Patients are simply not referred by their general practitioner if they suffer from disabling cognitive impairment.

The analysis of free-text comments also identified concepts missing in guidelines but regularly used by physicians to state or justify their decisions. Most of these concepts are very specialized and refer to the specific recruitment of the unit (high prevalence of secondary and/or complicated hypertension), whereas guidelines are mainly intended for the management of primary and uncomplicated hypertension in general practice. For example, the concept of renin – aldosterone dissociation, related to specific causes of hypertension, has 80 occurrences in free-text comments and is an important determinant of patients' management, with respect to local decision rules or habits.

4 Discussion – Conclusion

4.1 The Need for a Customized Domain Ontology

An ontology intended to support a specific task in a specialized domain requires a definite expressive power (scope and granularity level) and processing potential (structure). The specific shortcomings hindering a direct reuse of SNOMED-CT for our application were highlighted during concept mapping. Some SNOMED-CT terms are ambiguous and unequivocal mapping was sometimes difficult. For example, the concept of glomerulopathy, found as a kidney disorder in the record form, could be mapped with two brother SNOMED-CT terms, which are clearly synonyms for a clinician: "renal glomerular disease" and "glomerular disease". As expected, the scope of SNOMED-CT is not broad enough. Thirty-one of the 243 fundamental clinical concepts coming from the form items were lacking. For example, there is no term in SNOMED-CT referring to monogenic hypertension or drug induced hypertension. Finally, the structure of SNOMED-CT is presumably inadequate for our purposes, because it doesn't reflect the way clinicians picture the field of hypertension medicine. For example, the structuring concept of target organ damage, which determines many management decisions, is not found in SNOMED-CT.

4.2 The Need for a Customized Modeling Approach

Our ontological modeling approach relates with former bottom-up achievements relying on large corpora of texts produced during clinical practice [4,7]. This framework ensures that the ontology comprises the concepts actually used by physicians. Nonetheless, we customized this approach for our particular needs: (i) Corpora used in former works were collections of discharge summaries whereas we worked with medical records; (ii) We added top-down steps, exploiting preexistent experts' conceptualizations of the field.

Knowing that a concept is found among the record form items or in guidelines is a strong hint in favor of its contribution to standardized hypertension management, following explicit or implicit rules. We also considered frequently occurring concepts in free-text comments, which must be related to some aspect of routine hypertension management. The remaining, less frequently occurring concepts in free-text comments are likely to play a role in the individualization of management decisions.

4.3 Future Work

We are currently organizing our ontological core in a strictly taxonomic hierarchy. We will then complete the terminological analysis of free-text comments, in order to integrate the less frequent concepts in the ontology. After a strong initial focus on concepts not related to practice individualization, it should be easier to discern and incorporate concepts more directly related to our purpose.

We will then use the ontology to represent clinical cases managed in the hypertension unit. Free-text answers can not be automatically represented with the concepts they issued. The association of matching concepts to each case can only be automated for the structured part of medical records. Indeed, the form items are unequivocally linked with the concepts they gave birth to. Once this partial representation achieved, it will be possible (i) to identify cases managed outside guidelines' recommendations and (ii) to cluster similar such cases thanks to semantic similarity measures.

Within cluster comparisons, i. e. qualitative analysis of similar cases, will lead to uncover further pertinent characteristics of cases and their ontological representation will be manually complemented. After a first loop, it will be possible to compute semantic similarity anew on the extended ontological representations. Finer clusters of similar cases will ensue, allowing a finer qualitative analysis. The process may be looped as many times as useful and possible.

Acknowledgements. Didier Bourigault provided the statistical results of corpus analyses with SYNTEX and UPÉRY. This work benefits from a grant of the AP-HP and from a subsidy of the French Society for Hypertension.

References

1. Green, J., Britten, N.: Qualitative Research and Evidence Based Medicine. BMJ 316, 1230–1232 (1998)
2. Coiera, E.: Medical Informatics. BMJ 310, 1381–1387 (1995)
3. Degoulet, P., Chatellier, G., Devriès, C., Lavril, M., Ménard, J.: Computer-Assisted Techniques for Evaluation and Treatment of Hypertensive Patients. Am J Hypertens. 3, 156–163 (1990)
4. Charlet, J., Bachimont, B., Jaulent, M.: Building Medical Ontologies by Terminology Extraction from Texts: An Experiment for the Intensive Care Units. Comput Biol Med. 36, 857–870 (2006)
5. Troncy, R.: Differential Ontology Editor (Last accessed on the 2006/09/09) http://homepages.cwi.nl/~troncy/DOE/
6. Virginia-Maryland Regional College of Veterinary Medicine: SNOMED-CT Browser (Last accessed on the 2007/01/30) http://snomed.vetmed.vt.edu/sct/menu.cfm
7. Baneyx, A., Charlet, J., Jaulent, M.: Building An Ontology Of Pulmonary Diseases With Natural Language Processing Tools Using Textual Corpora. Int J Med Inform. 76, 208–215 (2006)

Analyzing Differences in Operational Disease Definitions Using Ontological Modeling

Linda Peelen[1], Michel C.A. Klein[2], Stefan Schlobach[2], Nicolette F. de Keizer[1], and Niels Peek[1]

[1] Dept. of Medical Informatics, Academic Medical Center, Amsterdam
{l.m.peelen,n.b.peek,n.f.keizer}@amc.uva.nl
[2] Dept. of Artificial Intelligence, Vrije Universiteit Amsterdam
michel.klein@cs.vu.nl, schlobac@few.vu.nl

Abstract. In medicine, there are many diseases which cannot be precisely characterized but are considered as natural kinds. In the communication between health care professionals, this is generally not problematic. In biomedical research, however, crisp definitions are required to unambiguously distinguish patients with and without the disease. In practice, this results in different operational definitions being in use for a single disease. This paper presents an approach to compare different operational definitions of a single disease using ontological modeling. The approach is illustrated with a case-study in the area of severe sepsis.

1 Introduction

In medicine, many diseases cannot be unequivocally defined by etiology or anatomical localization, but are instead described by a combination of signs and symptoms that are common in patients believed to be suffering from that disease.

An example of such a disease is the syndrome of *severe sepsis*. In this disease the immune system of the patient overreacts to an infection. If untreated, the patient becomes severely ill, which may result in organ failure and eventually death. The cause of severe sepsis is largely unknown, and the disease is not restricted to an exact anatomical localization, which hinders the precise characterization of the patient.

In daily patient care and in communication between health care professionals such a lack of precision is often not problematic. However, when the purpose of describing patients is to select patients for medical research or to automatically reason with patient data (e.g., in triggering computerized guidelines) a crisp disease definition is required, which unambiguously distinguishes patients with the disease from persons without the disease. In practice, this often results in ad hoc, operational definitions that largely cover the intended patient group.

It is questionable to which extent patients selected by different definitions can be compared. This is an important issue in, for instance, statistical aggregation of data, meta-analysis of medical scientific evidence, and in the design of clinical studies.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 297–302, 2007.

In previous work we have shown that nine recent clinical trials in the area of severe sepsis all used different operational definitions. Applying these definitions onto real clinical data resulted in the selection of patient groups with different outcome characteristics [1].

In this paper we present an approach to systematically compare different operational definitions of a single disease using *ontological modeling*. First we present a general abstraction hierarchy which indicates the levels at which the concepts related to the operational definitions are expressed. Subsequently we propose a method that uses this hierarchy to compare complex definitions at different levels of abstraction.

Throughout the paper we will use two operational definitions of severe sepsis as an example, which are depicted in Table 1. When comparing these definitions, we note that both definitions have *polythetic* aspects: a list of signs and symptoms is given, of which a particular number has to be fulfilled, and some of which are necessary conditions.

Table 1. Definitions for severe sepsis used in the PROWESS[2] and Kybersept[3] trial

	PROWESS	
Known or suspected infection or signs of pneumonia	At least three of the modified SIRS criteria: 1)temperature $\geq 38°$C or temperature ≤ 36 °C 2) heart rate ≥ 90/min 3) respiratory rate ≥ 20/min or PaCO$_2$ ≤ 32 mmHg or mechanical ventilation 4)leukocyte count $\geq$12,000/mm^3 or leukocyte count $\leq$4,000/mm^3 or $>$10 % immature neutrophils	At least one out of: 1) pH ≤ 7.3 or base deficit ≥ 5.0 mmol/L with plasma lactate > 1.5 times higher than normal 2) urine output < 0.5 mL/kg/hr 3) thrombocyte count $< 80 \cdot 10^3$ /mm^3 4) PaO$_2$/FiO$_2$ ≤ 250 or ≤ 200 if no other organ dysfunction present 5) systolic blood pressure < 90 mmHg or mean arterial pressure ≥ 70 mmHg or use of vaso-active medication
	Kybersept	
Suspected infection	temperature $> 38.5°$C or temperature < 35.5 °C AND leukocyte count $>$10,000/mm^3 or leukocyte count $<$3,500 /mm^3	At least three out of: 1) heart rate > 100/min 2) respiratory rate $>$24/min or mechanical ventilation 3) plasma lactate higher than normal or pH < 7.30 or base excess -10 mmol/L 4) urine output < 20 mL/hr 5) thrombocyte count $<100 \cdot 10^3$/mm^3 6) systolic blood pressure < 90 mmHg or use of vaso-active medication

2 Analyzing Differences in Definitions

This section describes the abstraction levels that can be distinguished in operational disease definitions (Section 2.1), and explains how these levels are used in comparing different definitions (Section 2.2).

2.1 Levels of Operationalization

When describing a disease which is considered a natural kind, in fact, operationalization takes place at different levels. These levels form an abstraction hierarchy for concepts that are used in operational disease definitions, which is depicted in Table 2. Four different levels are distinguished. On the first, most abstract, level, concepts are expressed in terms of the *condition* of the patient. The second level focuses on *signs and symptoms*. Signs and symptoms are further operationalized using terms related to *measurements* that are performed in the patient. On the fourth, most concrete, level, terms are used to describe the *threshold value* for measurements, which distinguishes patients with the sign or symptom from patients without.

When moving through the hierarchy from top to bottom the concepts become more explicit. In daily patient care, it mostly suffices to use terms from the 'Condition' and 'Sign / Symptom' levels. Instead, in operational definitions used for purposes of selection, concepts are mostly expressed in terms of the measurements with their threshold values.

2.2 Using the Operationalization Hierarchy in Comparing Definitions

Differences between operational definitions occur at different levels of operationalization. Identifying differences at the Threshold level is relatively straightforward, e.g., both severe sepsis definitions use the Measurement 'Thrombocyte count', but Kybersept uses '< 100' as a cut-off value, whereas PROWESS requires the thrombocyte count to be lower than 80 $\cdot 10^3/\text{mm}^3$. It is however much more complicated to see to which extent a definition that requires two specific criteria to be present relates to a definition that requires three out of a list of four criteria, such as in our example. Our approach helps to compare the definitions not only at the lowest level of operationalization, but also at higher levels.

The approach consists of three steps. First, a 'disease ontology' is created, which describes all possible operationalization choices for a specific disease in a formalized way. Concepts for similar conditions that are operationalized in different ways

Table 2. Abstraction hierarchy for concepts that are used in the operational definitions of medical conditions

Level	Description	Example
Condition	A collection of symptoms and/or signs of which a given number has to be present for the condition to be present.	Severe sepsis
Sign / Symptom	A characteristic of the patient which is experienced by the patient or can be measured by the physician.	Platelet disorder
Measurement	Result of a measurement performed by the physician.	Low thrombocyte count
Threshold value	Threshold which determines whether the result of the measurement indicates the sign / symptom to be present.	Thrombocyte count $\leq$ 80,000 mm^3

in the definitions are given different names, e.g., KS-Thrombocytopenia and PW-Thrombocytopenia. As formalization language we have used OWL DL, a language recently recommended as a standard for ontology modeling based on Description Logics [4], because it is able to 1) formalize concepts and complex relations (e.g. number restrictions); 2) reason and query at different levels; 3) be used in combination with real patient data. A detailed description of the formalization of the severe sepsis definitions is found in [5].

Second, each disease definition is formulated in terms of the disease ontology. For each element of the definition is determined at which level it is specified, and the appropriate concept from the disease ontology is chosen. In this step, we also specify the polythetic conditions using number restrictions, e.g., KS-SevSepsis $\sqsubseteq$ conditionOf.($>$ 3 hasSymptom.(or KS-Thrombocytopenia KS-Hypotension KS-AbnormalHeartRate KS-AbnormalRespiratoryState KS-Oliguria KS-Acidosis)).

In the third step we compare the reformulated definitions automatically. The question, "can the definitions be considered similar", is easily answered by checking for equivalence between the disease definitions. As this is often not the case, in the second phase further comparisons aim to discover in which parts of the definitions the differences are located. This comparison is based on enforcing equivalence between concepts from different definitions and checking for inconsistencies. This can be done at different levels. To verify whether the thresholds are similar, equivalence is enforced at the Measurements level. This will lead to an inconsistency when the definitions make use of different thresholds. To investigate whether the problem is located at the threshold level only, the ontology can be 'pruned' unto the level of Measurements (i.e., at the threshold level the concepts in both definitions are forced to be equivalent) and again checked for equivalence of the definition and for inconsistencies. These comparisons can be repeated at the higher, more abstract, levels.

For example, Figure 1 depicts a part of the ontology with some elements of the Kybersept and PROWESS definitions of severe sepsis. When we test for equivalence of the PROWESS and Kybersept concepts of 'severe sepsis' we will find that these two definitions are not equivalent (cf. Table 1). To find out which concepts or relations cause the differences, we start at the lowest level. When enforcing equivalence between both 'Low TC count' measurements, we will find an inconsistency, as the trials use different threshold values.

3 Discussion

In this paper we present an approach to systematically compare different operational definitions of a single disease using ontological modeling. Its use has briefly been illustrated with an example in the area of severe sepsis. More extensive examples of reasoning possibilities are given in [5].

The approach we have presented can be applied for several purposes. We are implementing a web-service which assists trial designers in operationalizing

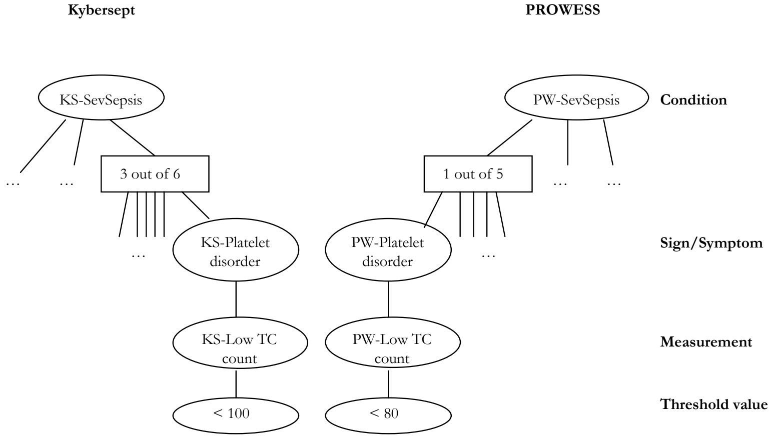


Fig. 1. Part of the disease ontology with some elements of the Kybersept and PROWESS definitions of severe sepsis. (TC count = thrombocyte count).

disease definitions.[1] Current decision-support systems for trial designers focus mainly on procedural, safety, and ethical aspects of the trial protocol (e.g, [6]), whereas in our approach the focus is on the operational definition of the disease. The approach can also be used in meta-analysis of scientific medical results, and in the area of development of computerized versions of clinical guidelines.

In the current DL model we did not use datatype properties to model the 'Threshold' level, but instead created artificial concepts which were in a subsumption relation (e.g., we used VeryLowTCCount-lt80, which is a subclass of LowTCCount-lt100) . In future work we will enhance our approach to allow for more complex reasoning at this lowest level. Furthermore, we will extend the current model with an A-box with real patient data to combine querying the knowledge-based model with querying patient data.

Acknowledgements. The authors would like to thank dr. Ronald Cornet for his contribution to the modeling. MK and NP received a grant from the Netherlands Scientific Organisation (NWO) under project number 634.000.020.

References

1. Peelen, L., De Keizer, N., Peek, N., De Jonge, E., Bosman, R., Scheffer, G.: Influence of entry criteria on mortality risk and number of eligible patients in recent studies on severe sepsis. Crit Care Med 33, 2178–2183 (2005)
2. Bernard, G., Vincent, J.L., Laterre, P.F., et al.: Efficacy and safety of recombinant human activated protein C for severe sepsis. N Engl J Med 344, 699–709 (2001)

[1] The user-interface of the web service is found at `http://prauw.cs.vu.nl/sepsis-trials/`. We are currently implementing the connection to the reasoning engine.

3. Warren, B., Eid, A., Singer, P., et al.: High-dose antithrombin III in severe sepsis. A randomized controlled trial. JAMA 286, 1869–1878 (2001)
4. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. (eds.): The DL Handbook. Cambridge University Press, Cambridge (2003)
5. Peelen, L., Klein, M., Schlobach, S., De Keizer, N., Peek, N.: Analyzing differences in operational disease definitions using ontological modeling with an application in severe sepsis. Technical Report 2007-02, Dept. of Medical Informatics, Academic Medical Center - Universiteit van Amsterdam (2007)
6. Modgil, S., Hammond, P.: Decision support for clinical trial design. Artificial Intelligence in Medicine 27, 181–200 (2003)

Part VI

Decision Support Systems

Adaptive Optimization of Hospital Resource Calendars

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

[1] Centre for Mathematics and Computer Science (CWI),
Amsterdam, The Netherlands
I.B.Vermeulen@cwi.nl
[2] Academic Medical Center, University of Amsterdam, The Netherlands

Abstract. As demand for health care increases, a high efficiency on limited resources is necessary for affordable high patient service levels. Here, we present an adaptive approach to efficient resource usage by automatic optimization of resource calendars. We describe a precise model based on a case study at the radiology department of the Academic Medical Center Amsterdam (AMC). We model the properties of the different groups of patients, with additional differentiating urgency levels. Based on this model, we develop a detailed simulation that is able to replicate the known scheduling problems. In particular, the simulation shows that due to fluctuations in demand, the allocations in the resource calendar must be flexible in order to make efficient use of the resources. We develop adaptive algorithms to automate iterative adjustments to the resource calendar. To test the effectiveness of our approach, we evaluate the algorithms using the simulation. Our adaptive optimization approach is able to maintain overall target performance levels while the resource is used at high efficiency.

1 Introduction

Hospitals continuously aim to improve their patient-oriented care. They want to provide their patients with high service levels. However, the demand for health care is increasing, and more patients must be treated with the same capacity. High efficiency on resources is necessary for high service levels.

Traditional approaches to logistical improvement are usually not suited to the medical domain. The distributed authority in hospitals [1] makes improvements involving many departments difficult to implement. Furthermore, scheduling decisions must be made depending on the individual patient's specific attributes. Efficient scheduling of patient appointments on expensive resources is a complex and dynamic task.

Hospital resources are many: ranging from CT and MRI scanners, to hospital beds, to attending staff. A resource is typically used by several patient groups with different properties [2]. There are groups of inpatients (admitted to the hospital) and outpatients (not admitted), with different levels of urgency [3]. The total hospital resource capacity is allocated to these groups, explicitly or implicitly. Either way, due to fluctuations in demand, this allocation must be flexible to make efficient use of the resources.

To allocate hospital resources, electronic calendar-systems are widely applied. However, they are mostly just storing the patient appointments. The actual scheduling approach largely influences whether resources are efficiently used. Efficient resource usage

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 305–315, 2007.

and short waiting times are of great importance to the hospital departments. Department managers can control options such as changing the capacity, or adjusting the performance goals. It is typically hard to determine the effect of such changes a priori.

In this paper, we study the case of scheduling Computer Tomography scans (CT-scans) at the radiology department within the Academic Medical Centre Amsterdam (AMC). The AMC is a large hospital with approximately one thousand beds, and treating around 350,000 outpatients annually. Currently the radiology department makes more than 15,000 CT-scans per year. Diagnostic resources such as the CT-scanners are literally central in the clinical pathways of many patients. Long access times to such resources are immediately felt as bottlenecks for health care processes in the hospital.

In recent years the logistical process around the CT-scan has improved already substantially [8]. The actual scheduling of appointments is (still) done manually. A calendar supervisor determines a long time in advance how the scanner capacity is allocated to different groups of patients. This allocation is determined based on experience, future expectation, and in cooperation with medical experts.

The calendar supervisor also monitors and adjusts the allocations in the calendar on a regular basis (at least daily) to maintain its efficiency. Often, the actual realization of patient arrival does not match the allocation. This results in inefficient use of the capacity and/or increased waiting time for patients. The calendar supervisor can adjust the calendar to counter such problems. With constant active – and time consuming – supervision of the calendar, the manual scheduling practice performs satisfactory.

However, making good adjustments is critically dependent on the supervisor's expertise: even a short vacation or illness of the calendar supervisor leads to immediate significant deterioration of the resource efficiency. From a planning and sustainability perspective, this is highly unsatisfactory.

Here, we develop an approach to automatically determine effective optimizations of a resource calendar we derive from our case study. Our approach enables the calendar supervisor to quickly implement calendar adjustments, and anticipate – and remedy – the impact of current demand trends of future resource efficiency.

As a model of hospital resource scheduling, we present a precise model for our application case of CT-scan scheduling. Our model and its parameter values are determined from extensive case analysis. Sources include historical data and extensive discussion with experts at various levels in the organization.

We have implemented an extensive computer simulation of the application case. This allows us to study various problem scenarios and scheduling approaches. It can serve as a prototype as the final step towards application.

Furthermore we present our adaptive approach to automatic optimization of resource calendars. In our approach the allocation of capacity to different patient groups is flexible and adaptive to the current and future situation. To maintain high performance levels, our system exchanges capacity between different urgent and non-urgent patients groups. Additionally, resource openings hours can be reduced to increase capacity usage while maintaining high performance levels, or extended to counter increasing waiting time. We extensively evaluate our adaptive approach in our simulated environment,

Other approaches consider how to coordinate patient scheduling, such as [4], and [5]. In this work, human schedulers still make local decisions based on their experiences and knowledge of the individual patient. It supports the current process, making it suitable for fast application. In [6] the author considers updating the allocation of hospital resources to the departments; here, we provide a more operational approach at the level of scheduling patients. In [7] the authors present a first step towards a general model for solving resource conflicts represented as a constraint satisfaction problem, our approach directly focuses on maintaining high performance levels. Our adaptive model can straightforwardly be applied to a wide range of similar situations where resources are shared between patient groups for which an appointment calendar is used.

2 CT-Scan Scheduling Model

In this Section, we define our CT-scan scheduling model. From the AMC electronic calendar system, we have collected the historical data of the appointments made (from October 2005 until March 2006). We have complemented this with data from actual production of CT-scans (November 2005 until January 2006). During this period, some scans were taken without an appointment. We derive patient distributions and scheduling practice from this data, site-visits, and extensive discussions with the human schedulers, the calendar supervisor, and resource manager.

Our model consists of three main parts, discussed in the following sections. One part is the set of arriving **patients** that need to be scheduled for an appointment. Secondly, we have the available resource, and the way the associated appointment **calendar** is structured. The third part is the **scheduling** process that determines how appointments are made by assigning patients to timeslots on the calendar.

2.1 Patients

An important issue is that there is a great variety in patients and scan attributes. We make the abstraction that a patient always needs to be scheduled for exactly one CT-scan. We therefore model the patient and his/her scan as a unity, which we from now on refer to as ’patient’. Patient attributes are listed in Table 1.

In practice, patients and their attributes are structured in different patient groups. Table 2 lists these groups and their specific properties. The group size is given relative to the total number of patients.

The largest group – **out+ivc** – is comprised of non-urgent outpatients who need intravenous contrast (ivc) injected before taking the CT-scan. Non-urgent outpatients who do not need intravenous-contrast are in the group **out–ivc**. All urgent outpatients form the group **urgent**. The fourth group – **clinic** – consists of all inpatients.

Besides these four groups, there are a number of smaller, highly specific groups: $special_n$. These include patients taking part in special programs, and patients who need a specific treatment for making a CT-scan. E.g., one special group is the group of patients, usually children, who need to be sedated while making the CT-scan.

Urgency of patients is defined in terms of planning windows (PLANWIN) with different sizes. Outpatients with normal urgency (out+ivc, and out–ivc) have a planning window of $(2, 14)$, which means that the appointment must be scheduled between 2 days and 14 days after the request for the scan is made. Urgent outpatients and clinic

Table 1. Patient attributes

attribute	**description**
request time	date and time when request for CT-scan is made
in-/outpatient	is the patient an in- or outpatient?
contrast needed? ($\pm ivc$)	does intravenous-contrast need to be injected?
planning window ($planwin$)	expresses urgency of patient.
duration	of the needed appointment

Table 2. Patient groups

group	**urgency**	**planwin** (fraction)	**duration**	**size(%)**
out+ivc	normal	$(2, 14)$	15 mins	$52\% \pm 6\%$
out–ivc	normal	$(2, 14)$	15 mins	$23\% \pm 4\%$,
urgent	high	$(0, 1)(33\%)$, $(0, 2)(33\%)$, or $(0, 3)$	15 mins	$10\% \pm 3\%$
clinic	high	$(0, 1)(40\%)$, or $(0, 2)$	30 mins	$6\% \pm 2\%$
$special_n$	n.a.	n.a.	$duration_n$	$9\% \pm 2\%$

patients have high urgency and have planning windows of a few days. Patients from special groups are always scheduled to the first available timeslot of matching type.

2.2 Resource Calendar

Patients must be scheduled to a timeslot on the calendar. The total resource capacity is given by the number of actual CT-scanners m (in our case $m = 2$) and the opening hours. The hospital's emergency room has an additional CT-scanner, which is used as a walk-in facility for emergencies, which we do not consider in our model.

A standard calendar is used, structured in days and weeks. The time on the calendar is partitioned into timeslots of different sizes. All timeslots have a size of a multitude of the unit size us. (In our case us is 15 minutes, and we use timeslots of sizes $1us$ up to $4us$.) The parameters in Table 3 define the resource calendar. The parameters m and us are fixed for long periods of time, the remaining can vary. Openings hours must be known at least one week in advance to plan staff. In general we assume that the m actual resources are interchangeable.

Timeslot Type Specification. CT-scan capacity is reserved for different patients groups and these allocations serve medical restrictions (e.g. due to preparation constraints for narcosis), as well as a scheduling goal (e.g. reserve timeslots for urgent patients). E.g., on the actual calendar, three timeslots are reserved on all Thursday mornings for patients from a $special_n$ group, who need to be sedated while making the CT-scan. During lunch time, radiologists schedule meetings and other activities. Therefore, out+ivc patients, who need to be injected with intravenous contrast for which a radiologist must be present, cannot be scheduled during lunch. In the afternoon of every day a number of timeslots is reserved for urgent outpatients.

We model this allocation by using a timeslot-type specification (TTS). A timeslot-type specifies which patient can be scheduled to a certain timeslot (Table 4). The TTS

Table 3. Calendar parameters

parameter	description
m	number of resources
$o_{j,d}$	opening time of resource j on day d
$c_{j,d}$	closing time of resource j on day d
us	unit size timeslots
TTS	timeslot type specification

Table 4. Timeslot Types

Timeslot-Type	allowed patients	size
TTout	out+ivc, out–ivc, urgent	$1us$
TT-ivc (during lunch)	out–ivc, urgent(with no ivc)	$1us$
TTurgent	urgent	$1us$
TTclinic	clinic	$2us$
$TTspecial_n$	$special_n$	1–$4us$

thus determines how much of the resource capacity is allocated to the patient groups. The TTS is not necessarily fixed as the capacity allocation can be dynamically altered.

The $TTspecial_n$ type of timeslots can only be used by very specific types of patients. For each of these types there is a rule which states that if there are still any free slots remaining r_n days in advance, these slots are changed to TTout type of timeslots. This rule is currently the only automatic TTS adjustments in operation at the hospital.

2.3 Scheduling

Scheduling is the process of assigning patients to timeslots, i.e. making appointments. In the case we describe, scheduling performance of different approaches is influenced by two things: first, by how well the TTS matches the actual situation, and second, by the actual scheduling method (the selection of a timeslot per patient given the TTS).

As in many hospitals, for the AMC CT-scanners the actual scheduling of appointments is done manually. Human schedulers schedule patients in turn, by looking on the calendar for an available slot, or using the search function of the electronic calendar system. The search returns a list of available timeslots. Human schedulers have expertise in taking the individual patient's attributes into account. They can also use a patient's preference (e.g. for a specific day, or time). However, the human schedulers have little overview on how their local decisions will influence overall performance goals

The calendar supervisor can adjust the TTS in case there is a mismatch between the TTS and the actual realization of patient arrivals. The human schedulers are used to working with different types of timeslots on the calendar. They use their expertise and long-time experience to select timeslots within the scheduling rules.

3 Simulation

Based on the model we have implemented a patient scheduling simulation (PSS). We use the PSS in the evaluation of different scheduling and resource management approaches. The PSS takes as input distributions of patients attributes, the standard resource openings hours and TTS, a scheduling method, a performance measure, and an adaptive model of how to adjust the TTS and openings hours. The PSS generates a patient stream, a filled-in calendar, the used openings hours and a performance value.

Patient Arrival Simulation. With our model of patient properties and the relative request proportions, we can simulate the arrival of all patients during a week. In the simulation, we have structured the arrival process by the following steps for each week:

1. A standard random walk with a drift τ towards the average $\bar{n}$ fits the distribution over the number of patients arrivals per week. The number of patients for next week (n_{w+1}) is determined as a function of the current patient arrivals n_w as:

   $$n_{w+1} = n_w + \mathcal{N}(0, \sigma) + \frac{\bar{n} - n_w}{\tau},$$

   where $\mathcal{N}(0, \sigma)$ a normally distributed fluctuation of patient arrivals. We set: $\bar{n} = 250$, $n_0 = \bar{n}$, $\sigma = 30$, and $\tau = 3$.
2. Given n_w divide the patients over the groups, using the distribution from Table 2, where out+ivc will get the remainder of $n_w \approx 52\%$.
3. Per patient determine request date within week (see below).
4. Per patient determine request time on request day, using a uniform distribution over the opening hours.
5. Per patient determine the planning window using the distributions from Table 2.
6. Order patients by increasing request time within week.

Because of extra rounds for inpatients on Monday and Friday, patient arrival (step 3 above) is slightly structured during the week. On Monday and Friday twice as many requests for CT-scans of inpatients are ordered compared to the other three weekdays. Requests for outpatients arrive uniformly over the week. Note that as resource is closed on Saturday and Sunday, urgent and clinic patients requested on a Friday with a PLANWIN of (0,1) or (0,2) must also be scheduled to that Friday.

Resource Calendar and Scheduling Approach. In our simulation we use a resource calendar, which is similar to the calendar used in practice. Opening time on the calendar is 8:30, while the resource closes at 16:45. To simulate current scheduling practice we use the following scheduling method:

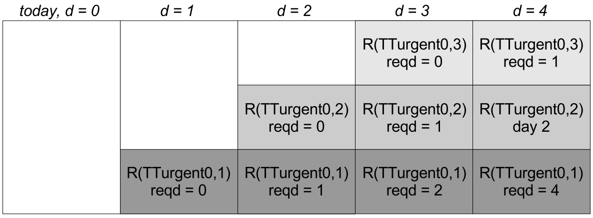
First Come Randomly Served (FCRS). Patients are scheduled in order of arrival. A patient is assigned to a timeslot within his planning window, randomly selected from all the free timeslots of the allowed types. If there are no free timeslots within the planning window, the first free timeslot after the planning window is selected.

For non-urgent patients FCRS simulates the scheduling process where patient preferences are taken into account. We represent this by random allocation to free slots. Urgent and clinic patients have high urgency and thus patient preferences are of little importance. In current practice however, the human schedulers do not take the individual urgency of these patients into account and are therefore also scheduled randomly. We will present a dynamic approach to the scheduling of urgent and clinic patients in the next Section. Additionally our adaptive model for calendar adjustments is input for the scheduling approach used in the PSS.

Performance Measure. Based on discussions with hospital experts, we want our performance measure to express that patients must be scheduled within their planning windows. It is important that each group (G) has a good service level. We define the minimum service level (MSL), over the four main groups of Table 2, as:

$$MSL = \min_G \left(\frac{|\text{patient } \in G = \text{ontime}|}{|G|} \right),$$

where ontime is defined as scheduled within the planning window.



today, d = 0	d = 1	d = 2	d = 3	d = 4

			R(TTurgent0,3)
			reqd = 0
				R(TTurgent0,3)
				reqd = 1

		R(TTurgent0,2)
		reqd = 0
			R(TTurgent0,2)
			reqd = 1
				R(TTurgent0,2)
				day 2

	R(TTurgent0,1)
	reqd = 0
		R(TTurgent0,1)
		reqd = 1
			R(TTurgent0,1)
			reqd = 2
				R(TTurgent0,1)
				reqd = 4

Fig. 1. Reservation within TTurgent

4 Adaptive Model

The TTS and total capacity, and the actual method of scheduling, determine the performance of a scheduling approach. To cope with uncertainty in patient arrival, an additional surplus of capacity above the expected number of urgent and clinic patients must be available. This allocation of capacity in the TTS must be dynamically managed for maximum efficiency. We propose a three-part approach to scheduling and calendar adjustments, to best fit the calendar to current and future situations. Our approach is adaptive to, first, the current (partly filled-in) calendar, and second, the current expectation of the arrival of patients and their attributes.

Adaptive Urgent Scheduling Method. To schedule urgent and clinic patients on time, the allocated capacity must be large enough. Patients with different PLANWINs use the same type of slots: urgent with (0,1); (0,2); (0,3) use TTurgent; clinic with (0,1); (0,2) use TTclinic. In practice, hospital schedulers do not distinguish between different PLANWINs; less urgent patients are regularly scheduled in place of high urgency ones.

To counter this problem we virtually divide urgent capacity while scheduling, by making reservations for different PLANWINs. Within the slots for TTurgent we make reservations (R): $R(TTurgent_{reqd}^{(0,1)})$, $R(TTurgent_{reqd}^{(0,2)})$, and $R(TTurgent_{reqd}^{(0,3)})$, for all $reqd$ ($reqd$ is the request day of the patients relative to today (0), 1 is tomorrow, etc.). To make sure patients are scheduled on time, these reservations are placed on the last day of the PLANWIN: $R(TTurgent_0^{(0,1)})$ on day 1, $R(TTurgent_0^{(0,2)})$ on day 2, etc., see Fig. 1. The same is done for clinic patients: within TTclinic reservations: $R(TTclinic_{reqd}^{(0,1)})$ and $R(TTclinic_{reqd}^{(0,2)})$ are added for all $reqd$. Note that since there can be no reservations on weekends, a large number of reservations are made on Friday; in practice, a corresponding large capacity allocated to urgent patients is found.

Given these reservations, patients are scheduled in first come first serve order, as long as enough timeslots are still available for patients with higher urgency (smaller PLANWIN). Algorithm 1 describes this method for urgent patients. We use a similar algorithm for clinic patients within TTclinic.

By dividing the total capacity as above, we also increase the variance in its usage. To deal with possible occurring problems, we allow for a reservation violation only if the patient is not scheduled on time otherwise. Specifically, in that case the patient is scheduled to the day within his PLANWIN with the most available timeslots regardless of reservations, see Algorithm 2 for urgent patients. We use a similar algorithm for clinic patients. This method makes the capacity division by reservations more flexible.

Algorithm 1. Reservations for urgent patient within TTurgent.

```
 1: $p$ is the current to be scheduled patient at day 0
 2: $R(TTurgent^{pw}_{reqd})$ is the number of TTurgent slots reserved for patients with PLANWIN $pw$
    and have a request date of $reqd$.
 3: $FREE(TTurgent_d)$ is the number of free TTurgent timeslots on day $d$
 4: $TS$ is the first available timeslot within TTurgent
 5: if $planwin == (0,2)$ OR $planwin == (0,3)$ then
 6:   if ($TS$ is on day 1) AND $(FREE(TTurgent_1) \leq R(TTurgent^{(0,1)}_{reqd=0})$ then
 7:     $TS$ = the first available TTurgent timeslot after day 1
 8: if $planwin == (0,3)$ AND $TS$ is on day 2 AND
    $(FREE(TTurgent_2) \leq R(TTurgent^{(0,1)}_{reqd=1}) + R(TTurgent^{(0,2)}_{reqd=0})$ then
 9:   $TS$ = the first available TTurgent timeslot after day 2
10: schedule $p$ to $TS$
```

Algorithm 2. Additional steps to insert between line 9 and 10 in Algorithm 1.

```
1: if $TS$ is outside $planwin$ then
2:   $D$ is day within $planwin$ with most free TTurgent slots
3:   if $FREE(TTurgent_D) > 0$ then
4:     $TS$ = the first available TTurgent timeslot on day D
```

Managing Urgent Capacity. In the previous section we discussed how use reservations in scheduling urgent and clinic patients. If timeslots within the reservations are not used, these could be made available for other groups. In our approach we dynamically manage the surplus capacity allocated to deal with uncertain patient arrival. In general, to maintain high MSL, we can shift capacity between the groups urgent, clinic, and out+ivc. At the start of each day, the total surplus capacity is reallocated between groups, by the following four steps:

1. Change all remaining free TTout capacity of day 0 and 1 into TTurgent.
2. Change free TTurgent capacity on day 0 and 1 above the reservations into TTclinic.
3. Change free TTclinic capacity on day 0 and 1 above the reservations into TTurgent.
4. If free TTurgent capacity on day 0, 1, and 2 is above the capacity of the reservations, this amount of timeslots on day 2 is changed into TTout.

By using this specific order, capacity can be shifted between the three types of TTS. Empty slots of type TTout on day 0 and 1 – which can not be used by out+ivc patients (they have a PLANWIN of (2,14)) – can result in extra TTout timeslots on day 2.

Adjusting Opening Hours. In busy periods, when the total demand reaches or exceeds resource capacity, waiting time can increase rapidly. With a little extra capacity this can usually be avoided. In addition to the adaptive scheduling method described above, we can use a directed search method on our Patient Scheduling Simulator (PSS) to find the required opening hours (OH) for a given desired MSL.

Table 5. Reservation sizes (expected n.o. patients)

	TTurgent	**TTclinc**	**TTclinc**
planwin	all days	mon, fri	tue, wed, thu
(0,1)	3 (1.6)	4 (2)	2 (1)
(0,2)	2 (1.6)	2 (2)	1 (1)
(0,3)	2 (1.6)		

Table 6. Average performance with 41h15min openings hours per week

approach	**perf. (MSL)**	**cap. usage**
FCRS static calendar	0.77 ±0.24	0.90 ±0.04
Reservations	0.80 ±0.27	0.90 ±0.04
Flexible Reservations	0.83 ±0.27	0.90 ±0.04
Fully Adaptive	0.94 ±0.15	0.91 ±0.04
FCRS static, +2,5h	0.93 ±0.15	0.86 ±0.04

5 Experiments

We compare the performance of our fully adaptive approach to benchmark approaches with various levels of adaptivity. We conduct computer experiments to evaluate our adaptive optimization of the scheduling process. In PPS simulations, realistic problem runs are generated. We average performances over 40 runs. Within each run patients arrive during 20 weeks. To avoid start-up effects, we start with a partially filled-in calendar, and measure average performance (MSL) over the last 10 weeks. We use a TTS optimized for an average arrival of patients. In this TTS, there are 18 TTclinic timeslots reserved for an average of 14 (±4) clinic patients per week. There are 34 TTurgent timeslots reserved for an average of 25 (±8) urgent patients per week. Note that patients also arrive randomly during the week.

In various experiments, we have determined the best sizes of the reservations, see Table 5. The shortest planning window needs the most surplus, and patients with lower urgencies can also make use of this surplus (as a result from Alg. 2). First we show results for scheduling methods and adaptive management of urgent capacity, for fixed openings hours. Secondly we show how opening hours can be adjusted to maintain high MSL or increases resource usage.

Fixed Capacity. In Table 6 we present the average performances for four approaches. The first benchmark is the baseline approach using FCRS for all patients, with a fixed resource calendar. This resembles the practical case where the calendar supervisor is absent. The second benchmark is the baseline plus the reservation blocks of Alg. 1. The third benchmark uses flexible reservations (Alg. 2). We compare this with our fully adaptive approach, that additionally manages the urgent capacity. For baseline performance to match our adaptive approach, 2.5 hours per week openings hours are required.

Adaptive Opening Hours. When more patients arrive than expected, waiting time increases exponentially. Adding extra capacity temporarily can prevent this from happening. Our approach can then propose OH changes to resource managers to maintain high performance. In the following experiment we study a specific scenario of 16 weeks with a short busy period: $n_w = 200 | w \leq 4, n_w = 300 | 6 \leq w \leq 11, n_w = 250 | w = 5, w \geq 12$. In Fig. 2 we show the performance (averaged over 10 runs) of the baseline approach with fixed OH, our fully adaptive approach with fixed OH, and our fully adaptive approach with variable OH. We also plot the extra OH (in minutes) used by the fully adaptive approach with variable OH.

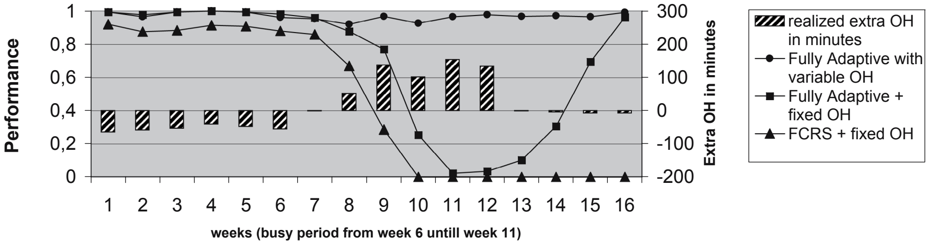


Fig. 2. Performance over weeks with variable and fixed OH

It is clear that a busy period would result in a great decline in performance for the baseline approach. Our fully adaptive approach with fixed OH does decline in performance but reaches good performance quickly after the busy period. The fully adaptive approach with variable OH can adjust the OH such that high performance is maintained over all weeks. Summed over all 16 weeks, it uses almost the same amount of OH as the approaches with fixed OH.

6 Conclusions

We presented a detailed model for the CT-scan scheduling practice at the AMC. This case has similar scheduling problems as other places in the hospital. We describe how the resource calendar is structured and how various patient groups with different levels of urgency are scheduled. The resulting Patient Scheduling Simulator enables us to model various scenarios and to evaluate different allocation and scheduling approaches.

We developed an adaptive approach to the scheduling process and resource calendar management. We showed that this enables us to effectively schedule patients with different urgencies and make efficient use of capacity. By dynamically managing surplus capacity, overall, all patient groups benefit. In current practice this task requires constant attention and is critically dependent on the expertise of the calendar supervisor. Additionally we have shown that we can adjust the opening hours automatically to maintain high service levels. This is an important contribution, because currently it is very hard to determine when and by how much capacity should be extended or reduced to achieve certain patient service levels.

We are currently extending the presented work to cases where appointments must be coordinated between multiple departments, and we are looking into incorporating more patient preferences into appointment scheduling.

Acknowledgement. We would like to thank Han Noot for his programming contributions, and Anieke Eikelenboom for all of her expert knowledge.

References

1. Vissers, J., Beech, R.: Health Operations Management. Routledge (2005)
2. Maruster, L., Weijters, T., de Vries, G., van den Bosch, A., Daelemans, W.: Logistic-based patient grouping for multi-disciplinary treatment. AI in Medicine 26(1-2) (2002)
3. Bowers, J., Mould, G.: Managing uncertainty in orthopaedic trauma theatres. Europ. J. of Operational Research 154 (2004)
4. Decker, K., Li, J.: Coordinating mutually exclusive resources using gpgp. In: Alonso, E., Kudenko, D., Kazakov, D. (eds.) Adaptive Agents and Multi-Agent Systems. LNCS (LNAI), vol. 2636, Springer, Heidelberg (2003)
5. Vermeulen, I., Bohte, S., Somefun, D., Poutr J.L., é.: Improving patient activity schedules by multi-agent pareto appointment exchanging. In: Proc. IEEE International Conference on E-Commerce Technology (CEC/EEE) (2006)
6. Vissers, J.M.: Patient flow-based allocation of inpatient resources: A case study. Europ. J. of Operational Research 105 (1998)
7. Oddi, A., Cesta, A.: Toward interactive scheduling systems for managing medical resources. Artificial Intelligence in Medicine 20(2) (2000)
8. Elkhuizen, S., van Sambeek, J., Hans, E., Krabbendam, J., Bakker, P.: Applying the variety reduction principle to management of ancillary services. Health Care Management Review 32 (2007)

On the Behaviour of Information Measures for Test Selection

Danielle Sent[1] and Linda C. van der Gaag[2]

[1] Department of Electrical Engineering, Mathematics and Computer Science,
University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
danielle.sent@utwente.nl

[2] Department of Information and Computing Sciences, Utrecht University,
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
linda@cs.uu.nl

Abstract. In diagnostic decision-support systems, a test-selection facility serves to select tests that are expected to yield the largest decrease in the uncertainty about a patient's diagnosis. For capturing diagnostic uncertainty, often an information measure is used. In this paper, we study the Shannon entropy, the Gini index, and the misclassification error for this purpose. We argue that for a large range of values, the first derivative of the Gini index can be regarded as an approximation of the first derivative of the Shannon entropy. We also argue that the differences between the derivative functions outside this range can explain different test sequences in practice. We further argue that the misclassification error is less suited for test-selection purposes as it is likely to show a tendency to select tests arbitrarily. Experimental results from using the measures with a real-life probabilistic network in oncology support our observations.

Keywords: Shannon entropy, Gini index, misclassification error, test selection.

1 Introduction

In many fields of medicine, physicians have to establish a diagnosis and have to decide upon an appropriate therapy in relative uncertainty about a patient's true condition. To assist physicians in their complex reasoning processes, sophisticated decision-support systems are being developed. Such a system is often equipped with a test-selection facility that serves to indicate which tests had best be performed to decrease the uncertainty about the patient's diagnosis [1,2]. The two most commonly used measures for capturing diagnostic uncertainty in decision-support systems, are the Shannon entropy and the Gini index [6]; in other contexts, also the misclassification error is used for measuring uncertainty [5]. The three measures are defined for a probability distribution over a designated diagnostic variable and express the expected amount of information that is required to establish the value of this variable with certainty.

The Shannon entropy and the Gini index are generally considered to behave alike for test-selection purposes, in particular for diagnostic variables with a small number of values [3]. In fact, common knowledge has it that the two measures are interchangeable in practice. In this paper, we compare the Shannon entropy, the Gini index and

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 316–325, 2007.

the misclassification error from a fundamental perspective. By studying the first derivatives of the three functions, we argue that for a large range of probability distributions over the main diagnostic variable, the Shannon entropy and the Gini index are indeed expected to behave alike. For the more extreme probability distributions, however, the two measures are expected to result in different test sequences. We further argue that the misclassification error is less suited for test-selection purposes as it is likely to show a tendency to select tests randomly.

We studied the Shannon entropy and the Gini index also from an experimental perspective. For this purpose, we implemented the two measures in a decision-support system for the domain of oesophageal cancer and performed test selection for 162 real patients. Upon analysing the sequences of tests yielded, we found that for 71% of the patients, already the first or second test selected differed between the two measures. In contrast with common knowledge, therefore, the Shannon entropy and the Gini index gave rise to quite different test-selection behaviour. All differences could be explained, however, from the insights that we had gained from our more fundamental analysis of the Shannon entropy, the Gini index, and their first derivatives.

The paper is organised as follows. Section 2 reviews the Shannon entropy, the Gini index, and the misclassification error, and details how these measures are used for test selection. Section 3 summarises our fundamental analysis of the three measures, and of their first derivatives more specifically. In Section 4 we report on the experimental results obtained with the Shannon entropy and the Gini index, and explain the observed differences. The paper ends with our conclusions in Section 5.

2 Information Measures and Test Selection

In a diagnostic decision-support system, test selection generally amounts to selecting tests that are expected to yield the largest decrease in the uncertainty about a patient's diagnosis. For capturing diagnostic uncertainty, typically an information measure is used. The three most commonly used measures are the Shannon entropy, the Gini index, and the misclassification error. These measures are defined for a probability distribution Pr over a set of stochastic variables. We distinguish a diagnostic variable D, modelling the diagnoses of interest; the possible values of D are denoted d_j, $j = 1,\dots,m, m \geq 2$. We further distinguish $n \geq 2$ test variables T_i, modelling diagnostic tests whose results can influence the uncertainty in D; the results of a test T_i are denoted t_i^k, $k = 1,\dots,m_i$, $m_i \geq 2$. Each of the three measures attains its maximum when the uncertainty about the value of the diagnostic variable is the largest, that is, when the probability distribution over this variable is a uniform distribution. For a distribution with $\Pr(d_i) = 1$ for some value d_i of D and $\Pr(d_j) = 0$ for all $d_j \neq d_i$, the uncertainty about the value of the diagnostic variable is resolved and the measures yield their minimum value of 0.

The *Shannon entropy* $H(\Pr(D))$ of the probability distribution Pr over the diagnostic variable D is the expected amount of information that is required to establish the value of D with certainty; more formally, the entropy is defined as

$$H(\Pr(D)) = -\sum_{j=1,\dots,m} \Pr(D = d_j) \cdot {}^2\log \Pr(D = d_j)$$

where $0 \cdot {}^2\log 0$ is taken to be 0. Now suppose that some diagnostic test T_i is performed and that the result t_i^k is yielded. Because of this additional information, the probability distribution over D will change from the prior distribution to the posterior distribution given $T_i = t_i^k$. The entropy of the distribution over D will then change as well, to the entropy of the posterior distribution:

$$H(\Pr(D \mid T_i = t_i^k)) = -\sum_{j=1,\ldots,m} \Pr(D = d_j \mid T_i = t_i^k) \cdot {}^2\log \Pr(D = d_j \mid T_i = t_i^k)$$

Prior to performing the test T_i, however, we do not know for certain which result will be obtained: each possible result t_i^k is yielded with a probability $\Pr(T_i = t_i^k)$. Before actually performing the test, therefore, we expect the entropy of the posterior probability distribution over D to be

$$H(\Pr(D \mid T_i)) = \sum_{k=1,\ldots,m_i} H(\Pr(D \mid T_i = t_i^k)) \cdot \Pr(T_i = t_i^k)$$

We now have that the decrease in uncertainty in the diagnostic variable D by performing the test T_i is expected to be $\widetilde{H}(T_i) = H(\Pr(D)) - H(\Pr(D \mid T_i))$. A test that maximises $\widetilde{H}$ thus is the best test to perform. We assume that upon selecting a test that maximises the expected decrease in uncertainty, ties are broken at random.

The *Gini index* $G(\Pr(D))$ of the probability distribution Pr over the variable D is defined as

$$G(\Pr(D)) = 1 - \sum_{j=1,\ldots,m} \Pr(D = d_j)^2$$

The expected Gini index $G(\Pr(D \mid T_i))$ after performing a test T_i is defined as the expected value of the Gini index where the expectation is taken over all possible results:

$$G(\Pr(D \mid T_i)) = \sum_{k=1,\ldots,m_i} G(\Pr(D \mid T_i = t_i^k)) \cdot \Pr(T_i = t_i^k)$$

The best test to perform again is a test that is expected to result in the largest decrease in diagnostic uncertainty, that is, a test that maximises $\widetilde{G}(T_i) = G(\Pr(D)) - G(\Pr(D \mid T_i))$.

Occasionally also the misclassification error is used for capturing uncertainty [5]; in the sequel we will argue that this measure is less suited for the purpose of test selection, however. The *misclassification error* $M(\Pr(D))$ of the probability distribution Pr over the diagnostic variable D captures the difference between the probability of a certain diagnosis, that is, a probability equal to 1, and the probability of the most likely diagnosis; more formally, it is defined as

$$M(\Pr(D)) = 1 - \max\{\Pr(D = d_j) \mid j = 1, \ldots, m\}$$

The expected misclassification error after performing a diagnostic test T_i is

$$M(\Pr(D \mid T_i)) = \sum_{k=1,\ldots,m_i} M(\Pr(D \mid T_i = t_i^k)) \cdot \Pr(T_i = t_i^k)$$

The decrease in uncertainty in D by performing the test T_i thus is expected to be $\widetilde{M}(T_i) = M(\Pr(D)) - M(\Pr(D \mid T_i))$. A test that maximises $\widetilde{M}$ again is the best test to perform.

3 The Measures from a Fundamental Perspective

To provide for predicting the test-selection behaviour of the Shannon entropy, the Gini index and the misclassification error, we study the three measures from a fundamental perspective. Upon doing so, we focus on a binary diagnostic variable only; our considerations, however, also hold for non-binary variables. For a binary diagnostic variable D, with values d_1 and d_2, the Shannon entropy, the Gini index and the misclassification error can be written as

$$\begin{aligned} H(\Pr(D)) &= -\sum_{j=1,2} \Pr(D=d_j)\cdot {}^2\!\log \Pr(D=d_j) = \\ &= -x\cdot {}^2\!\log x - (1-x)\cdot {}^2\!\log(1-x) \\[1em] G(\Pr(D)) &= 1-\sum_{j=1,2} \Pr(D=d_j)^2 = \\ &= 2x-2x^2 \\[1em] M(\Pr(D)) &= 1-\max\{\Pr(D=d_j)\mid j=1,2\} = \\ &= \begin{cases} x & \text{, if } x\in[0,\frac{1}{2}] \\ 1-x & \text{, if } x\in\langle\frac{1}{2},1] \end{cases} \end{aligned}$$

where $x=\Pr(D=d_1)$; the functions are shown in Figure 1(a), (b) and (c) respectively.

From Figure 1(a) and (b), we observe that the Shannon entropy has a higher value than the Gini index. To formally support this observation, we consider the second derivatives of the two functions:

$$\begin{aligned} H''(x) &= -\frac{1}{x\cdot\ln 2} - \frac{1}{(1-x)\cdot\ln 2} \\ G''(x) &= -4 \end{aligned}$$

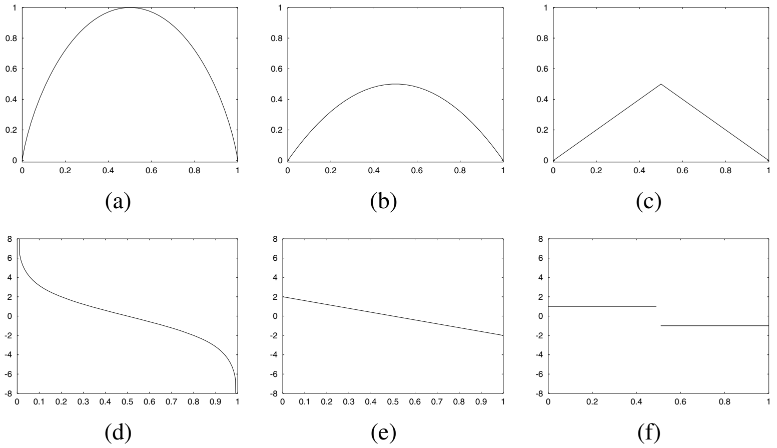
We observe that $H''(x) < G''(x)$ for all $0<x<1$. Since both measures attain their maximum at $x=\frac{1}{2}$, we thus have that in the interval $\langle 0,\frac{1}{2}\rangle$ the ascent of the Shannon entropy is steeper than that of the Gini index; in the interval $\langle\frac{1}{2},1\rangle$, the Shannon entropy shows a steeper descent than the Gini index. We further observe that the two functions attain the same value at $x=0$ and at $x=1$. We conclude that the two functions do not otherwise intersect and, hence, that $H(x)>G(x)$ for all $0<x<1$. We also compare the misclassification error and the Gini index. Within the interval $[0,\frac{1}{2}]$, we have that

$$G(x) = 2x-2x^2 \geq x$$

since from $0\leq x\leq\frac{1}{2}$ we can conclude that $2x^2\leq x$. The Gini index, therefore, lies above the misclassification error. Within the interval $\langle\frac{1}{2},1]$, we find that

$$G(x) = 2x-2x^2 = 2x\cdot(1-x) > 1-x$$

since from $\frac{1}{2}<x\leq 1$ we can conclude that $2x>1$. Again, the Gini index lies above the misclassification error. We thus conclude that $G(x)\geq M(x)$. In fact, the misclassification error can be looked upon as a piece-wise linear interpolation of the three points $G(0), G(\frac{1}{2})$ and $G(1)$ of the Gini index.



(a) (b) (c)

(d) (e) (f)

Fig. 1. The Shannon entropy (a), the Gini index (b), and the misclassification error (c) of a distribution over a binary variable, and their first derivatives (d), (e) and (f)

Now, for test-selection purposes, we are not so much interested in the precise values that the Shannon entropy, the Gini index and the misclassification error attain for a specific probability distribution over the diagnostic variable D. We are more interested in the way they value a *shift* in the distribution that is occasioned by a test result. We therefore also study the first derivatives of the three functions:

$$\begin{aligned} H'(x) &= -{}^2\log x + {}^2\log(1-x) \\ G'(x) &= 2 - 4x \\ M'(x) &= \begin{cases} 1 & \text{, if } x \in [0, \frac{1}{2}\rangle \\ -1 & \text{, if } x \in \langle\frac{1}{2}, 1] \end{cases} \end{aligned}$$

These derivative functions are depicted in the Figures 1(d), (e) and (f) respectively.

The first derivative of the Gini index can be regarded as an approximation of the first derivative of the Shannon entropy for a large range of values of x. To support this observation, we consider the first three terms of the Taylor expansion of $H'(x)$ around $x = \frac{1}{2}$, divided by $G'(x)$. For the quotient, we find that

$$\frac{H'(x)}{G'(x)} = 1.44 + R$$

where the rest term R equals

$$R = 2.85 \cdot (x - \tfrac{1}{2})^2$$

We observe that the rest term is dependent upon the value of x at which we compare the two derivatives. Within the interval $[0.37, 0.63]$, for example, the rest term is smaller than 0.05. Within this interval, therefore, we have that the Taylor approximation of the first derivative of the Shannon entropy differs from the first derivative of the Gini index by a multiplicative factor only, with an error of at most 0.05. This finding is supported by Figure 1(d), from which we observe that the first derivative of the Shannon entropy approximates the linear derivative function of the Gini index in the middle part of the $[0,1]$-interval. From the figure, we further observe that this property no longer holds for the more extreme values. From the rest term R, we find, for example, that for the value $x = 0.3$ the approximation error is less than 0.19 while for the value $x = 0.25$ it has grown to 0.49. For x approaching the extremes, therefore, the quotient $H'(x)/G'(x)$ grows excessively in favour of $H'(x)$.

To compare the first derivatives of the Gini index and of the misclassification error, we begin by observing that G' is a linear function and M' is a piecewise constant function. We further observe that $G'(\frac{1}{4}) = M'(\frac{1}{4})$ and $G'(\frac{3}{4}) = M'(\frac{3}{4})$. We conclude that the first derivative of the misclassification error is a two-point approximation of the first derivative of the Gini index. We now briefly address the suitability of the misclassification error for the purpose of test selection. We observe that within the interval $x \in [0, \frac{1}{2}]$, the misclassification error for the probability distribution over the diagnostic variable D equals $M(\Pr(D)) = \Pr(D = d_1) = x$. Now suppose that for a test variable T_i, we have that $\Pr(D = d_1 \mid T_i = t_i^k) \in [0, \frac{1}{2}]$ for all possible results t_i^k of T_i. We then find that the expected value of the misclassification error after performing the test equals

$$M(\Pr(D \mid T_i)) = \sum_{k=1,\ldots,m_i} \Pr(D = d_1 \mid T_i = t_i^k) \cdot \Pr(T_i = t_i^k) = \Pr(D = d_1) = x$$

The expected misclassification error $M(\Pr(D \mid T_i))$ of the posterior distribution thus equals the misclassification error $M(\Pr(D))$ of the prior distribution, and the expected decrease in uncertainty in D by performing the test T_i is $\widetilde{M}(T_i) = M(\Pr(D)) - M(\Pr(D \mid T_i)) = 0$. Similar observations hold for $\Pr(D = d_1) = x \in \langle \frac{1}{2}, 1]$. Only if the posterior probabilities of a diagnosis given the possible results t_i^k of T_i, are distributed over both intervals can the expected decrease in diagnostic uncertainty $\widetilde{M}(T_i)$ be larger than 0. Now, if at some stage in the test-selection process, for all remaining diagnostic tests the expected decrease in diagnostic uncertainty equals 0, the misclassification error will select a test at random. Since the probability distribution over the diagnostic variable is likely to become less uniform as the test-selection process progresses, the probability that a test will induce a shift to the other interval decreases. The misclassification error will then show a tendency to select tests rather arbitrarily; this tendency has been noted before by Breiman et al. [4]. We note that the tendency of the misclassification error to select tests at random may be quite undesirable for real-life decision-support systems.

We conclude by reviewing the implications of our findings for the test-selection behaviour of the Gini index and the Shannon entropy. The two measures value a test based upon the shifts that its results induce in the probability distribution over the diagnostic variable and upon the probabilities with which these results are expected to be found. Tests that induce a large shift in the probability distribution with a high probability, are valued as more informative than tests that result in a minimal shift with a high probability or in a large shift with just a small probability. Since the first derivative of the Gini index is a decreasing linear function, we find that it values a shift in distribution concavely by a constant factor. Since the first derivative of the Shannon entropy approximates a linear function within the interval $[0.37, 0.63]$, it values a shift in a distribution where x stays within this interval in the same way as the Gini index. We conclude that the Shannon entropy and the Gini index will yield the same diagnostic test upon test selection as long as the tests under consideration are unlikely to result in a rather extreme distribution over the diagnostic variable. Since the Shannon entropy values a shift to an extreme distribution disproportionally more than the Gini index, the two measures may select different tests if a test is likely to result in such an extreme distribution. We note that several other researchers [4,6] also described this difference in behaviour between the Gini index and the Shannon entropy. Glasziou and Hilden for example argue that the Shannon entropy overestimates the gain in information for shifts in an already extreme probability distribution.

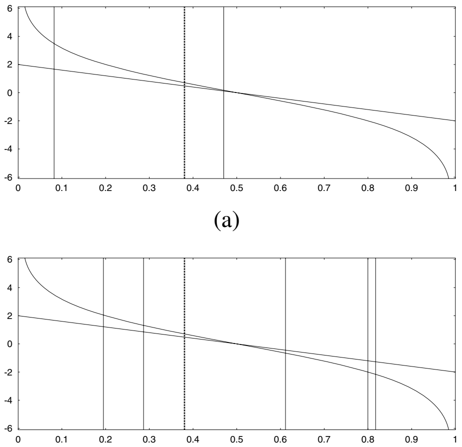
4 The Experimental Results

We formulated, in the previous section, the differences to be expected in the test-selection behaviour of the three measures. Based upon our findings, we concluded that the misclassification error is not as suitable for test selection as the other two measures. In this section we therefore focus on the Shannon entropy and the Gini index. To study the differences between the two measures in a practical setting, we conducted a test-selection experiment using the measures in the context of a real-life decision-support system in oncology. We briefly introduce the system that we used for our experiment before presenting the results that we obtained.

With the help of two experts in gastrointestinal oncology from the Netherlands Cancer Institute, we developed a decision-support system for the staging of cancer of the oesophagus [7]. The kernel of the system is a probabilistic network that models the various presentation characteristics of an oesophageal tumour and the pathological processes involved in its growth. The network currently includes 42 statistical variables, for which almost 1000 probabilities are specified. The main diagnostic variable of the network is the variable *Stage* that summarises the depth of invasion of the primary tumour and the extent of its metastasis; this variable has six possible values. The oesophageal cancer network further includes 25 variables to represent the results of diagnostic tests. For the staging of a patient's oesophageal cancer, typically a number of tests are performed. The various tests differ considerably in their reliability characteristics.

To study the behaviour of the Shannon entropy and the Gini index in the context of the oesophageal cancer network, we extended our decision-support system with a sequential test-selection facility. With the facility, we conducted two experiments. For the first experiment, we extended the oesophageal cancer network with a new binary variable *Operable* that summarises the six possible values of the original diagnostic variable *Stage* by classifying a patient's oesophageal cancer as operable or inoperable. For the second experiment, we used the test-selection facility with the original six-valued diagnostic variable. For our experiments, we had available the medical records of 162 patients diagnosed with cancer of the oesophagus. To simulate a realistic setting, we entered, for each patient, the results of a gastroscopic examination into the network prior to using the facility; in daily practice, the physicians also start selecting tests based upon the initial findings from this standard test.

As an illustration of the results that we found from our first experiment, we discuss the test-selection behaviour of the two measures for a specific patient. When the test



(a)

(b)

Fig. 2. The effects of the results of a CT-scan of the liver (a) and of an endosonography of the oesophageal wall (b), against the first derivatives of the Gini index and the Shannon entropy

selection is started, the probability of the cancer of this patient being operable, equals 0.38; the Gini index of the distribution over the variable *Operable* equals 0.471 and the Shannon entropy equals 0.958. For the next test to perform, the Gini index and the Shannon entropy suggest different tests. The Shannon entropy indicates that a CT-scan of the liver is expected to result in the largest decrease in diagnostic uncertainty, whereas the Gini index selects an endosonography of the oesophageal wall. More specifically, the expected Shannon entropy is computed to be 0.862 for the CT-scan and 0.899 for the endosonography; the expected values of the Gini index are 0.418 and 0.412 respectively.

To explain the observed difference in behaviour between the two measures, we study the shifts in the probability distribution over the diagnostic variable *Operable* that are occasioned by the various test results. Figure 2(a) shows, by means of vertical lines, the shifts in distribution that are yielded by the two possible results of a CT-scan of the liver; the shifts occasioned by the five different values of the endosonography of the oesophageal wall are shown in Figure 2(b). The prior probability of the patient's cancer being operable is indicated by a bold vertical line in both figures. From Figure 2(a), we observe that the leftmost vertical line, indicating the probability 0.082 of the patient's tumour being operable given that the result of the CT-scan of the liver is *yes*, is well within the range in which the first derivative of the Shannon entropy no longer approximates a linear function. The shift in the probability distribution over the variable *Operable* that is occasioned by this test result, therefore, is valued much higher by the Shannon entropy than by the Gini index. The result moreover is relatively likely to be found, with a probability of 0.231. The result *no* of the CT-scan is valued more or less the same by both measures. Two results of the endosonography, on the other hand, are valued more or less concavely by a constant factor by the Shannon entropy as well as by the Gini index. The other three results of the endosonography are not within the range where they are valued more or less the same by the two measures. These three

Table 1. The step, in the test-selection process, at which the Shannon entropy and the Gini index select different tests

Step	*Frequency*	*Step*	*Frequency*	*Step*	*Frequency*	*Step*	*Frequency*	*Step*	*Frequency*
1	24	5	4	9	1	13	0	none	3
2	90	6	6	10	1	14	0		
3	11	7	1	11	1	15	0		
4	19	8	1	12	0	16	3		

results have very low probabilities, of 0.034, 0.097 and 0.005, however. The result that serves to shift the probability of interest to 0.612, on the other hand, has a probability of 0.252, whereas the result that serves to yield a shift to 0.287 has a probability of 0.612. Note that although the probability 0.287 is not within $[0.37, 0.63]$, it is quite close to this interval. The shift to this probability is therefore valued more by the Shannon entropy than by the Gini index, yet not to a large extent. Since the shift occasioned by the endosonography is expected to result in a larger decrease of the uncertainty involved than that occasioned by the CT-scan of the liver, the Gini index selects the endosonography as the best test to perform. The expected decrease in diagnostic uncertainty by the CT-scan, however, is disproportionally larger with the Shannon entropy than with the Gini index, thereby explaining the Shannon entropy selecting the CT-scan. Note that these findings are conform the expectations from our fundamental analysis.

So far, we discussed in detail the differences in test-selection behaviour of the two measures under study for a binary diagnostic variable. We also studied the differences in behaviour for the original six-valued diagnostic variable *Stage*. Table 1 summarises, over all patients, the step in the test-selection process at which the Shannon entropy and the Gini index first selected a different diagnostic test. From the table we observe that for 24 patients (15%), already the first test differed. For 90 patients (56%), the measures selected the same diagnostic test for the first one to be performed, yet chose different tests for the second one. The range of tests selected in the first two steps was quite limited, however. The Shannon entropy selected the endosonography of the local region of the primary tumour for 44% of the patients as the most informative test, the endosonography of the oesophageal wall for 21% of the patients, and the CT-scan of the liver for 28% of the patients. The Gini index selected the endosonography of the local region of the primary tumour and of the oesophageal wall respectively, for 44% and 40% of the patients as the most informative test.

Since the Shannon entropy and the Gini index are commonly taken to be interchangeable for practical purposes, it is remarkable that for just three patients the two measures selected the same tests in the same order. The analysis from the previous section serves to explain why the two measures can select different tests. To explain the large number of differences found, we recall that, before the test-selection process is started for a patient, we entered the results from the gastroscopic examination into the network. Since these results tend not to influence the probability distribution over the diagnostic variable much, the test-selection process was started with a rather similar probability distribution for many patients. The example patient discussed in the previous section in fact belongs to this large group of similar patients.

5 Conclusions

In diagnostic decision-support systems, test selection amounts to selecting tests that are expected to yield the largest decrease in the uncertainty about a patient's diagnosis. For capturing this uncertainty, often an information measure is used. In this paper, we studied the Shannon entropy, the Gini index, and the misclassification error for this purpose. We argued that the first derivative of the Gini index can be regarded as an approximation of the first derivative of the Shannon entropy for a large range of values. We observed that, although a shift in many probability distributions over the diagnostic variable is valued similarly by the Gini index and the Shannon entropy, a shift to rather extreme distributions is valued much higher by the Shannon entropy than by the Gini index. Based upon this observation, the two measures are expected, at least occasionally, to select different tests. We feel that, despite their possible differences in behaviour, both measures are equally suited for use in a decision-support system. We furthermore concluded that the misclassification error should not be used for test-selection purposes due to its tendency to select tests randomly when all possible shifts in the probability distribution over the diagnostic variable are within the same half of the $[0,1]$-interval.

We conducted an experiment to study the behaviour of the Shannon entropy and the Gini index in a real-life setting. The results from our experiment served to corroborate the differences in behaviour expected from our more fundamental analysis. A sensitivity analysis with respect to the selection of tests based upon the two information measures, moreover, showed that test selection based on the Shannon entropy and the Gini index is quite robust. Since our analysis of the two measures is independent of our application domain, we feel that the differences in test-selection behaviour observed in our experiments in the domain of oesophageal cancer, are likely to show in other domains as well.

References

1. Andreassen, S.: Planning of therapy and tests in causal probabilistic networks. Artificial Intelligence in Medicine 4, 227–241 (1992)
2. Ben-Bassat, M.: Myopic policies in sequential classification. IEEE Transactions on Computers 27(2), 170–174 (1978)
3. Breiman, L.: Technical note: some properties of splitting criteria. Machine Learning 24, 41–47 (1996)
4. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadworth & Brooks, Pacific Grove (1984)
5. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer, New York (2001)
6. Hilden, J., Glasziou, P.: Test selection measures. Medical Decision Making 9, 133–141 (1989)
7. Van der Gaag, L.C., Renooij, S., Witteman, C.L.M., Aleman, B.M.P., Taal, B.G.: Probabilities for a probabilistic network: A case-study in oesophageal cancer. Artificial Intelligence in Medicine 25(2), 123–148 (2002)

Nasopharyngeal Carcinoma Data Analysis with a Novel Bayesian Network Skeleton Learning Algorithm

Alex Aussem[1], Sergio Rodrigues de Morais[1], and Marilys Corbex[2]

[1] Université de Lyon,
LIESP, Université de Lyon 1,
F-69622 Villeurbanne France
{aaussem,sergio.rodrigues-de-morais}@univ-lyon1.fr
[2] International Agency for Research on Cancer (IARC)
150 cours Albert Thomas
F-69280 Lyon Cedex 08 France
CORBEXM@emro.who.int

Abstract. In this paper, we discuss efforts to apply a novel Bayesian network (BN) structure learning algorithm to a real world epidemiological problem, namely the Nasopharyngeal Carcinoma (NPC). Our specific aims are : (1) to provide a statistical profile of the recruited population, (2) to help indentify the important environmental risk factors involved in NPC, and (3) to gain insight on the applicability and limitations of BN methods on small epidemiological data sets obtained from questionnaires. We discuss first the novel BN structure learning algorithm called Max-Min Parents and Children Skeleton (MMPC) developped by Tsamardinos et al. in 2005. MMPC was proved by extensive empirical simulations to be an excellent trade-off between time and quality of reconstruction compared to most constraint based algorithms, especially for the smaller sample sizes. Unfortunately, MMPC is unable to deal with datasets containing approximate functional dependencies between variables. In this work, we overcome this problem and apply the new version of MMPC on Nasopharyngeal Carcinoma data in order to shed some light into the statistical profile of the population under study.

Keywords: Bayesian networks, machine learning, epidemiology.

1 Introduction

The last twenty years have brought considerable advances in the field of computer-based medical systems. These advances have resulted in noticeable improvements in medical care, support for medical diagnosis and computer assisted discovery. Decision support systems based on Bayesian Networks (BN) have proven to be valuable tools that help practitioners in facing challenging medical problems, such as diagnosis by identifying the relevant factors (also called features) involved in the disease, illness or disorders under study from experimental data. These probabilistic graphical models offer a coherent and intuitive representation of uncertain domain knowledge. One of the main advantages of BN over other AI schemes

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 326–330, 2007.

for reasoning under uncertainty is that they readily combine expert judgment with knowledge extracted from the data within the probabilistic framework. In this paper, we discuss efforts to apply a new BN learning method to a real world epidemiological problem, namely the Nasopharyngeal Carcinoma (NPC) [1]. The objective is to investigate the role of various environmental factors in the aetiology of NPC in the Maghrebian population.

The graphical part of BN reflects the structure of a problem (ideally a graph of causal dependencies in the modelled domain), while local interactions among neighboring variables are quantified by conditional probability distributions. All independence constraints that hold in the joint distribution represented by any Bayesian network with structure $\mathcal{G}$ can be identified from the structure itself under certain conditions. However, the problem of learning the skeleton from data is worst-case NP-hard [7]. Very recently, a new powerful constraint-based learning algorithm has been proposed by L. Tsamardinos et al. [10] particularly well suited to smaller data sets. The algorithm, known as Max-Min Parents and Children (MMPC), learns the BN skeleton, i.e., the graph of the BN without regard to the direction of the edges. MMPC identifies first the parents and children $\mathbf{PC}_T$ of each target variable T and then pieces together the identified edges into the network skeleton. MMPC employs a smart search strategy for identifying conditional dependencies that exhibits better sample utilization compared to other procedures (e.g. TPDA [3], PC [9]). The algorithm is sound in that it returns the true set provided there is a graph *faithful* to the same distribution and the statistical tests performed are reliable.

Although very encouraging results have been reported with MMPC with smaller datasets, it suffers from one difficulty : the method fails to reconstruct correctly the skeleton when some *approximate functional dependencies* exists among groups of variables. A functional dependency (written $\mathbf{X} \rightarrow Y$) is a constraint between a set of variables, such that, given the value for all $X_j \in \mathbf{X}$, one can functionally (and deterministically) determine the corresponding value of Y. More generally, $\mathbf{X} \rightarrow Y$ is an *approximate* functional dependency (AFD) if it does not hold over a small fraction of the tuples [4]. AFDs are pitfalls to watch out for when MMPC is run on data because it causes the method to miss weakly associated pairs of variables. They are often observed in questionnaire data owing to hidden redundancies in the questions or misunderstanding. As MMPC fails to work properly in the presence of AFDs, the algorithm was modified. Unfortunately, restriction of space has precluded description of our modifications to overcome the above problem. In this paper, we just analyse the graph obtained on the NPC data. In [1], the new MMPC version on small was validated by extensive experiments. Results are not reported here for conciseness.

2 Application to NPC Data

We briefly discuss the application of MMPC2 (the new version of MMPC) to a real-world problem : the Nasopharyngeal carcinoma (NPC) epidemiological data. Epidemiological studies have suggested a large number of environmental

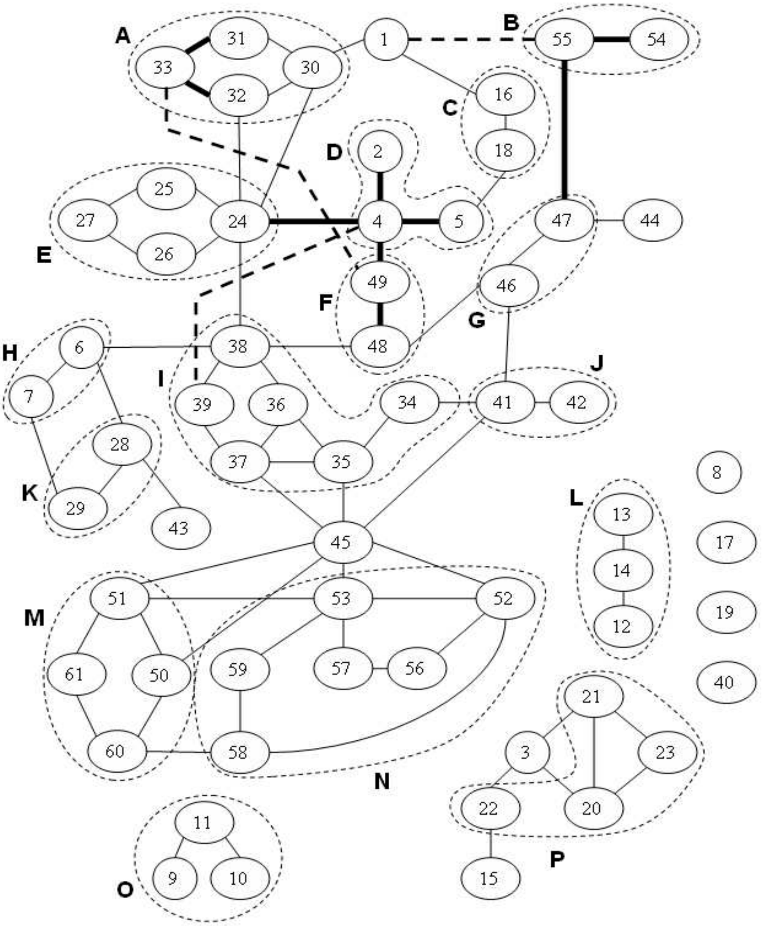


Fig. 1. The skeleton obtained with MMPC2. Bold edges are approximate functional dependencies (AFD) detected by MMPC2, dotted edges are weaker associations. For instance the edge $1 - 55$ would have been hidden by the AFD $47, 44 \rightarrow 55$. In dotted line: the groups of thematic variables. Lexical : **NPC** 1, age of interview for control individuals and age at cancer for cases 2, sex 3, instruction 4, professional category 5, lodging ch. and ad. 6 7, parents consanguinity 8, otitis 9, pharyngitis 10, cold 11, asthma 12, eczema 13, allergy 14, chemical manure and pesticide 15, chemical products 16, smoke 17, dust 18, formaldehyde 19, alcohol 20, tabac 21, neffa 22, cannabis 23, housing type ch. and ad. 24 25, separated beds ch. and ad. 26 27, animal in the house ch. and ad. 28 29, kitchen ventilation ch. and ad. 30 31, house ventilation ch. and ad. 32 33, incense ch. and ad. 34 35, kanoun and tabouna ch. and ad. 36 37, wood fire ch. and ad. 38 39, brest feeding and age of weaning and way of weaning 40 41 42, contact with adult saliva 43, traditional childhood treatments 44, hot pepper 45, smen and fat ch. and ad. 46 47, vegetables and fruits ch. and ad. 48 49, house made harrissa ch and ad. 50 51, industrial harrissa ch. ad. 52 53, house made proteins ch. and ad. 54 55, industrial proteins ch. and ad. 56 57, industrial canned vegetables ch. and ad. 58 59, house made canned vegetables ch. and ad. 60 61. ch.=childhood and ad=adult.

risk factors for NPC, including dietary components as well as household and occupational exposures (see legend of Figure 1). A multi-center case-control study has been undertaken in 2004 by the International Agency for Research on Cancer (IARC) in the Maghreb (Morocco, Algeria and Tunisia), the endemic region of North Africa. The data is made up from 986 individuals older than 35, 61 discrete variables and 5% missing data. The discrete variables have 2 or 3 modalities except age with 4 modalities. We adopt for simplicity the *available case analysis* method to handle missing data although this solution is known to introduce potentially dangerous biases in the estimates (see [8] for a discussion).

MMPC2 on the data yields the skeleton in Figure 1. Bold edges are the approximate functional dependencies detected in the data, dotted edges are the weaker associations that would have been missed by MMPC. As may be seen, The relation between NPC (variable 1) and all other variables is mediated by 30 (bad kitchen ventilation during childhood), 16 (exposure to chemical products and the latter is linked to dust exposure 18 and professional category 5) and 55 (house made proteins at adult age). According to the expert domain, the skeleton confirms that the NPC is associated with: 1) a low socio-economic status with poor housing condition characterized by overcrowding and lack of ventilation; 2) low professional category and chemical product exposure 3) a monotonous diet including the regular consumption of traditionally preserved food (e.g., smen, house made proteins) since very early age. As may be seen, 16 coherent groups of variables are extracted. They are denoted by upper-case letters A to P. A reflects the house and kitchen ventilation; B, house made proteins; C, the exposure to chemical products and dust; D, reflects a strong and interesting dependence between age at cancer, professional category and instruction in these countries; E, is the housing type; F, vegetables and fruits consumption; G are specific traditionally preserved protein and fat; H, lodging condition; I is the exposure to fumes; J, age and way of weaning; K, animals and pets in the house; L, are allergies; M, house made food; N, industrial food; O, are ear, nose and throat infections; P, are the local drugs (tabacs, neffa, cannabis, alcohol) consumed essentially by men. The way groups are related is also informative and the edges lend themselves to interpretation : men are more inclined to smoke and take drugs ; the house type, the overcrowded lodging conditions and the socio-professional conditions are clearly related, exposure to dust/chemical products are related to professional category, smen, vegetables and wood fire are statistically related, domestic animals are present in poor housing conditions, smen (fat) is used as a traditional childhood treatments, the consumption of hot pepper and harrissa is common, poor housing condition is characterized by overcrowding and lack of ventilation etc. More generally, the habits during childhood are reproduced at adult age.

3 Conclusion

In this paper, we discuss the application of a new algorithm called MMPC algorithm developed by Tsamardinos et al. in 2005 to a small real-world nasopharyngeal carcinoma data set. The found skeleton provides the statistical profile of the population.

Acknowledgment

This work is supported by "Ligue contre le Cancer, Comité du Rhône, France". The NPC data was kindly supplied by the International Agency for Research on Cancer, Lyon, France.

References

1. Aussem, A., Rodrigues de Morais, S., Corbex, M.: Analysis of Nasopharyngeal Carcinoma Data with a Novel Bayesian Network Learning Algorithm. In: IEEE Int. Conference on Research, Innovation and Vision for the Future, RIVF'07, March 5-9, 2007, Hanoi, Vietnam, pp. 281–287 (2007)
2. Brown, L.E., Tsamardinos, I., Aliferis, C.F.: A Comparison of Novel and State-of-the-Art Polynomial Bayesian Network Learning Algorithms. In: Proceedings of the Twentieth National Conference on Artificial Intelligence AAAI, pp. 739–745 (2005)
3. Cheng, J., Greiner, R., Kelly, J., Bell, D., Liu, W.: Learning Bayesian Networks from Data: An Information-Theory Based Approach. Artificial Intelligence 137, 43–49 (2002)
4. King, R.S., Legendre, J.J.: Discovery of Functional and Approximate Functional Dependencies in Relational Databases. Journal of Applied Mathematics & Decision Sciences 7(1), 49–59 (2003)
5. Leray, P., Francois, O.: BNT Structure Learning Package: Documentation and Experiments. Reasearch report Laboratoire PSI, INSA Rouen France (2004) http://bnt.insa-rouen.fr/programmes/BNT
6. Murphy, K.: The BayesNet Toolbox for Matlab. Computing Science and Statistics: Proceedings of Interface, vol. 33, pp. 33–40 (2001) www.ai.mit.edu/~murphyk/Software/BNT/bnt.html
7. Neapolitan, R.E.: Learning Bayesian Networks. Prentice-Hall, Englewood Cliffs (2004)
8. Ramoni, M., Sebastiani, P.: Robust Learning with Missing Data. Machine Learning 2 45, 147–170 (2001)
9. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. The MIT Press, Cambridge (2000)
10. Tsamardinos, I., Aliferis, C.F.: The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Machine Learning 1 65, 31–78 (2006)

Enhancing Automated Test Selection in Probabilistic Networks

Danielle Sent[1] and Linda C. van der Gaag[2]

[1] Department of Electrical Engineering, Mathematics and Computer Science,
University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
danielle.sent@utwente.nl

[2] Department of Information and Computing Sciences, Utrecht University,
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
linda@cs.uu.nl

Abstract. Most test-selection algorithms currently in use with probabilistic networks select variables myopically, that is, test variables are selected sequentially, on a one-by-one basis, based upon expected information gain. While myopic test selection is not realistic for many medical applications, non-myopic test selection, in which information gain would be computed for all combinations of variables, would be too demanding. We present three new test-selection algorithms for probabilistic networks, which all employ knowledge-based clusterings of variables; these are a myopic algorithm, a non-myopic algorithm and a semi-myopic algorithm. In a preliminary evaluation study, the semi-myopic algorithm proved to generate a satisfactory test strategy, with little computational burden.

Keywords: diagnostic test selection, probabilistic networks, semi-myopia.

1 Introduction

To support the entire process of a patient's management, a decision-support system should not only provide information about the most probable diseases or the best suitable therapy, it should also provide information about which diagnostic tests had best be performed to reduce the uncertainty about a patient's condition. In the context of probabilistic networks, an automated test-selection facility is usually composed of an information measure, a test-selection loop, and a criterion for deciding when to stop gathering further information. The information measure is defined on the probability distribution over the main diagnostic variable and essentially captures diagnostic uncertainty. With respect to the actual test-selection loop, most algorithms in use with probabilistic networks serve to select diagnostic tests myopically [2]. In each iteration, the most informative variable is selected from among all possible test variables to indicate the next test to perform. The user is prompted for the value of the selected variable, which is entered into the network and propagated to establish the posterior probabilities for all variables. From the set of test variables still available, the next variable is selected. This process of selecting test variables and propagating their results is continued until a stopping criterion is met or until results for all test variables have been entered.

We feel that the test-selection strategy that is induced by a myopic algorithm is an oversimplification of the problem-solving strategies found in many fields of medicine. Based upon interviews with two experts in the field of oesophageal cancer, we in fact

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 331–335, 2007.

identified several aspects where myopic test selection does not match daily routines. In the strategy of our experts, different subgoals are identified that are addressed sequentially, such as discovering the characteristics of the primary tumour and establishing the absence or presence of metastases. We feel that a more involved test-selection facility should take such subgoals into account. Moreover, our experts order tests in packages to reduce the length in time of the diagnostic phase of a patient's management. For the latter purpose, especially, a non-myopic algorithm would be required in which in each step multiple tests can be selected. A fully non-myopic algorithm is computationally very demanding, however, and may easily prove infeasible for practical purposes. Based upon these considerations, we present in this paper three new test-selection algorithms that take a fixed clustering of test variables into account. These algorithms retain some of the idea of non-myopia, yet stay computationally feasible.

The paper is organised as follows. Section 2 reviews the basic test-selection algorithm currently in use with probabilistic networks. Section 3 presents our new algorithms for test selection. In Section 4 we briefly describe the experiments that we conducted with our new algorithms. The paper ends with our conclusions in Section 5.

2 Preliminaries

Before presenting our new algorithms for test selection with probabilistic networks, we briefly review the myopic algorithm in use for this purpose [1,3,4]. This algorithm takes for its input a set $\mathscr{T}$ of test variables. For its output, it sequentially prompts the user to supply a value for a selected variable $T_i \in \mathscr{T}$. The value entered by the user then is propagated through the network at hand before the next variable is selected and presented to the user. The algorithm amounts to the following in pseudo-code:

```
Myopic test selection
input: $\mathscr{T}$ is a list of test variables $T_i$
$Stop = false$
while $\mathscr{T} \neq \varnothing$ and $Stop \neq true$ do
    compute most informative $T_i \in \mathscr{T}$ and remove $T_i$ from $\mathscr{T}$
    prompt for evidence for $T_i$ and propagate
    compute $Stop$
od
```

We assume that the algorithm employs the Gini index of the probability distribution over the disease variable; other information measures can also be used, however. The *Gini index* $G(\Pr(D))$ of the probability distribution Pr over the diagnostic variable D is defined as

$$G(\Pr(D)) = 1 - \sum_{j=1,\ldots,m} \Pr(D = d_j)^2$$

The expected Gini index $G(\Pr(D \mid T_i))$ after obtaining a value for the variable T_i is defined as the expected value of the Gini index where the expectation is taken over all possible values:

$$G(\Pr(D \mid T_i)) = \sum_{k=1,\ldots,m_i} G(\Pr(D \mid T_i = t_i^k)) \cdot \Pr(T_i = t_i^k)$$

The best test variable to select is one that maximises the decrease $G(\Pr(D)) - G(\Pr(D \mid T_i))$ in diagnostic uncertainty. From a computational point of view, the most expensive step in the algorithm is that in which the most informative variable is selected. In this step, the probability distributions $\Pr(D \mid T_i)$ are computed for all test variables T_i. Using Bayes' rule, the number of propagations required equals the number of values of D.

3 Enhanced Test-Selection Algorithms

The test-selection strategy implied by the basic myopic algorithm seems to be an oversimplification of the test-selection routines found in many fields of medicine. In the domain of oesophageal cancer, for example, we found that physicians order tests for specific subgoals. They start gathering general information about the patient and about the tumour. Having this information, they focus on establishing the presence or absence of distant metastases and order tests accordingly. The physicians further order physical tests such as a CT-scan, even though the results of the scan are modelled in the network by multiple variables such as *CT-liver*, *CT-loco*, *CT-lungs*, *CT-organs*, and *CT-truncus*.

To arrive at a test-selection facility that fits in more closely with daily practice, we enhance the basic myopic algorithm to take a list $\mathscr{S}$ of subgoals S_i into consideration. The algorithm performs test selection per subgoal, that is, for each subgoal it focuses on the test variables that provide information about that particular goal. For this purpose, the algorithm is provided with a list of subsets $\mathscr{T}(S_i)$ of $\mathscr{T}$, each of which includes all test variables that pertain to a specific goal. Our first algorithm now computes the most informative test to be performed by investigating single test variables. The user is prompted for just the selected test variable and only the evidence for this variable is propagated throughout the network, before the test-selection process is continued:

```
Algorithm $A_1$: myopic test selection with subgoals
input: $\mathscr{S}$ is a list of subgoals $S_i$,
          $\mathscr{T}$ is a list of test variables $T_j$, organised in sublists $\mathscr{T}(S_i)$ per subgoal $S_i$
Stop-subgoal$(S_i)$, Stop-overall $=$ false
while $\mathscr{S} \neq \varnothing$ and Stop-overall $\neq$ true do
        select next $S_i$ from $\mathscr{S}$ and remove $S_i$ from $\mathscr{S}$
        while $\mathscr{T}(S_i) \neq \varnothing$ and Stop-subgoal$(S_i)$, Stop-overall $\neq$ true do
                compute most informative $T_j \in \mathscr{T}(S_i)$ and remove $T_j$ from $\mathscr{T}$
                prompt for evidence for $T_j$ and propagate
                compute Stop-subgoal$(S_i)$, Stop-overall
        od
od
```

The algorithm selects a subgoal S_i from the list of subgoals. From the associated set of test variables, it selects the variable T_j that is expected to yield the largest decrease in diagnostic uncertainty. The user is prompted to enter evidence for T_j, which is subsequently propagated through the network. The process of selecting test variables continues until the stopping criterion for the subgoal S_i or that for the overall goal has been met, or all tests for S_i have been performed. When the stopping criterion for S_i is satisfied or its set of test variables has been exhausted, the algorithm selects the next subgoal. As soon as the overall stopping criterion is satisfied, the entire process is halted.

Algorithm A_1 is still strictly myopic: test variables are selected sequentially on a one-by-one basis and the next variable is selected only after the user has entered evidence for the previous one. We have argued above that a myopic test-selection strategy may not be realistic for many applications in medicine. A fully non-myopic algorithm, in which the expected Gini index given every possible subset of test variables is established, on the other hand, may be infeasible for practical purposes. Our second algorithm now is non-myopic in nature, yet uses a predefined clustering of the test variables where each cluster is associated with a single physical test. The clustering of the test variables is given as part of the input to the algorithm:

```
Algorithm $A_2$: non-myopic test selection with subgoals
input: $\mathscr{S}$ is a list of subgoals $S_i$,
       $\mathscr{T}$ is a list of clusters $C_j$ of test variables, organised in sublists $\mathscr{T}(S_i)$ per subgoal
$Stop\text{-}subgoal(S_i)$, $Stop\text{-}overall = false$
while $\mathscr{S} \neq \varnothing$ and $Stop\text{-}overall \neq true$ do
    select next $S_i$ from $\mathscr{S}$ and remove $S_i$ from $\mathscr{S}$
    while $\mathscr{T}(S_i) \neq \varnothing$ and $Stop\text{-}subgoal(S_i)$, $Stop\text{-}overall \neq true$ do
        compute most informative cluster $C_j \in \mathscr{T}(S_i)$ and remove $C_j$ from $\mathscr{T}$
        prompt for evidence for all $T_k \in C_j$ and propagate
        compute $Stop\text{-}subgoal(S_i)$, $Stop\text{-}overall$
    od
od
```

Again driven by subgoals, the algorithm selects the cluster C_j of variables that is expected to yield the largest decrease in uncertainty. The user is prompted for evidence for each separate variable T_k from the cluster. We note that algorithm A_2 is much more computationally demanding than algorithm A_1. The increase in computation time stems from computing the most informative cluster. To this end, the probability distributions $\Pr(D \mid C_j = c)$ and $\Pr(C_j = c)$ are computed for all combinations of values c of the test variables in C_j, which requires an exponential number of propagations.

Algorithm A_2 in essence is non-myopic in its test-selection strategy and may become computationally too demanding if a meaningful clustering would result in clusters of relatively large size. To save computation time yet retain some of the idea of non-myopia, we designed an algorithm that implies a semi-myopic test-selection strategy:

```
Algorithm $A_3$: semi-myopic test selection with subgoals
input: $\mathscr{S}$ is a list of subgoals $S_i$,
       $\mathscr{T}$ is a list of clusters $C_j$ of test variables, organised in sublists $\mathscr{T}(S_i)$ per subgoal
$Stop\text{-}subgoal(S_i)$, $Stop\text{-}overall = false$
while $\mathscr{S} \neq \varnothing$ and $Stop\text{-}overall \neq true$ do
    select $S_i$ from $\mathscr{S}$ and remove $S_i$ from $\mathscr{S}$
    while $\mathscr{T}(S_i) \neq \varnothing$ and $Stop\text{-}subgoal(S_i)$, $Stop\text{-}overall \neq true$ do
        compute most informative $T_j \in \mathscr{T}(S_i)$
        prompt for evidence for $T_j$ and for all $T_k \in C_m$ with $C_m$ such that $T_j \in C_m$,
            propagate and remove $C_m$ from $\mathscr{T}$
        compute $Stop\text{-}subgoal(S_i)$, $Stop\text{-}overall$
    od
od
```

Algorithm A_3 very much resembles the myopic algorithm A_1 presented above. Driven by subgoals, it selects the variable T_j that is expected to yield the largest decrease in diagnostic uncertainty. The main difference is, however, that algorithm A_3 prompts not just for evidence for T_j, but also for evidence for all test variables T_k that belong to the same cluster as T_j. Entering evidence for physical tests rather than for just one test variable fits in more closely with the daily routines of the physicians. Physicians think in terms of physical tests even when they are interested mainly in the value of a single variable. After performing the test, therefore, it seems logical to enter not just the result that is currently of interest, but all other results obtained from the same test as well.

4 Preliminary Experimental Results

To compare the performance of the three algorithms for test selection described above, we conducted a preliminary experimental study in the domain of oesophageal cancer. We found that all three algorithms resulted in rather similar sequences of tests, with just occasional differences. To explain this finding, we observe that, if a single test variable is expected to result in a large decrease in diagnostic uncertainty, then it is likely that the test to which it pertains will be quite informative as well. We presented the sequences of test variables constructed by the algorithms to our domain experts. They indicated that they felt most comfortable with the sequences generated by the semi-myopic algorithm. They indicated more specifically that the sequence generated by the myopic algorithm appeared somewhat unnatural.

5 Conclusions

Most test-selection algorithms currently in use with probabilistic networks select variables myopically. We argued that, while myopic test selection is not realistic for many medical applications, non-myopic test selection would be too demanding. We presented new test-selection algorithms which all employ knowledge-based clusterings of variables. Both from the perspective of fitting in with physicians' daily routines and from a computational perspective, we feel that our semi-myopic algorithm provides an appropriate mean by introducing a concept of restricted non-myopia.

References

1. Andreassen, S.: Planning of therapy and tests in causal probabilistic networks. Artificial Intelligence in Medicine 4, 227–241 (1992)
2. Ben-Bassat, M.: Myopic policies in sequential classification. IEEE Transactions on Computers 27(2), 170–174 (1978)
3. Doubilet, P.: A mathematical approach to interpretation and selection of diagnostic tests. Medical Decision Making 3(2), 177–195 (1983)
4. Glasziou, P., Hilden, J.: Test selection measures. Medical Decision Making 9, 133–141 (1989)

ProCarSur: A System for Dynamic Prognostic Reasoning in Cardiac Surgery

Niels Peek[1], Marion Verduijn[1,2], Winston G. Tjon Sjoe-Sjoe[1],
Peter J.M. Rosseel[3], Evert de Jonge[4], and Bas A.J.M. de Mol[2,5]

[1] Dept. of Medical Informatics, Academic Medical Center, Amsterdam
[2] Dept. of Biomedical Engineering, University of Technology, Eindhoven
[3] Dept. of Anesthesia and Intensive Care, Amphia Hospital, Breda
[4] Dept. of Intensive Care Medicine, Academic Medical Center, Amsterdam
[5] Dept. of Cardio-thoracic Surgery, Academic Medical Center, Amsterdam
n.b.peek@amc.uva.nl

Abstract. We present the ProCarSur system for prognostic reasoning in the domain of cardiac surgery. The system has a three-tiered architecture consisting of a Bayesian network, a task layer, and a graphical user interface. In contrast to traditional prognostic tools, that are usually based on logistic regression, ProCarSur implements a dynamic, process-oriented view on prognosis. The system distinguishes between the various phases of peri-surgical care, explicates the scenarios that lead to different clinical outcomes, and can be used to update predictions when new information becomes available. To support users in their interaction with the Bayesian network, a set of predefined prognostic reasoning tasks is implemented in the task layer. The user communicates with the system through an interface that hides the underlying Bayesian network and aggregates the results of probabilistic inferences.

1 Introduction

Systems for outcome prediction are receiving increasing attention in healthcare [1,2]. This is especially true in fields such as critical care and cardiac surgery, where both the risks and the costs are high. An accurate prognostic system can assist in making clinical decisions, in selecting patients for clinical studies, for purposes of planning and allocating resources, and to correct for differences in case-mix when assessing the quality of delivered care [3].

Prognostic systems are usually developed by applying supervised data analysis methods such as multivariate logistic regression analysis or decision tree induction. Subsequently, the resulting model is supplied with a user interface that allows for entering patient data and prognostic queries. Examples are the risk assessment tool for estimating the 10-year risk of developing coronary heart disease [4] based on the Framingham Heart Study [5], the Apache IV System [6] for estimating the death risk of ICU patients [7], and the EuroSCORE Risk Calculator [8] for estimating the death risk when undergoing cardiac surgery [9].

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 336–340, 2007.

The main drawback of these systems is that they allow only for prediction of one event or outcome (typically, death) at a single, predefined point in time (typically, prior to treatment). In clinical practice, however, there is often a need for simultaneously predicting multiple events (e.g. different types of complications), and to adjust predictions as the treatment process progresses. This also means that some variables may change from *predictee* (i.e. that what is predicted) to *predictor* (i.e. that what is used to predict). A typical example is the occurrence of complications during surgery, which may ultimately lead to death.

This paper present a prognostic system, named *ProCarSur*, that is based on the Bayesian network methodology [10,11] and overcomes these limitations by implementing a dynamic, process-oriented view on prognosis. A demonstration of the system will be given during the AIME07 conference.

2 Clinical Context

Cardiac surgery is a complex medical procedure that is applied to patients with severe insufficiency of the cardiac functioning. Most of cardiac surgical interventions involve coronary artery bypass grafting (CABG), repair or replacement of heart valves, aorta surgery, or a combination of these procedures. These procedures are embedded in a health care process that consists of the stages of pre-assessment, operation, and recovery. During these different stages highly specialized clinical personnel are involved, such as cardiologists, cardiac surgeons, anaesthetists, and intensive care unit (ICU) physicians.

During the operation and the postoperative stay at the ICU and nursing ward, several complications may occur that extend the operation time, delay the recovery process, and may lead to permanent disabilities or death. The risks of complications and death play an important role in decision making prior to and during surgery, and these types of outcome are often used to evaluate the quality of the delivered care. Since the mid-1980s, a large number of models have been developed for predicting peri-operative death, with the *Euro*SCORE as predominant model [9]. Most of them have been developed using the statistical method of logistic regression analysis, and aim at preoperative assessment of death risk.

3 System Description

The ProCarSur system was developed by a multidisciplinary team consisting of a medical informatician, a computer scientist, software engineers, and three clinical specialists that are involved in the cardiac surgical process (cardiac surgeon, anaesthetist, and ICU physician). It has a three-tiered architecture, consisting of (i) a Bayesian network and associated inference algorithms, (ii) a task layer that defines four typical prognostic reasoning tasks, and (iii) a user interface through which the user can enter patient data and view the results of prognostic inference. We implemented the Bayesian network in the Netica software (Norsys Software Corp., Vancouver, 2003). The task layer of ProCarSur was written

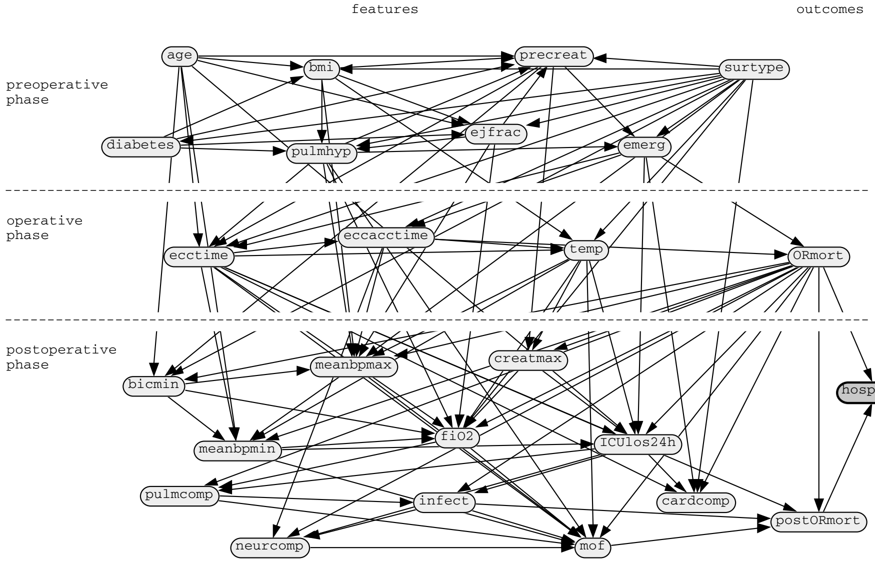
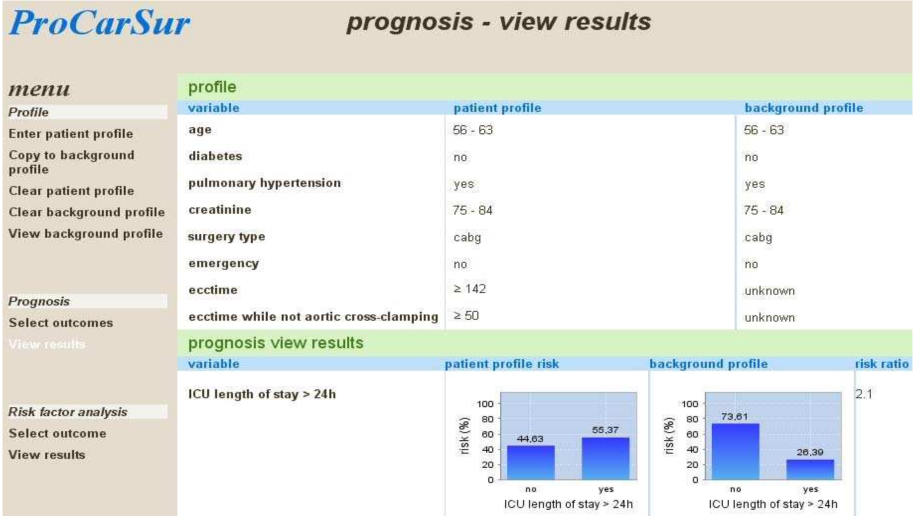


Fig. 1. The prognostic Bayesian network that is part of ProCarSur

in Java, and accesses the Bayesian network through the Netica Java-API. The user interface is made up of HTML-forms. The system runs under Windows XP and requires Netica 1.12, Java Runtime Environment 1.5.0_04, Apache Tomcat 5.5.16, and Internet Explorer 6.0.

Figure 1 displays the prognostic Bayesian network that is part of ProCarSur. The network includes 8 preoperative variables, 3 operative variables, and 12 physiological and complication variables from the postoperative phase. In addition, the network contains an outcome variable that represents death during hospitalization (`hospmort`). The network was induced from data of 10,114 patients who underwent cardiac surgery at the Amphia Hospital (Breda, The Netherlands) using a dedicated network learning procedure that is described elsewhere [12]. Evaluation of the network on an independent test set yielded an area under the ROC curve [13] of 0.767 for preoperative prediction of death, 0.782 for the same type of prediction at ICU admission, and 0.834 after 24h of ICU stay. This resembles the level of performance achieved by most prognostic models in the field of cardiac surgery.

Four reasoning tasks are implemented in ProCarSur's task layer. First, the user can enter patient data and ask for the prognosis on one or more outcome variables, such as death, length of ICU stay, and occurrence of complications. There are no restrictions on the type and amount of patient information that is entered into the system for this purpose: one may, for instance, even choose to make a prediction just by age and type of surgery, in emergency settings. Second, earlier predictions can be updated by the system once new information has become available, and the two prognoses can be compared in terms of relative risks. Third, the system can also compare completely different patient profiles



profile

variable	patient profile	background profile
age	56 - 63	56 - 63
diabetes	no	no
pulmonary hypertension	yes	yes
creatinine	75 - 84	75 - 84
surgery type	cabg	cabg
emergency	no	no
ecctime	≥ 142	unknown
ecctime while not aortic cross-clamping	≥ 50	unknown

prognosis view results

variable	patient profile risk	background profile	risk ratio
ICU length of stay > 24h			2.1

Fig. 2. The output screen of the ProCarSur system

with respect to expected outcomes. For instance, one can compare the risks of neurological complications between coronary artery bypass grafting (CABG) and heart valve operations. The fourth reasoning task is a risk factor analysis and inverses the direction of inference. It identifies important predictors of an unfavorable event or outcome, selected by the user, for a given patient profile.

In Fig. 2, a sample of ProCarsur's output screen is displayed. In this case, a prognosis was made for a 62-year-old non-diabetic patient who has undergone an elective (i.e., non-emergent) CABG operation; this patient had pulmonary hypertension and a preoperative serum creatinine value of 80 μmol/l. These data were available for prognostic assessment in the preoperative stage of the process; the results hereof for the variable ICU length of stay longer than 24h are visible in the right-hand diagram of the lower pane. The operation of this patient took relatively long, resulting in high values for the variables ecctime and ecctime while not aortic cross-clamping, and this information was used to update the prognosis after the operation. The results are shown in the left-hand diagram of the lower pane. The prolonged operation time indicates surgical complications and therefore the risk of an ICU stay longer than 24h has increased from 26.39% to 55.37%, an increase with a factor of 2.1.

4 Discussion and Future Work

In care processes where patients undergo various phases of treatment, prognosis is not a one-shot activity but a set of recurring tasks, often involving a multitude of related outcomes. We have presented a new type of prognostic system that supports all these tasks, using a single underlying model. This is possible through exploiting the declarative nature of knowledge representation in Bayesian networks, and effectively separating domain and task knowledge.

The ProCarSur system currently has a prototype status and is not used in routine medical care. The intended users are physicians and management staff of departments that are involved in cardiac surgery. To be clinically relevant and trustworthy, several adjustments to the Bayesian network model are needed: the current model ignores several important prognostic factors, and is based on data from a single hospital. In the future, we hope to improve this domain model, using a larger, multi-center dataset.

Acknowledgements. Niels Peek receives a grant from the Netherlands Organization of Scientific Research (NWO) under project number 634.000.020. We would like to thank Bas Groot, Maarten Stenvers, and Maurits Visser for their contribution to the system.

References

1. Wyatt, J., Altman, D.G.: Prognostic models: clinically useful or quickly forgotten? Br Med J 311, 1539–1541 (1995)
2. Abu-Hanna, A., Lucas, P.J.F.: Prognostic models in medicine. Methods of Information in Medicine 40, 1–5 (2001)
3. DesHarnais, S.I., Forthman, M.T., Homa-Lowry, J.M., Wooster, L.D.: Risk-adjusted quality outcome measures: indexes for benchmarking rates of mortality, complications, and readmissions. Qual Manag Health Care 5(2), 80–87 (1997)
4. http://hp2010.nhlbihin.net/atpiii/calculator.asp (Last accessed April 17, 2007)
5. Kannel, W.B.: Fifty years of Framingham Study contributions to understanding hypertension. J Hum Hypertension 14, 83–90 (2000)
6. http://www.icumedicus.com/icu_scores/apacheIV.php (Last accessed April 17, 2007)
7. Zimmerman, J.E., Kramer, A.A., McNair, D.S., Malila, F.M.: Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 34(5), 1297–1310 (2006)
8. www.euroscore.org/calc.html (Last accessed April 17, 2007)
9. Nashef, S.A.M., Roques, F., Michel, P., Gauducheau, E., Lemeshow, S., Salomon, R.: European system for cardiac operative risk evaluation (EuroSCORE). European Journal of Cardio-Thoracic Surgery 16, 9–13 (1999)
10. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA (1988)
11. Cowell, R.G., Dawid, A.P., Lauritzen, S.L., Spiegelhalter, D.J.: Probabilistic Networks and Expert Systems. Springer, New York (1999)
12. Verduijn, M., Rosseel, P.J.M., Peek, N., de Jonge, E., de Mol, B.A.J.M.: Prognostic Bayesian networks. I: Rationale, learning procedure, and clinical use. II: An application in the domain of cardiac surgery (Submitted for publication)
13. Metz, C.E.: Basic principles of ROC analysis. Sem Nucl Med 8, 283–298 (1978)

Content Collection for the Labelling of Health-Related Web Content

K. Stamatakis[1], V. Metsis[1], V. Karkaletsis[1], M. Ruzicka[2], V. Svátek[2], E. Amigó[3],
M. Püllä[4], and C. Spyropoulos[1]

[1] National Centre for Scientific Research "Demokritos"
{kstam,vmetsis,vangelis,costass}@iit.demokritos.gr
[2] University of Economics, Prague
{ruzicka, svatek}@vse.cz
[3] Universidad Nacional de Educacion a Distancia
enrique@lsi.uned.es
[4] Teknillinen Korkeakoulu – Helsinki University of Technology
mpolla@cis.hut.fi

Abstract. As the number of health-related web sites in various languages increases, so does the need for control mechanisms that give the users adequate guarantee on whether the web resources they are visiting meet a minimum level of quality standards. Based upon state-of-the-art technology in the areas of semantic web, content analysis and quality labelling, the MedIEQ project, integrates existing technologies and tests them in a novel application: the automation of the labelling process in health-related web content. MedIEQ provides tools that crawl the web to locate unlabelled health web resources, to label them according to pre-defined labelling criteria, as well as to monitor them. This paper focuses on content collection and discusses our experiments in the English language.

Keywords: content labelling, health information quality, web content collection, focused crawling, spidering, content classification, machine learning.

1 Introduction

The number of health information web sites and online services is increasing day by day. Different organizations around the world are currently working on establishing quality labelling criteria for the accreditation of health-related web content [9, 10]. The European Council supported an initiative within eEurope 2002 to develop a core set of “Quality Criteria for Health Related Web Sites” [8]. However, self-adherence to such criteria is nothing more than a claim with little enforceability. It is necessary to establish rating mechanisms which exploit such labelling criteria.

Based upon state-of-the-art technology in the areas of semantic web, content analysis and quality labelling, the EC-funded project MedIEQ[1] aims to pave the way

[1] MedIEQ: Quality Labelling of Medical Web Content using Multilingual Information Extraction. Project site: http://www.medieq.org/

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 341–345, 2007.

towards the automation of quality labelling process in medical web sites by: a) adopting the use of the RDF[2] model for producing machine readable content labels (at the current stage, the RDF-CL[3] model is used); b) creating a vocabulary of criteria, reusing existing ones from various Labelling Agencies; this vocabulary is used in the machine readable RDF labels; c) developing AQUA (Assisting Quality Assessment) [11], a system through which a labelling expert will be able to identify unlabelled resources having health-related content, visit and review the identified resources, generate quality labels for the reviewed resources and monitor the labelled resources.

Our approach necessitates a robust web content collection mechanism with powerful classification skills. A substantial amount of previous work in the area includes various Web crawling or spidering techniques. A Focused Crawler (term introduced by Chakrabarti et al. in 1999 [2]) is a hypertext resource discovery system, which has the goal to selectively seek out pages that are relevant to a pre-defined topic or set of topics. Aiming to enhance crawling, methods that combine link-scoring (ranking of hyperlinks) with reinforcement learning (InfoSpiders [6], "Intelligent crawling" [1]) or others that link the crawler to domain specific linguistic resources, have been proposed. The latter approach was implemented in two, slightly different, ways. First, for the Crossmarc focused crawler [7], a domain specific ontology, linked to several language specific lexicons, provides the crawling start points, defining thus the subset of the web to be crawled. Second implementation: a domain specific ontology [4] or glossary [5] (MARVIN[4] of HON [10]), gives the crawler filtering capabilities: every accessed resource's relevance is estimated and irrelevant resources are excluded.

Section 2 outlines the AQUA system and describes its web content collection methodology, while section 3 discusses our evaluation methodology and experimental results. Section 4 gives our concluding remarks and suggests the future steps.

2 AQUA and Its Web Content Collection Subsystem

As already said, MedIEQ develops AQUA, a system designed to support the work of the labelling expert by providing tools that help the identification of unlabelled web resources, automate a considerable part of the labelling process and facilitate the monitoring of already labelled resources. AQUA is an enterprise-level, web application, which supports internationalization and implements an open architecture.

This paper focuses on the Web Content Collection subsystem of AQUA which involves the following components:

a) the Focused Crawler (identifying health related web sites),
b) the Spider (navigating web sites) with link-scoring and content-classification capabilities (the Spider utilizes a content classification component which consists of a number of classification modules, statistical and heuristic ones),
c) tools assisting the formation of corpora (to train/test classification algorithms),
d) a mechanism producing trained classification/scoring models (to be used by the Spider).

[2] http://www.w3.org/TR/rdf-schema.

[3] RDF-CL will be refined by the W3C POWDER WG (http://www.w3.org/2007/powder/).

[4] http://www.hon.ch/Project/Marvin_specificities.html

3 Evaluation Methodology and Results

A first set of 11 criteria, to examine our methodology and test our tools, was decided, by the Labelling Authorities participating in the project consortium. This set of criteria will soon expand to include additional quality aspects[5]. From the initial 11 criteria, the classification mechanism our Spider exploits has been examined using statistical classification techniques for all criteria depicted in Table 1. In addition, for the last criterion, a method using heuristic detection was examined.

Table 1. The MedIEQ criteria upon which our classification components were evaluated

Criterion	MedIEQ methodology
The target audience of a web site	Classification among three possible target groups: adults, children and professionals
Contact information of the responsible of a web site must be present and clearly stated	Detection of candidate pages during the spidering process and forwarding for information extraction
Presence of virtual consultation services	Detection of parts of a web site that offer such services during the spidering process
Presence of advertisements in a web site	Detection of parts of a web site that contain advertisements during the spidering process

For the statistical classification, pre-annotated corpora were used and three different classifiers provided by the Weka[6] classification platform have been tested: SMO (Weka implementation of SVM), Naïve Bayes and Flexible Bayes. The HTML pages were pre-processed and tokenized in two different methods: a) all HTML tags were removed and only the clear text content of the document was used for the classification and b) both HTML tags and textual content were used. The performance of all classifiers was evaluated using 1-grams and 1/2/3-grams (our results in Tables 2, 3).

Heuristic classification was investigated only for the advertisement detection (our results in Table 4). A large part of current advertising in internet is associated with a reasonably small group of domains; a simple advertisement detection test can be performed by extracting all links on a web page and matching these to a known list of advertisement-providing domain names.

The classification performance for web pages of specific type seems satisfactory, especially if we consider the fact that the corpora were relatively small and the structural information of the pages was not used for the classification. Regarding the performance of the tested classifiers, the obtained values are generally balanced. According to our needs for better precision or recall we can vary the threshold between 0 and 1. All the results presented below use 0.5 as threshold.

The usage of 1/2/3-grams, once the HTML tags removed, gives better results in Target audience classification. A combination of HTML tags and 1/2/3-grams seems to be more helpful in the classification of Contact info and Virtual consultation pages. This latter seems reasonable if we consider that n-grams like “email address”, “phone number”, or sequences of HTML tags which indicate the existence of a communication form

[5] The final set of criteria will be announced through the project website: http://www.medieq.org

[6] http://www.cs.waikato.ac.nz/ml/weka/

Table 2. Target audience (Adults: 102 / Children: 98 / Professionals: 96), F-measure values

		1-grams, no tags	1-grams, with tags	1/2/3-grams, no tags	1/2/3-grams, with tags
Adults	NB	0.77	0.71	0.77	0.83
	FB	0.75	0.76	**0.84**	0.76
	SMO	0.80	**0.84**	0.83	0.79
Childr.	NB	0.90	0.83	**0.93**	0.86
	FB	0.88	0.82	0.91	0.86
	SMO	0.92	0.90	0.91	0.91
Prof.	NB	0.90	0.83	0.91	0.91
	FB	0.88	0.82	**0.95**	0.91
	SMO	0.92	0.90	0.89	0.87

Table 3. CI: Contact info (109 pos. / 98 neg.) – VC: Virtual Consultation (100 pos. / 101 neg.) – AD: Advertisements - *Statistical classification* (100 pos. / 104 neg.), F-measure values

		1-grams, no tags	1-grams, with tags	1/2/3-grams, no tags	1/2/3-grams, with tags
CI	NB	0.72	0.83	0.83	**0.88**
	FB	0.81	0.80	0.83	0.86
	SMO	0.84	0.81	**0.88**	0.87
VC	NB	0.83	0.85	0.85	**0.87**
	FB	0.83	0.83	0.83	0.83
	SMO	0.86	0.84	0.85	0.84
AD	NB	0.88	0.86	0.86	0.84
	FB	**0.89**	0.85	0.88	0.81
	SMO	**0.89**	0.83	0.85	0.82

Table 4. Advertisements - *Heuristic classification*

Precision	Recall	F-measure
0.84	0.72	0.78

may boost the classification performance. On the contrary, we have indications that in Advertisements case such standard sequences of tokens do not occur.

As for the heuristic classification method used for the detection of web pages that contain advertisements, it gives moderate performance results when used independently. The not-so-high precision value is owed to the fact that the known lists of advertisement-providing domain names contain also domain names that provide various "tracking" services instead of advertisements. A potential filtering of those domains would enhance the performance.

4 Conclusions and Future Work

MedIEQ employs existing technologies in a novel application: the automation of the labelling process in health-related web content. Such technologies are semantic web technologies to describe web resources and content analysis technologies to collect domain-specific web content and extract information from it.

Effective spidering using content classification is a vital part of the web content collection process. Our experimental results, investigating the performance of different learning and heuristic methods, clearly indicate that we are in the right direction. They also make appear even more feasible one of the big challenges of the MedIEQ project, that is, provide the infrastructure and the means to organizing and support the daily work of labelling experts by making it computer assisted. Such a system or platform aims to become AQUA.

Nevertheless, there is follow-up work to be done on content collection. In particular, it would be interesting to combine machine learning with heuristics and examine whether classification accuracy is boosted. At the same time, to scale-up AQUA, we should test our methodology in more languages and evaluate our mechanisms in additional quality criteria.

Acknowledgements

The authors would like to particularly thank the people from AQUMED and WMA for collecting the corpora necessary in our experiments, as well as I. Nieminen, V. Rentoumi, I. Kostopoulou, D. Bilidas and A. Skarlatidis for their support at the realization of the experiments.

References

1. Aggarwal, C., Al-Garawi, F., Yu, P.: Intelligent Crawling on the World Wide Web with Arbitrary Predicates. In: Proceedings of the 10th International WWW Conference, Hong Kong, May 2001, pp. 96–105 (2001)
2. Chakrabarti, S., van den Berg, M., Dom, B.: Focused Crawling: a New Approach to Topic-Specific Web Resource Discovery. Computer Networks 31(11-16), 1623–1640 (1999)
3. Curro, V., Buonuomo, P.S., Onesimo, R., de, R.P., A., V., di Tanna, G.L., D'Atri, A: A quality evaluation methodology of health web-pages for non-professionals. Med Inform Internet Med 29(2), 95–107 (2004)
4. Ehrig, M., Maedche, A.: Ontology-focused crawling of Web documents. In: Proc. of the 2003 ACM symposium on Applied computing, pp. 1174–1178 (2003)
5. Gaudinat, A., Ruch, P., Joubert, M., Uziel, P., Strauss, A., Thonnet, M., Baud, R., Spahni, S., Weber, P., Bonal, J., Boyer, C., Fieschi, M., Geissbuhler, A.: Health search engine with e-document analysis for reliable search results. Int J Med Inform. 75(1), 73–85 (2006)
6. Menczer, F., Belew, R.K.: Adaptive Retrieval Agents: Internalizing Local Context and Scaling up to the Web. Machine Learning 39(2/3), 203–242 (2000)
7. Stamatakis, K., Karkaletsis, V., Paliouras, G., Horlock, J., Grover, C., Curran, J., Dingare, S.: Domain Specific Web Site Identification: The CROSSMARC Focused Web Crawler. In: Proc. of the 2nd International Workshop on Web Document Analysis (WDA) (2003)
8. European Commission. eEurope 2002: Quality Criteria for Health related Web sites. europa.eu.int/information_society/eeurope/ehealth/doc/communication_acte_en_fin.pdf
9. WMA, Web Mèdica Acreditada http://wma.comb.es/
10. HON: Health on the Net Foundation http://www.hon.ch
11. Stamatakis, K., Chandrinos, K., Karkaletsis, V., Mayer, M.A., Gonzales, D.V., Labsky, M., Amigo, E., Pöllä, M., AQUA,: a system assisting labelling experts assess health web resources. In: Proc. of the 12th International Symposium for Health Information Management Research (iSHIMR), 18-20 July 2007, Sheffield, UK (to appear, 2007)

Bayesian Network Decomposition for Modeling Breast Cancer Detection

Marina Velikova[1], Nivea de Carvalho Ferreira[2], and Peter Lucas[2]

[1] Department of Radiology, Radboud University Nijmegen Medical Centre
6525 GA, Nijmegen, The Netherlands
m.velikova@rad.umcn.nl

[2] Institute for Computing and Information Sciences, Radboud University Nijmegen
6525 ED Nijmegen, The Netherlands
{nivea,peterl}@cs.ru.nl

Abstract. The automated differentiation between benign and malignant abnormalities is a difficult problem in the breast cancer domain. While previous studies consider a single Bayesian network approach, in this paper we propose a novel perspective based on Bayesian network decomposition. We consider three methods that allow for different (levels of) network topological or structural decomposition. Through examples, we demonstrate some advantages of Bayesian network decomposition for the problem at hand: (i) natural and more intuitive representation of breast abnormalities and their features (ii) compact representation and efficient manipulation of large conditional probability tables, and (iii) a possible improvement in the knowledge acquisition and representation processes.

1 Introduction

The automated differentiation between benign and malignant abnormalities is a difficult problem due to the inherent uncertainty of the detection of breast cancer using mammograms as obtained in breast-cancer screening. Probabilistic approaches based on Bayesian networks (BNs) appear to be useful and promising tools in modeling the problem. Their power lies in the efficient encoding of the causal structure of a domain, which facilitates both the representation and inference of probabilistic knowledge.

In the recent years there have been developed a number of automated systems based on BNs for breast cancer detection [1,2]. In this paper we propose a novel perspective based on BN decomposition. More precisely, we consider three methods that allow for different (levels of) network topological or structural decomposition for tackling classification of breast abnormalities. We exploit the presence of contextual independence between mammographic findings and breast abnormalities as well as the causal independence between various breast cancer risk factors. We also study the presence of class information in this domain in order to obtain compact structure representation and efficient inference in BNs.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 346–350, 2007.

2 Bayesian Networks

Definitions. Consider a finite set $\mathbf{U}$ of random variables, where each variable in $\mathbf{U}$ takes on values from a finite domain $dom(X)$. Let P be a joint probability distribution of $\mathbf{U}$ and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be subsets of $\mathbf{U}$. We say that $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$, denoted by $\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$, if for all $\mathbf{x} \in dom(\mathbf{X})$, $\mathbf{y} \in dom(\mathbf{Y})$, $\mathbf{z} \in dom(\mathbf{Z})$, $P(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) = P(\mathbf{x} \mid \mathbf{z})$, whenever $P(\mathbf{y}, \mathbf{z}) > 0$. In short, $P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})$.

A Bayesian network BN $= (G, P)$ is defined as an acyclic directed graph (ADG) $G = (V, E)$ with set of nodes V, corrresponding to the random variables in $\mathbf{U}$, and set of arcs E, representing the direct causal relationships between the variables. We say that G is an *I–map* of P if any independence represented in G by d-separation, denoted by $A \perp\!\!\!\perp_G B \mid C$, for disjoint sets of nodes $A, B, C \subseteq V$, is satisfied by P, i.e., $A \perp\!\!\!\perp_G B \mid C \Longrightarrow \mathbf{X}_A \perp\!\!\!\perp_P \mathbf{X}_B \mid \mathbf{X}_C$. Here, $\mathbf{X}_W$ represent the random variables from $\mathbf{U}$ that correspond to the set of nodes W. A BN provides a compact representation of independence information about P by specifying a *conditional probability table* (CPT) for each node. The joint probability distribution can be computed by simply multiplying the CPTs.

Contextual and causal independence. Contextual independence refers to conditional independence that holds only in certain contexts, i.e., given the assignment of values to certain variables. Suppose that $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{C}$ are disjoint sets of variables. According to [3], $\mathbf{X}$ and $\mathbf{Y}$ are contextually independent given $\mathbf{Z}$ and context $\mathbf{c} \in dom(\mathbf{C})$, denoted by $\mathbf{X} \perp\!\!\!\perp_P^{\mathbf{c}} \mathbf{Y} \mid \mathbf{Z}$, if $P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}, \mathbf{c}) = P(\mathbf{X} \mid \mathbf{Z}, \mathbf{c})$, whenever $P(\mathbf{Y}, \mathbf{Z}, \mathbf{c}) > 0$, and $\exists \mathbf{c}' \in dom(C)$, $\mathbf{c}' \neq \mathbf{c}$: $P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}, \mathbf{c}') \neq P(\mathbf{X} \mid \mathbf{Z}, \mathbf{c}')$.

Exploiting contextual independence allows one to obtain a finer grained factorization of the joint distribution by further decomposition of conditional probabilities. If we split up a BN into separate BNs, one for each possible context $\mathbf{c}$, then the $\perp\!\!\!\perp_G^{\mathbf{c}}$ is defined by the ADG associated with the individual BN. Such a representation, called a *similarity network*, has been exploited by Geiger and Heckerman in [4]. In [3], Boutilier *et al.* propose another approach for exploiting contextual independence by a compact represention of CPTs as decision trees.

The approach described in [3] is also a way of representing causal independence by qualitative encoding within the network structure. Causal independence arises in cases where multiple causes (parent nodes) lead to a common effect (child node). Intuitively, causal independence allows further decomposition of the conditional probabilities by using functions (e.g., *decomposable* causal independence models [5]).

Language Representation. Having built a BN for a problem domain implies that the entire model is used each time an inference is performed. In addition, as a language BNs are essentially propositional in nature.

The *knowledge-based construction model* attempts to address both limitations [6]. The idea is to use a general knowledge representation language to specify domain facts and relationships. Here, we concentrate on using a probabilistic

Horn clause-like language for representing expert knowledge. A simple example illustrates this language.

$$\begin{aligned}&\mathit{has\_gene}(X,G) \leftarrow \mathit{parent}(Y,X), \mathit{has\_gene}(Y,G).\\&\mathit{parent}(Y,X) \leftarrow \mathit{mother}(Y,X).\\&\mathit{parent}(Y,X) \leftarrow \mathit{father}(Y,X).\end{aligned}$$

3 Structure Representation of the Breast Cancer Domain

One of the problematic areas in the domain of breast cancer detection is the distinction of the main breast abnormalities—mass, architectural distortion, and asymmetry—based on direct and indirect mammographic findings. In the previously developed BN approaches, the authors represent architectural distortion and asymmetry as two-state nodes with values *present* and *absent* [1,2]. However, taking into account various mammographic features and history information, such as previous biopsy, can yield different likelihood for the suspiciousness of these abnormalities.

The specific problem we focus on is the distinction between malignant/benign mass, malignant/benign architectural distortion and focal/benign asymmetry, whose cartesian product of values represents the class variable *Breast cancer* (BC). First, similarly to [2] we consider a number of factors that lead to increase in the risk of breast cancer. These are *Age*, *Number of first-degree relatives with breast cancer*, *Age at menarche*, *Age at first live birth*, *Previous biopsy*. In our study, we add one more risk factor–the presence of BRCA1/2 genes. Next, following the definitions given in the BI-RADS lexicon [7], we consider the following set of mammographic features: *Density*, *Location*, *Margin*, *Orientation*, *Size*, *Shape*. All the causal and feature-abnormality relationships are presented by a single BN in Fig. 1.

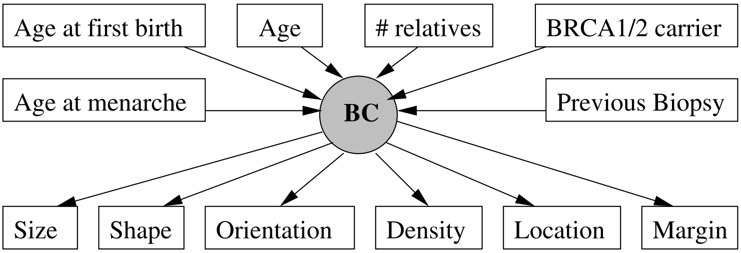


Fig. 1. A single Bayesian network representing the causal and feature-abnormality relationships in the domain of breast cancer detection

The problem with a single BN approach is that it does not naturally represent the causal and contextual independence present in the domain of breast cancer detection. Therefore, here we suggest a finer and better representation of this information by applying a decomposable BN approach.

Given the number of risk factors, the size of the CPT of the node BC increases considerably. To obtain a more compact representation we exploit causal independence of the risk factors. It is known that being a $BRCA1/2$ ($brca$) carrier already establishes a very high probability for developing breast cancer, irrespective of the values of the other factors. If that genetic sign is not present, having a first-degree relative ($nrelat$) determines another probability value. Finally, having neither of those factors, the age and the history of previous biopsy become the factors to pay attention to. Women older than 35 years with known previous biopsy are in a high-risk group compare to women younger than 35. A compact representation of this causal independence is given by:

$$P(bc(X, present) \mid brca(X, yes)) = p_1$$
$$P(bc(X, present) \mid brca(X, no), nrelat(X, 1)) = p_2$$
$$P(bc(X, present) \mid brca(X, no), nrelat(X, 0), age(X) > 35, pbiopsy(X, yes)) = p_3$$

where variable X should be instantiated with the name of a patient.

Having determined the probability of breast cancer given the different causes we then consider contextual independence on the mammographic findings. Malignant masses in contrast to benign masses usually are with irregular or lobular shape, ill-defined or spiculated margins, average size of 1.5 cm, highly-dense, located mostly in the upper-outer quadrant, and have vertical orientation. When a mass is not visible and there is no history about previous biopsy, spiculation can be the only sign of malignant architectural distortion. The central difference between a focal asymmetry (a sign of malignancy) and normal asymmetry is the appearance of the former's density to be concentrated toward its center, i.e., with high density in relatively smaller size. We exploit this contextual information by using the similarity network method mentioned in Section 2. We first build the so-called connected cover of mutually exclusive class variables. The cover contains three triplets (BNi, $i = 1, 2, 3$) corresponding to the distinction between each of abnormalities with its level of suspiciousness and normal breast tissue: BN1{malignant/benign mass, normal breast tissue}; BN2{malignant/benign architectural distortion, normal breast tissue}; BN3{focal/benign asymmetry, normal breast tissue}. Each triplet is a BN with a specific structure given in Fig. 2.

The global network decomposition facilitates the breast cancer detection by grouping the variables that help to discriminate only a specialized group of classes. Furthermore, for well-isolated subdomains, such as microcalcifications and masses, separate similarity networks can be constructed. This leads to a natural and intuitive way of the knowledge representation in the domain.

Finally, we propose an approach for taking into account class information in providing better structure representation of the domain knowledge. In particular, we suggest to extend the ideas in previous related systems such as [1,2] by using different classes in the representation of both MLO and CC views.

Following the way radiologists work, the presence of a highly suspicious mass on MLO view yields a high probability that the same mass is present on the CC view, i.e., based on the class information for one view we can update our belief

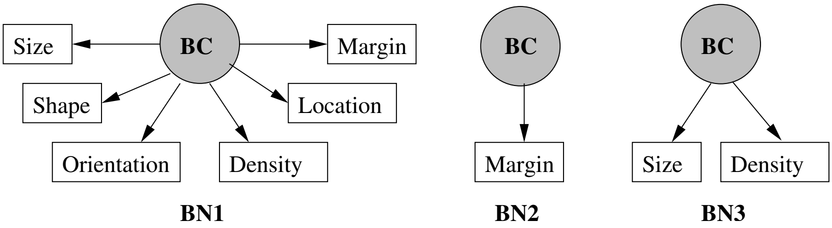


Fig. 2. Local Bayesian networks for the breast cancer detection domain

for the class information on the other view. One possible representation of this information is given by:

$$
\begin{aligned}
&mass(Mass\colon Lesion, View\colon Mammog, Patient\colon Person, Value)\\
&\qquad val(mass) = \{no, benign, malignant\}\\
&margin(Mass\colon Lesion, View\colon Mammog, Patient\colon Person, Value)\\
&\qquad val(margin) = \{NA, circumscribed, ill\!-\!defined, obscured, spiculated\}\\
&P(mass(Y, W, X, malignant) \mid margin(Y, W, X, spiculated)) = p.
\end{aligned}
$$

4 Conclusion

Throughout this paper we examined different techniques for BN decomposition based on the concepts of contextual independence and probabilistic first-order languages for the detection of breast cancer. Our study has a bearing on other medical domains as well, as in many cases the anatomy and functionality of organ systems are looked at from different angles. Such information gives more insight into how variables really influence each other, which is necessary in producing good quality probabilistic models in medicine. Naturally, further development of what we here propose is our main interest at the moment.

References

1. Burnside, E.S., Rubin, D.L., Fine, J.P., Shachter, R.D., Sisney, G.A., Leung, W.K.: Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results. Journal of Radiology (2006)
2. Kahn, C.E., Roberts, L.M., Shaffer, K.A., Haddawy, P.: Construction of a bayesian network for mammographic diagnosis of breast cancer. Computers and Biology and Medicine 27(1), 19–30 (1997)
3. Boutilier, C., Friedman, N., Goldszmidt, M., Koller, D.: Context-specific independence in bayesian networks. In: Proc. of the 12th UAI Conference (1998)
4. Geiger, D., Heckerman, D.: Knowledge representation and inference in similarity networks and bayesian multinets. Artificial Intelligence 82, 45–74 (1996)
5. Heckerman, D., Breese, J.S.: A new look at causal independence. In: Proc. of the 10th UAI Conference, pp. 286–292 (1994)
6. Wellman, M.P., Breese, J.S., Goldman, R.P.: From knowledge bases to decision models. The Knowledge Engineering Review 7(1), 35–53 (1992)
7. BI-RADS: Breast Imaging Reporting and Data System (BI-RADS). American College of Radiology, Reston, VA (1993)

A Methodology for Automated Extraction of the Optimal Pathways from Influence Diagrams

A.B. Meijer

Department of Radiology and Department of Medical Decision Making, Leiden University Medical Center, Albinusdreef 2, Leiden, The Netherlands[*]
a.b.meijer@lumc.nl

Abstract. The influence diagram (ID) is a powerful tool for modelling medical decision-making processes, like the optimal application of diagnostic imaging. In this area, where safety and efficacy are determined by a number of aspects varying in nature and importance, it is difficult for humans to relate all available pieces of evidence and consequences of choices. IDs are well suited to provide evidence-based diagnostic pathways. However, medical specialists cannot be expected to be familiar with IDs and their output can be difficult to interpret. To overcome these shortcomings, a methodology is developed to automatically extract the optimal pathways from an ID and represent these in a tree shaped flow diagram. It imposes a few general rules on the structure of the model, which determine the relation between decisions to perform imaging and the availability of test results. Extracting the optimal pathways requires post-processing the results of an ID, leaving out the sub optimal choices and irrelevant scenarios. Predictive value of tests are vital information in medical protocols, they are at hand in the ID for each relevant scenario. The methodology is illustrated by the problem of diagnosing acute chest pain.

1 Introduction

Clinical guidelines and protocols provide evidence-based recommendations. But with the rapid advances in modern technology, these become quickly outdated and there is a real danger of non-adherence from the medical practitioners. In order to prevent this, two challenges have been formulated [1], also called "living guidelines": (1) Clinical guidelines must become flexible and adaptable. The aim is to develop guidelines, which present up-to-date and state-of-the-art knowledge. (2) Guidelines should be supported by computerized tools aiming at integrating them in the daily work processes of health practitioners. In addition, O argues that most guidelines only take into account a small part of the medical process and that symptoms, diagnostics and treatment all must be considered [2]. Graphical probabilistic models would allow for the integration of these aspects.

Indeed, Bayesian networks and influence diagrams (IDs) have found their place in medical diagnosis. An advantage of IDs for the development of outcome-based

[*] This work was financially supported by the EC-EURATOM 6 Framework Programme (2002-2006) and forms part of the CT Safety & Efficacy (Safety and Efficacy of Computed Tomography (CT): A broad perspective) project, contract FP/002388.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 351–358, 2007.

practice guidelines is that they provide a direct link between the intervention and outcomes [3]. A more recent application of an ID was in the development of a practice guideline on developmental dysplasia of the hip [4].

In this paper, a methodology is described for the automated extraction of the optimal pathways from an ID for symptom-based diagnosis. The aim is to consider all important aspects of safety and efficacy in the medical process and still meet the requirements of a living guideline, as well as representing the optimal pathways in the common medical format of a flow diagram. The methodology should, in principle, be generally applicable to problems of diagnosing symptom-based indications and is illustrated by the problem of diagnosing acute chest pain.

2 Automated Extraction of Optimal Pathways

There have been few attempts for automated extraction of clinical guidelines from decision models. One method uses augmented transition networks to derive a decision tree for therapeutic strategies in the management of chronic diseases [5], but it actually improves already existing guidelines. ALCHEMIST [6] is a web-based system that creates an annotated flowchart algorithm automatically by analyzing the underlying decision model. Although Sanders focuses on decision trees, he proves the feasibility of automated protocol extraction.

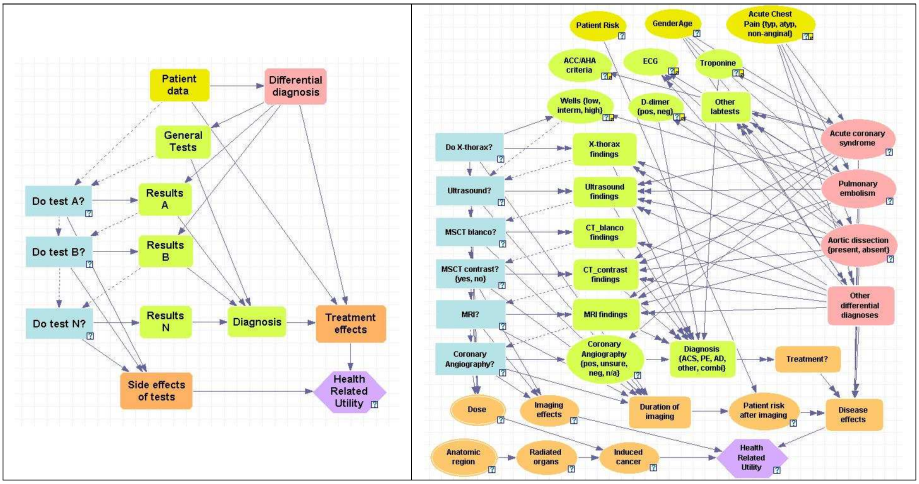


Fig. 1. Left: Relations of causal dependence (solid arcs) and temporal order (dashed) between groups of variables in a general influence diagram for symptom-based diagnosis. **Right**: Influence diagram for acute chest pain.

O [2] has outlined a broad approach to the modelling of symptom-based diagnosis and the distillation of a guideline from an ID. As opposed to most guidelines, she does not start with a strong suspicion for a particular disease and then tries to prove or refute this assumption. Rather, the symptoms form the starting point. The main variables in her model for acute abdominal pain are the entry point (patient data and differential diagnosis), general tests, diagnostic imaging, treatment and outcome (effects of imaging and treatment). Urgency of the symptoms is also taken into account.

In general, an ID outputs the optimal decisions under different scenarios (throughout this article, a scenario means a history of test results and other information). As the tool is also flexible, it is a well suited for automated development of clinical protocols. One drawback of IDs is that the output format is hard to read. It is a table of expected utilities under all possible scenarios and decisions. Not only can this table become enormous, some entries represent impossible scenarios, for example the utility of a test when the decision was to not perform the test. In a medical protocol, only possible and optimal pathways are needed. Moreover, the output does not contain the predictive value of a test in a specific scenario, which is valuable information ideally included in medical protocols. The predictive values can directly be obtained from the ID, as will be described later.

The main reason behind the research in this article is to transform the ID's optimal decision paths to a flow diagram: a format, which is widely accepted by medical specialists. O [2] has written the only paper on exactly this topic, although it is nothing new to unfold an ID into a decision tree, for the purpose of *validating* the results, or as a means to actually *calculate* the utilities. The methodology in this paper consists of two parts, (1) an approach to structured ID development, tailored to medical decision making, and (2) a method for automated extraction of the optimal pathways and representation of these in a (tree-shaped) flow diagram.

2.1 Model Structure

The methodology adopts the modelling approach of O [2] with some changes, mainly on the decision part and the way of computing the diagnosis. Also, temporal orders between tests and decisions are imposed on the model. Although this may seem unnecessarily restrictive, there are reasons to do so. Most important, it is very difficult if not impossible to avoid cycles in the graphical structure of the ID otherwise - maybe asymmetric IDs can provide a solution, but at the very least, the clarity of the model will suffer under the search for generality. Further, a very welcome advantage is a relative reduction in computational complexity from n!, when leaving all possibilities open, to s^n in the proposed structure, where n is the number of decisions and s the average number of states (in the chest pain model s = 2, but in general this depends on the way how decisions are modelled). Anyhow, the proposed order is common in medical practice, going from easy, quick and cheap to invasive and expensive. Any other order will most probably lack evidence base (no clinical trials). Moreover, the order between two tests will only be an issue if the "real" optimal decision path contains both tests and, even when one has proven that, small differences in health-related utility may be expected. Simplicity has its own merits.

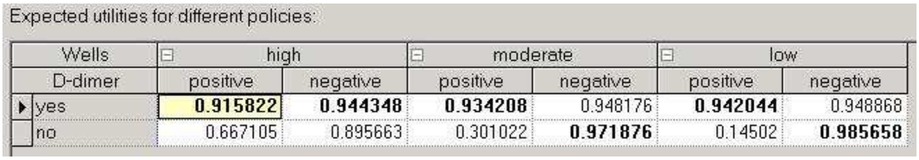
The left-hand side of figure 1 shows the structure of the ID. Patient data and general tests are assumed to be known before the first decision. They determine the prevalence of the diseases in the differential diagnosis. The decisions are whether or not to apply imaging modalities. The health-related utility combines the accuracy of the diagnosis with mortality and morbidity rates and also adds long term effects of radiation and potential negative side effects of imaging. The right-hand side of the figure shows these aspects and their relation in more detail for the problem of diagnosing acute chest pain. For the most important (groups of) nodes, the categorical states are given.

Dashed arcs between decisions are temporal relations, when from chance nodes to decisions they represent the availability of evidence. Solid arcs represent causal relations between (groups of) variables. The variables and causal relations in the left figure are considered the minimum to meet the requirements of a broad perspective. Other relations are allowed, for instance between tests, to model dependence between test results.

2.2 Automated Representation of the Optimal Decision Paths in a Tree-Shaped Flow Diagram

A diagnostic pathway or scenario in this article's approach to symptom-based diagnosis starts with patient characteristics and general tests, then alternately a decision (do test?) and the associated test result(s), ending in an expected utility. Optimal pathways maximize the utility in each scenario. As there is more than one test result possible after each optimal decision, together the pathways can be thought to form a tree. Figure 3 shows the output of the tree representation method for a special case of acute chest pain (only few nodes are expanded, unexpanded nodes are depicted by stacks of nodes).

A method to extract the tree of optimal pathways is given next. After the ID is computed, the space of decision alternatives and results is traversed in a depth-first manner. The incoming arcs to the first decision give the variables, which become the first layer of nodes under the root node "Start", which represents the entry point of the protocol. The optimal decision can always be directly read from the table with expected utilities, it is added (as a rectangular node) under the node, which represents the preceding test. In the next layer, tests results are added (as rounded rectangles) when they have a probability higher than some threshold. The process of alternately and depth-first adding a relevant result and the following optimal decision is repeated till a test is reached, which has no subsequent optimal decision. In this special case, the test result is the diagnosis (skewed rectangle). When an optimal choice is to not perform a test, the method discards it, as it is irrelevant for the protocol and jumps to the next. Pseudo code for automated extraction of a tree-shaped flow diagram is given below:



Expected utilities for different policies:

Wells	⊟ high		⊟ moderate		⊟ low	
D-dimer	positive	negative	positive	negative	positive	negative
▸ yes	**0.915822**	**0.944348**	**0.934208**	0.948176	**0.942044**	0.948868
no	0.667105	0.895663	0.301022	**0.971876**	0.14502	**0.985658**

Fig. 2. Expected utilities (here the fractions of correct diagnosis only) in the decision "Do MSCT contrast?" for mid-aged men with non-anginal chest pain, in six different scenarios of Well's test and D-dimer results

```
Run the influence diagram
Make a root node ("Start")
For all instances of Patient Data and General Tests do
    1. Add test instance, combine more test in one node
       if there's no decision separating them
       (especially under "Start")
    2. Depict disease prevalences of patient group in
       this scenario
    3. Add the optimal test under this scenario, skip
       "no"
    4. For all possible results of this test do
        • Add test result to the optimal test (=node
          of step 3)
        • If new variables become available, then for
          all instances of these variables: repeat
          step 1 to 4
        • If test was end point: add as diagnosis;
          end
        • Else: repeat step 3 and 4
```

Optionally, extra information can be added to certain nodes. It makes sense to add the predictive value of a diagnosis node. For the diagnosis that disease_x in the differential diagnosis is positive, the positive predictive value (PPV) is P(disease_x = present | diagnosis = positive), which can be directly obtained from the ID for any specific pathway. Something similar holds for the negative predictive value (NPV), which is P(all diseases are absent | diagnosis = all_negative). Another option is to depict the prevalence of diseases in tests, anywhere along an optimal pathway; and also similar to obtaining the PPV. Technically, both are a posteriori probabilities of a disease, only the naming convention is different.

3 Diagnosing Acute Chest Pain

The right-hand side of figure 1 shows the most important variables in the influence diagram for the problem of diagnosing acute chest pain. The three most life-threatening possible causes for acute chest pain are acute coronary syndrome (ACS), pulmonary embolism (PE) and aorta dissection (AD). Seven other diagnoses are hidden in the sub model "other differential diagnosis". The diagnosis is a challenge as symptoms can mimic other diseases and mortality is high if the ACS, PE or AD remains undiscovered or gets misdiagnosed. Moreover, the number of imaging modalities is rather high, so without a structured approach it is very difficult to find the diagnostic protocol which optimizes diagnostic efficacy, while minimizing harmful long term effects from radiation and side effects of imaging (e.g. allergy to contrast medium).

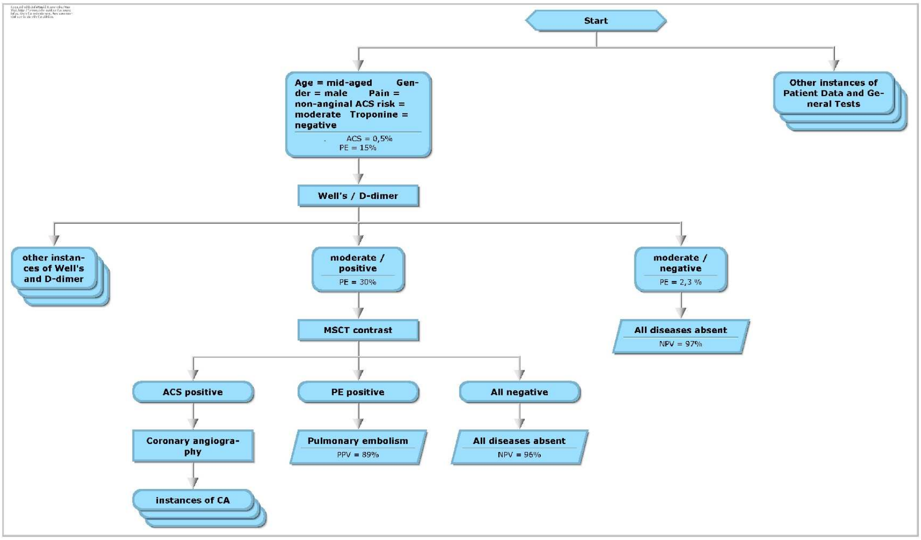


Fig. 3. Part of the diagnostic protocol for the patient group described in the case study

The model is adapted from the recommended structure of IDs for symptom-based diagnosis by O, more details can be found in [2]. The two most important changes are: (1) to each imaging modality a decision to perform it is associated, see discussion in the section "model structure", and (2) the diagnosis is directly based on the test results instead of a posteriori probabilities of a disease. Although the second adaptation might seem to miss the key point of established diagnostic (Bayesian) networks, one has to bear in mind that these networks do not provide a diagnosis as such, i.e. there is no node with states representing black and white conclusions. Rather, they compute the likelihood of all modelled diagnoses, from which one (or another) diagnosis may somehow be deduced. This very projection of posterior probabilities onto a conclusive diagnosis is not part of the Bayesian paradigm, which is one reason why in this article another approach is proposed; another reason is the problem of validating any such projecting function. In the chest pain model, the diagnosis is defined positive if any of the imaging is positive, with the only exception that invasive coronary angiography overrules CT on ACS to a negative diagnosis as well (note that this kind of exception *is* easily modelled within the Bayesian paradigm). Whether this diagnostic function is either good enough, needs further refinement towards medical practice or requires automated optimisation is another decision problem. Leaving more extensive philosophical discussion for future work, this article stays close to medical practice as in existing evidence-based diagnostic protocols. The end result of these is typically a conclusive diagnosis, ideally with positive or negative predictive value. The a posteriori's are used in the node "disease effects" to account for diagnostic accuracy, interacting with mortality and morbidity rates.

The methodology in this paper is illustrated by the case of a mid-aged man, presenting with non-anginal chest pain, having moderate risk for ACS according to the ACC/AHA criteria and negative cardiac troponines. The open scenarios here are defined by the results of Well's test and D-dimer. Other general tests are assumed unknown (modelled by removing the relevant dashed arcs from the ID). Also, decisions for Ultrasound, MSCT blanco and MRI are assumed "no", X-thorax is assumed "yes" and all diagnoses other than ACS and PE are assumed "absent" (in the model they are set manually). Because the model is under construction and some essential data for computing the health-related utility are missing at this stage, the utility function is reduced to the fraction of correct diagnoses. The associated assumptions are that imaging effects and induced cancer are absent and that treatment is always 100% successful and without complications, while withholding treatment to patients with ACS or PE is 100% fatal. Although the last assumptions are medical nonsense, obviously, it does not lead to any loss of generality regarding the proposed methodology, to which the precise nature of the utility is irrelevant.

Figure 2 gives the expected utilities of the decision "MSCT contrast?" under the six open scenarios. In four of the scenarios it is optimal to perform MSCT with contrast. Only in the unlikely case of finding ACS (±3,5%, due to 97% CT specificity and 0,5% prevalence), coronary angiography is the following optimal decision. Figure 3 gives the corresponding diagnostic pathways. This protocol is constructed by manually performing the steps of the method for automated extraction of optimal diagnostic pathways. As far as PE is concerned, it is comparable to the latest PIOPED protocols [7].

4 Discussion

A methodology is described for the automated extraction from influence diagrams of diagnostic pathways, which are optimized for safety and efficacy. Important variables in the model are differential diagnosis, patient data, imaging modalities, test results and diagnosis, treatment effects and side effects of imaging. The temporal and causal relations between these variables are specified. A method for the automated extraction of the optimal diagnostic protocol is described and pseudo code is given. The feasibility of the methodology is shown by a case study in acute chest pain.

Acknowledgements

The author would like to thank Lucia Kroft, Hildo Lamb and prof. de Roos for their assistance in the development of the model for acute chest pain. Special thanks goes to Ying-Lie O for continuous help and ideas on modelling issues and to project leader Koos Geleijns who is a busy man but always has time for questions and a chat. All are residing at the Leiden University Medical Center.

References

[1] ten Teije, A., Marcos, M., Balser, M., van Croonenborg, J., Duelli, C., van Harmelen, F., Lucas, P., Miksch, S., Reif, W., Rosenbrand, K., Seyfang, A.: Improving medical protocols by formal methods. Artificial Intelligence in Medicine 36(3), 193–209 (2006)

[2] Ying-Lie O.: Model-based guideline development for symptom-based indications. In: ten Teije, A., et al. (eds.) ECAI 2006, Workshop Clinical guidelines

[3] Owens, D.K., Nease Jr, R.F.: Development of outcome-based practice guidelines: a method for structuring problems and synthesizing evidence. Jt Comm J Qual Improv. 19(7), 248–263 (1993)

[4] Lehmann, H.P., Hinton, R., Morello, P., Santoli, J.: Developmental dysplasia of the hip practice guideline: technical report committee on quality improvement, and subcommittee on developmental dysplasia of the hip. Pediatrics 105(4), 57–71 (2000)

[5] Seroussi, B., Bouaud, J., Vieillot, J.: Automatic Derivation of a Decision Tree to Represent Guideline-Based Therapeutic Strategies for the Management of Chronic Diseases. In: Miksch, S. et al. (eds.) AIME 2005, LNAI, vol. 3581, pp. 131–135 (2005)

[6] Sanders, G.D., Nease Jr, R.F., Owens, D.K.: Design and pilot evaluation of a system to develop computer-based site-specific practice guidelines from decision models. Medical Decision Making 20(2), 145–159 (2000)

[7] Stein, P.D., Woodard, P.K., et al.: Diagnostic Pathways in Acute Pulmonary Embolism: Recommendations of the PIOPED II Investigators. Am. J. Med 119, 1048–1055 (2006)

Computer-Aided Assessment of Drug-Induced Lung Disease Plausibility

Brigitte Séroussi[1], Jacques Bouaud[2], Hugette Lioté[1,3], and Charles Mayaud[3]

[1] Université Paris 6, UFR de Médecine, Paris, France; AP-HP, Hôpital Tenon, Département de Santé Publique, Paris, France
[2] AP-HP, DSI, STIM, Paris, France; INSERM, UMR_S 872, eq. 20, Paris, France
[3] Université Paris 6, UFR de Médecine, Paris, France; AP-HP, Hôpital Tenon, Service de Pneumologie, Paris, France
brigitte.seroussi@tnn.aphp.fr

Abstract. Drug-induced lung disease (DILD), often suspected in pneumology, is still a diagnostic challenge because of the ever increasing number of pneumotoxic drugs and the large diversity of observed clinical patterns. As a result, DILD can only be evoked as a plausible diagnosis after the exclusion of all other possible causes. PneumoDoc is a computer-based decision support that formalises the evaluation process of the drug-imputability of a lung disease. The knowledge base has been structured as a two-level decision tree. Patient-specific chronological and semiological criteria are first examined leading to the assessment of a qualitative intrinsic DILD plausibility score. Then literature-based data including the frequency of DILD with a given drug and the frequency of the observed clinical situation among the clinical patterns reported with the same drug are evaluated to compute a qualitative extrinsic DILD plausibility score. Based on a simple multimodal qualitative model, extrinsic and intrinsic scores are combined to yield an overall DILD plausibility score.

1 Introduction

Awareness of drug-induced lung disease (DILD) is increasing: a review published in 1972 identified only 19 drugs as having the potential to cause pulmonary disease. Now at least 400 agents are identified when querying the Pneumotox database [1] and the list continues to grow. Early diagnosis is important, because stopping the drug usually reverses toxicity, whereas unrecognized toxicity can be progressive and even fatal.

However, recognition of DILD remains a diagnostic challenge because there is no gold standard test, and clinical, radiologic, and histologic findings are non-specific. Thus, the diagnosis of drug-induced injury is currently only assessed as a plausible hypothesis. As a consequence, there is no reliable data on which DILD diagnostic probabilities could be estimated, forbidding numerical approaches to model the process. In this work, we propose a non-numerical empirical model of uncertainty to assess the plausibility of DILD as a qualitative drug-imputability score. This model accounts for the heuristic principles that are used in clinical routine to diagnose DILD [2]: qualitative plausibility scores are locally assessed from patient data, or *intrinsic* factors, and drug knowledge, or *extrinsic* factors. Qualitative scores are then combined through an intermediate quantitative step to yield the overall DILD plausibility.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 359–363, 2007.

2 Reasoning Under Uncertainty

Plausibility is the state of being plausible, *i.e.* appearing worthy of belief, and is related to uncertainty. There has been considerable work to model uncertainty and belief, and how to combine uncertain information, or facts, to draw plausible inferences [3]. Uncertainty representations in AI fall into two basic categories: numerical (such as Bayes's, Dempster-Schafer's, and fuzzy theories) and non-numerical, or symbolic, approaches. Symbolic representations are mostly designed to handle the aspect of uncertainty derived from the incompleteness of information. Among these models, endorsement theory [4] relies on a heuristic approach of uncertainty. It was initially proposed as an alternative to the probabilistic handling of uncertainty: subjective degrees of belief are hardly quantified and generally do not behave as probabilities. The major advantage of this theory is that it makes *explicit* sources of uncertainty and the way they are combined so we may reason about them directly, instead of implicitly through some sort of numerical calculus. This qualitative reasoning about uncertainty is suited to model human expert-based knowledge, which can be context-dependent, when support data is missing. Some authors compared numerical theories and endorsement theory for a problem that combined data and heuristic knowledge [5].

3 Knowledge Model

The knowledge model (KM) is represented by a two-level tree-structured algorithm. The first level explores patient data to assess the intrinsic DILD plausibility. The second level explores bibliographical drug data to evaluate the potentiality of the suspected drug to induce pulmonary toxicity and assess the extrinsic DILD plausibility. This second phase has thus to be operated for each suspected drug.

3.1 Intrinsic Factors

Intrinsic factors include chronological and semiological criteria. The assessment of chronological criteria involved checking the suspected drug intake is before pulmonary manifestations occurred (otherwise the chronology is *incompatible*), searching for previous pulmonary episode with the same drug or hypersensitivity reaction following the suspected drug intake. The assessment of semiological criteria is only developed when the chronology is not *incompatible*. The first step consists in the exclusion of non drug-induced diagnoses for which pathognomonic characterizations exist. The 5 differential diagnoses modeled (pulmonary œdema, pulmonary embolism, infective pneumonia, and malignancy or systemic disease) are thus incrementally explored. As soon as one is proven, the DILD plausibility associated to semiology is *incompatible*. When not *incompatible*, the second step based on the evaluation of clinical, radiologic, and histologic (BAL) findings is performed. Opacities observed on X-rays are first characterized as localized or disseminated. When disseminated, the delay between the first manifestation of clinical and radiologic symptoms and the hospitalization date is assessed making the difference between overacute, acute, and chronic pneumonia. BAL findings as well as other clinical parameters (fever, acute respiratory distress syndrome, etc.) are integrated and the DILD plausibility score from semiological criteria is attached to the different patterns.

3.2 Extrinsic Factors

As opposed to intrinsic factors assessed from patient-specific data, extrinsic factors represent drug-based information. This information is evaluated using the Pneumotox database (*www.pneumotox.com*). The website offers a comprehensive catalog of drugs known to be responsible of DILD and gives for each drug a rough estimate of adverse effects frequency (as reported in the literature) scored with stars: '*', isolated case reports (1 to 5) which await confirmation; '**', about 10 available cases; '***', in the range of 20 to 100 cases; '****', more than 100 reported cases. No star means suspicious drug but no data published yet. For any suspected drug, absolute frequencies of pulmonary toxicity given by Pneumotox have been weighted to integrate frequencies of the observed clinical situation among the reported clinical patterns. The assessment of extrinsic DILD plausibility is summarized in figure 1.

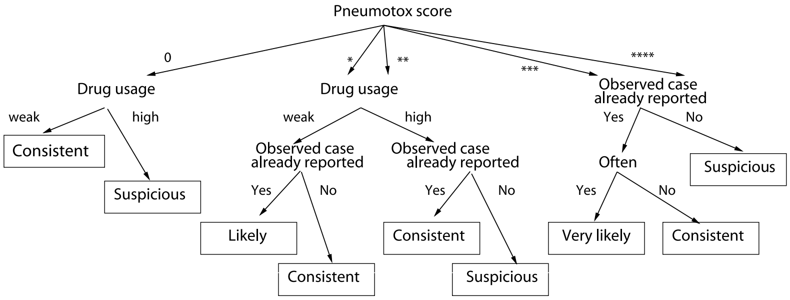


Fig. 1. Assessment of extrinsic DILD plausibility score

4 Plausibility Model

The plausibility representation of a statement related to drug imputability is based on an ordered set of qualitative values: *incompatible*, *suspicious*, *consistent*, *likely*, *very likely*. More specifically, *certain* is not considered in our application domain, since DILD certainty can never be established in the diagnostic process.

To manage the combination of plausible statements in a probability-like manner using the cross product, numerical values and intervals have been assigned to the qualitative plausibility values. In this context, *incompatible* means that the DILD hypothesis should be rejected whatever the plausibility of other statements: *incompatible* has thus to be the null element. *Consistent* has to be the neutral element: there is no specific argument to doubt or believe in DILD; when a DILD *consistent* statement is combined

Table 1. Numerical spaces assigned to symbolic plausibility values

Incompatible	*Suspicious*	*Consistent*	*Likely*	*Very likely*
0	$1/2$	1	2	> 2

with another statement, the resulting plausibility is that of this other statement. Table 1 reports such an assignment.

Let P_{chrono} denotes the chronology-related DILD plausibility, P_{semio} the semiology-related DILD plausibility, and $P_{intrinsic}$ the intrinsic DILD plausibility. Since $intrinsic = chrono \wedge semio$, and as chronology parameters and semiology data are considered independent, then $P_{intrinsic} = P_{chrono \wedge semio} = P_{chrono} \times P_{semio}$. As for extrinsic DILD plausibility denoted $P_{extrinsic}$, it is derived by considering literature citations and drug use as modeled in the KM (see figure 1). Thus, the overall DILD plausibility score $P_{overall}$ is obtained by the combination of intrinsic and extrinsic DILD plausibility scores: $P_{overall} = P_{intrinsic} \times P_{extrinsic}$.

However, whatever the strict value of the overall DILD plausibility score, the qualitative nature of the combination allows us to distinguish 5 different situations (Table 2) which lead to different interpretations and actions. There are 3 situations where intrinsic and extrinsic plausibilities are in the same range: (i) DILD is *suspicious* and the drug toxicity hypothesis will be rejected, (ii) DILD is *likely* or *very likely* and the DILD hypothesis will highly be considered with the immediate stop of the treatment, (iii) DILD is *consistent*, which is the most difficult case, there is neither evidence to support the hypothesis nor to reject it. In this case, any other administered drug that would better explain the clinical situation should be considered. By default, the suspected drug should be stopped to assess actual DILD.

In the 2 other situations, there is a mismatch between what is observed for the patient and what is currently known about the drug, this corresponds to "dissociation patterns". First, when extrinsic plausibility is *likely*, or *very likely*, while intrinsic plausibility is *suspicious*, extrinsic data should take precedence: although patient data should be carefully re-assessed, DILD should be considered and the treatment should be stopped. Second, when extrinsic plausibility is *suspicious* and intrinsic plausibility is *likely*, or *very likely*, if there is no other suspected drug, the actual clinical case could be a candidate for a new clinical pattern. The administered drug should be stopped for DILD assessment. In both situations, if DILD is confirmed, the clinical case is a new occurrence and should be ideally "published" to actualize the Pneumotox database.

Table 2. Overall DILD plausibility from the combination of intrinsic and extrinsic plausibilities

Intrinsic	*Extrinsic*			
	Suspicious	*Consistent*	*Likely*	*Very likely*
Suspicious	*Suspicious*	*Suspicious*	**Dissociation**	
Consistent	*Suspicious*	*Consistent*	*Likely*	*Very likely*
Likely	**Dissociation**	*Likely*	*Very likely*	*Very likely*
Very likely		*Very likely*	*Very likely*	*Very likely*

5 Conclusion

PneumoDoc is a computer-based decision support system designed to help physicians assessing the overall plausibility of DILD. Developed in the documentary paradigm of decision making initially proposed with OncoDoc [6], the system relies on a knowledge base (KB) which is interactively browsed by the user physician. Structured as a two-level decision tree, the KB implements the DILD diagnosis strategy proposed by the KM described. Based on the heuristic principles used in clinical routine, the system makes the most of patient data (intrinsic factors) and literature-based drug information (extrinsic factors). The proposed, specific, plausibility model is based on qualitative ordered values which are combined according to heuristic, domain-dependent, knowledge. Akin in spirit to the endorsement framework, the advantage is the explicit specification of how uncertainty is propagated, which allows for the contextual interpretation of the system's result in the case of an actual patient.

As DILD diagnosis relies on the successive exclusion of all other possible causes, beyond structuring the reasoning process, the sequence and the choice of the investigations proposed by PneumoDoc to eliminate the 5 main differential diagnoses stand for the appropriate etiologic search strategy. In addition, PneumoDoc may help the detection of new toxicity cases (new pneumotoxic drug or new clinical pattern of a known pneumotoxic drug) with the identification of dissociation patterns.

The system has been tested on 20 actual medical records of DILD and lead to 100% of *very likely* DILD plausibility score. A retrospective evaluation is currently under process on 50 randomly selected pneumological records (including known DILD and non DILD). A multicentric survey is planned to be carried out to measure the impact of PneumoDoc on medical practices evaluated in terms of medico-economical parameters (length of hospitalization, number and type of laboratory tests used in the etiologic search, etc.).

References

1. Camus, P., Fanton, A., Bonniaud, P., Camus, C., Foucher, P.: Interstitial lung disease induced by drugs and radiation. Respiration 71(4), 301–326 (2004)
2. Mayaud, C., Fartoukh, M., Parrot, A., Cadranel, J., Milleron, B., Akoun, G.: Les pneumopathies infiltrantes diffuses d'origine médicamenteuse : un probléme avant tout diagnostique. Rev Pneumol Clin 61(3), 179–185 (2005)
3. Bonissone, P.: Reasoning, plausible. In: Shapiro, S. (ed.) Enc Artif Intell, pp. 854–863. John Wiley and Sons, Inc., Chichester (1987)
4. Cohen, P.R.: Heuristic reasoning about uncertainty: an artificial intelligence approach. Pitman, Boston (1985)
5. Comber, A.J., Law, A.N.R., Lishman, J.R.: A comparison of Bayes', Dempster-Shafer and endorsement theories for managing knowledge uncertainty in the context of land cover monitoring. Computers, Environment and Urban Systems 28, 311–327 (2004)
6. Séroussi, B., Bouaud, J.: Using oncodoc as a computer-based eligibility screening system to improve accrual onto breast cancer clinical trials. Artif Intell Med 29(1–2), 153–167 (2003)

Part VII

Applications of AI-Based Image Processing Techniques

Segmentation Techniques for Automatic Region Extraction: An Application to Aphasia Rehabilitation

M.G. Albanesi[1], S. Panzarasa[2], B.Cattani[3], S. Dezza[1], M. Maggi[1], and S. Quaglini[1]

[1] Department of Computer Science and Systems, University of Pavia
[2] CBIM, Pavia
[3] IRCCS Foundation "S. Maugeri", Pavia, Italy

Abstract. We describe a system that facilitates speech therapists to administer cognitive rehabilitation exercises and to evaluate treatment outcomes. We started by augmenting a commercial tool with a more user-friendly interface, meeting the needs of the healthcare professionals involved. Then we integrated, into the same tool, a new type of exercise, that is particularly patient-tailored, being based on the recognition of familiar images within a picture (such as a relative, a domestic animal, a home object, etc). Segmentation techniques are used to elaborate an input picture and individuate areas including *interesting* objects, that will be semi-automatically linked to text and sound. The picture and associated information are then stored in the system database and may be subsequently used as objects for the new exercise. Any number of images may be elaborated, personalised and stored for each patient. The performance has been tested on voluntary subjects with good results.

1 Introduction

Aphasia is an acquired language deficit, often caused by cerebral lesion, such as ischemic or hemorrhagic stroke. Other causes may be trauma, cerebral tumours and encephalitis. Majority of lesions is on the left cerebral side, where the language area is located. Aphasia causes a communication disability. Patients may be impaired in speaking, comprehension, may have reading/writing deficits, and also disturbances in the non-verbal communication. Two or more deficits may be contemporary or not, and show different severity degrees. Some patients are totally unable to communicate while having intellectual functions fully preserved, leading to frustration and depression, heavily affecting the quality of life.

Comprehension problems arise at different levels: word recognition (phonemic level), word sensing (semantic level), syntactic structure decoding (morpho-syntactic level). Linguistic expression anomalies may consist in producing very brief sentences or isolated words, in using turns of phrase, distorting sounds, replacing correct words with words belonging to the same phonetic or semantic class, or even producing neologisms. The rehabilitation objective is the recovery of the communication capability, finalised to the reintegration of the individual in his social context, as stated by the SPREAD

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 367–377, 2007.

guidelines for stroke rehabilitation [1], where we read that "*Speech and language therapy is aimed at:*

- *recovering general and verbal communication, reading, writing and calculation;*
- *enhancing compensatory strategies for communication functions;*
- *instructing carers on methods for maximising communication.*

Most common treatments for aphasia are:

- *impairment-based approaches;*
- *recovery of communication functions according to neurocognitive models of language;*
- *stimulus-response approaches."*

Our system deals with the last item, thus from here on we refer to stimulus-response exercises. Traditionally, treatment is based on face to face encounters of patients with speech therapists. Stimuli are shown to patients through images and objects. Simple exercises consist in words/objects recognition, word/sentences comprehension, distinguishing among similar words/sounds, etc. One common problem is to find adequate material for adult and elderly people. In fact, while there is several commercial "stimuli kits" for the infancy language disturbances (see Figure 1), this is not the case for adult aphasia. Often, the same images and objects are used for children and adults, resulting in poor emotional involvement and frustration of the latter. Computer-based tools cannot fully replace encounters, but they may help therapists, improving personalisation and scalability, and may be useful to continue the therapy at home, since it has been shown that intensive therapy correlates with better outcomes [2].

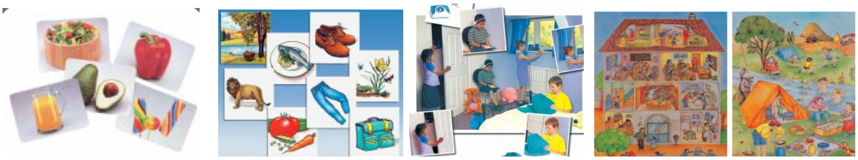


Fig. 1. Popular stimuli for aphasic children, often used also for adults

2 Background: Computer-Based Tools

In the last decades, several computer-based approaches have been proposed. Basically, they shift stimuli from the paper support to the computer screen. In addition to provide for a potentially infinite stimuli set, computerisation allows to register performance, in terms of response time and correctness, allowing for both patient-based and population-based statistics.

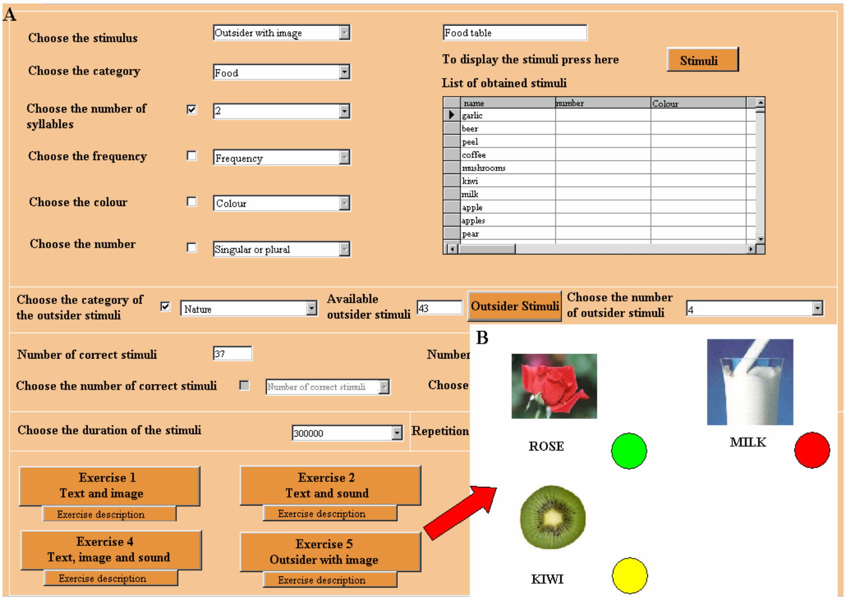
Computerised tools have been proposed by several groups:

- Grawemeyer et al [3] implemented a lexical decision system for patients with auditory perception deficits;
- Waller et al [4] describe TalksBac, a system for adult patients affected by non-fluent aphasia: it exploits the capability of recognising the most used words and short sentences; recently it has been transferred into the commercial sector as a

component of an integrated communication system called “Talk:About” marketed by Don Johnston Inc. (Chicago).

- Van de Sandt-Koenderman et al [5] developed PCAD, a portable system for improving daily life conversation capability;
- Bruce et al [6] and Wade et al [7] built and evaluate vocal recognition systems where patients may talk and see their pronounced words appearing on the screen;
- The German company Dr. Hein developed and distribute Evo-Care [8], a commercial telemedicine system that manages individual speech therapy sessions, particularly addressed to home therapy. Its client-server architecture allows clinicians to monitor patients' performance and to update treatment protocol from remote;
- TNP [9], distributed by BEAC Biomedical, is a more general tool for psychological and cognitive deficits. Its graphical interface allows building exercises particularly attractive for children and it is possible to update its archive to tailor the exercises for particular pathologies;
- E-Prime® [10] is a commercial tool that allows creating a wide set of exercises with different characteristics. Since it has been chosen for our application, it will be extensively described in the next section.

Each tool provides several common and specific functionalities; in our knowledge, however, there is no tool allowing easy generation of patient-tailored, scalable exercises, based on pictures from the patient’s own daily living.



Choose the stimulus: Outsider with image
Choose the category: Food
Choose the number of syllables ☑ 2
Choose the frequency ☐ Frequency
Choose the colour ☐ Colour
Choose the number ☐ Singular or plural

Food table
To display the stimuli press here
List of obtained stimuli

name	number	Colour
garlic		
beer		
peel		
coffee		
mushrooms		
kiwi		
milk		
apple		
apples		
pear		

Choose the category of the outsider stimuli ☑ Nature
Available outsider stimuli 43
Choose the number of outsider stimuli 4
Number of correct stimuli 37
Choose the number of correct stimuli ☐ Number of correct stimuli
Choose the duration of the stimuli 300000

Fig. 2. The interface created for facilitating the exercise generation with the tool E-Prime. In (A) the therapist chooses the type of exercise (in this case "outsider with image"), the category “food” with 2 syllables words for the stimuli, and the category “nature” for the outsider. The resulting patient’s interface is shown in (B).

3 The Solution Proposed

Unfortunately, no definite scientific evidence exists about the effectiveness of speech therapy tools. Thus, among the ones mentioned above, E-Prime has been chosen, due to its characteristics of flexibility, scalability and affordable cost. It is a powerful exercise generator and, with respect to its competitors, overcomes the two drawbacks of having only a standard set of exercises and/or being addressed mainly to children.

It is composed by two main modules:

- *E-Studio*, for creating the exercises through either a graphical interface or the proprietary *E-Basic* language, that allows fine modification and refinement of the objects that populate the exercises.
- *E-DataAid*, a data management module for analysing the patient's performance.

However, the tool requires a certain training for the therapists in order to learn how to generate exercises. In particular, using E-Studio is not easy for a non-expert indeed. To minimise training time and foster also the less computer-skilled therapists to use E-Prime, first of all a user interface has been developed for them. In a first step, the interface allowed to easily generate a set of *traditional* exercises, starting from the basic elements of the tool: windows, stimuli (textual, visual, acoustic), feedback, etc.

In a second step, which is the focus of the paper, we integrated an image elaboration tool, allowing to use familiar pictures, specific for each patient.

3.1 An Easier Interface for E-Prime

We exploited the E-prime utility that allows entering the exercise schema through a text file, and we developed a new interface for the therapist to easily create such opportune text files. The interface enables choosing one or more from a set of pre-defined exercises (decided by the speech therapists involved), covering the needs of most patients: word comprehension (both text and sound), lexical decision, i.e. association between sound and text, and various exercises of linguistic pragmatics (see Figure 2). To provide for system scalability and expandability, we built a database containing words, classified in categories, and the corresponding pictures and sounds. As shown in Table 1, each word is provided with length in syllables, use frequency, semantic class: these properties are useful for tailoring exercises to specific patients.

Table 1. Portion of word database created for E-Prime

Categories	*Properties*							
	Id	*Name*	*Picture*	*Colour*	*Syllables**	*Frequency*#	*Number*	*Sound*
Food	1	apple	Apple.bmp	red	2	FU	Sing.	Apple.wav
Objects	2	helmet	Helmet.bmp	black	2	HA	Sing.	Helmet.wav
Animals	3	cow	Cow.bmp	white	2	HU	Sing.	Cow.wav
Nature	4	rose	Rose.bmp	pink	2	FU	Sing	Rose.wav

#: FU = Fundamental, HA = High availability, HU = High use *: Number of syllables of Italian words.

3.2 The Image Elaboration Tool

To create patient-tailored, scalable exercises based on digital images, we developed a tool for adaptive image processing. It has been fully integrated into our E-Prime interface, in order to provide all the necessary data, extracted from the images, for defining and running the new type of exercise. The main goal of the Image Elaboration Tool (for brevity, IET) is to allow a simple method to identify different regions of the image; each of them corresponds to a particular object of the scene that has a specific semantic meaning for the patient: it may be the shape of a familiar object of daily life (a bag, a car, a vegetable, a lamp), a person (either a well known person, such as the patient's parents and/or children, or a generic cathegory, such as *baby*, *man*, *woman*) or an instance of a semantic category (animals, plants). The identification of several regions inside an image is a typical *segmentation task*. This is the core of the IET, but it is not the unique component. In fact, we have identified the following steps:

1) *Image Acquisition*. This step creates a collection of digital images which are revelant for patient rehabilitation. The images are preprocessed and resized in order to be in standarized and adjusted format for the next step.
2) *Image segmentation*. The goal is to generate a rectangular box which contains the target (object, person or category) that the patient has to recognize. We are forced to comply with the constraint of E-prime, that only manages rectangular regions. Unfortunately, this step is not strictly a typical image segmentation which can be perfomed only by considering colours, contours, texture or other pixel-based techniques. For example, in Figure 3, whatever classical low or medium level image processing segmentation algorithm would failed in detecting a proper box which comprehends all the parts of the women's shape (the hat, the face, the body, either the bare skin and the bath-cloth). We have called this task a *semantic segmentation*, because the output boxes has to take into consideration the semantic meaning of the several parts of the targets in the scene. The definition of the box is critical for the rehabilitation, as it corresponds to the sensible region on the screen the patient has to touch (or click over by the mouse). A touch/click outside the box means the failure of the recognition.

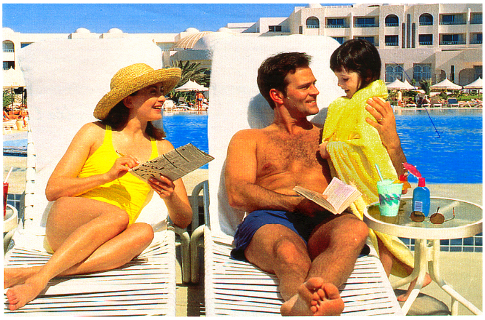


Fig. 3. A complex scene with three relevant targets: the women, the man, and the baby

3) *Region-Word Association*. Once a box has been identified for each target, it is associated to a proper word that represents the target. The same word will be presented to the patient (as a speech) in the E-prime environment during the exercise.

The classical solutions of image processing theory are not sufficient to realize an efficient application. In fact, there are several problems that are to be considered:

1) Scalability and size of the image collection for each patient. In order to assure the scalability of the system, the application has to allow the acquisition of several new images for each patients. This assures the possibility of presenting scenes of daily life provided from the specific past experience of each patient. By varying the number and the types of images (and consequently of the targets), we can enforce the efficacy of the rehabilitation exercises. From a strictly image processing point-of-view, this means that the segmentation process cannot have a-priori knowledge about the objects inside the scene or about their important properties, such as dimensions, colours, and spatial dispositions. For this reason, the segmentation task has to be the most general as possible. Conversely, as the great variety of the images, the segmentation task may be adaptive, to assure the best results. Of course, these two aspects (generality and adaptability) are full controversial, and the solution has to be an efficient compromise for this trade-off.
2) Poor quality of image sources. As we have previously explained, it is mandatory that the image collection is customized on the life experience of the patient and that it is as rich as possible. Unfortunately, very often the images provided by the patient's family have a very poor quality, in terms of classical image processing theory. In fact, they may be the output of scanning of printed pictures, or, even if they are in a digital form, they may have characteristics (poor brightness, bad light exposition, too much colours, poor contrast) which heavily compromise the next step (segmentation). Our tool performs automatic or semiautomatic pre-processing for the image adjustment in order to provide the best input image to the segmentation step.
3) Semantic scope of segmentation. To assure good performance, the interaction with the human expert is mandatory. The therapist must click by the mouse over one (or more) points of the scene which are relevant for the target semantic identification (for example, in Figure 3 the target women may be defined by three clicks, on the hat, on the bath – cloth and on the skin.)
4) Usability of the tool also for personnel which is not expert in image processing techniques. All the steps are automatic ore semi-automatic, with a very user-friendly interface. Our application keeps all the steps as the simplest as possible, with no a-priori knowledge of image processing algorithms. The therapist is asked only to click and confirm the results.

We have addressed and solved problems 1-4 by proposing a suitable compromise between the generality of classical general-purpose image processing algorithms and a customized solution. The general features of our proposed solution can be summarized in the following points, which underline the novelty of our approach in the field of image processing for rehabilitation applications: (a) a sufficient degree of generality (to assure scalability), (b) an adequate robustness against noise (to recover poor acquisitions), (c) an appropriate level of complexity for no–experts (to guarantee the practical usability of the system) and (d) a simple but efficient solution to assure a satisfactory semantic segmentation.

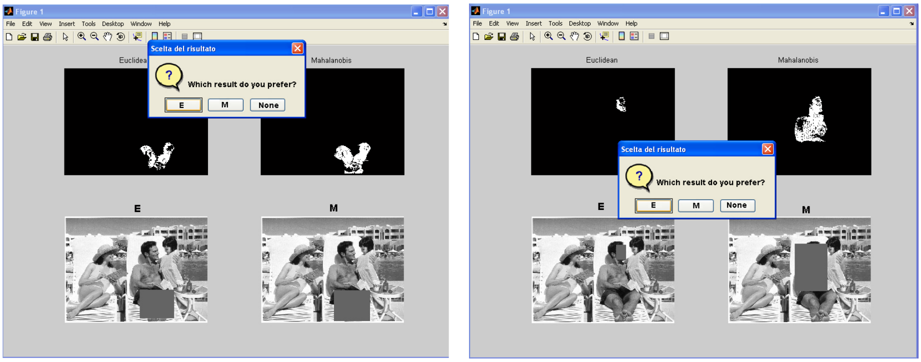
3.3 Novelty of the Approach

Our solution uses several techniques of classical image processing theory [11]: each of them is not new, but the novelty of our approach consists of three main aspects:

1) The choice of image processing algorithms and their tuning and customization. Among the several classical image processing techniques, we have chosen the ones which assure best results without asking the user to be expert in image processing: for this reason, even if some algorithms are parameter-dependent, the values of these parameters have been preliminarly set to reasonable figures. We have tested the choice of these values (thresholds, mask and window sizes) on a reasonable collection of 27 images. Chosen values are optimal in the sense that they minimize the number of user choices (number of interactions and feedback propagation, see third point). A significant part of our work has been extensive experimental tests and the consequent heuristic choice of all the parameters values involved in the image processing techniques. We have used the following [11] algorithms: *histogram equalization* (with automatic boundary intensity levels) for luminosity and contrast adjustment, *colour balancing* (in a semi-automatic form, by using bar indicators for each colour component) and *texture filters* (based on entropy computation in a 9X9 neighbourhood) to overcome the problem of over-segmentation in highly detailed objects. For the segmentation task, we have chosen a *region growing algorithm* (we have tested other algorithms, such as watershed techniques, but they did not give the same performance). Starting from the pixel pointed by the mouse click of the user (i.e. the therapist), a 6X6 window is considered. The average values of each colour component (RGB space) and a covariance matrix for all the pixels of the window are computed. Each average value of all the windows is compared to a proper value of threshold *T* in order to identify seed points for the region growing. As the choice of the distance measure is critical and could produce very different results, we have considered two options: Euclidean and Mahalanobis. The covariance matrix is used to set proper values of the threshold *T* for the seed point definition: if Max_C is the maximum value on the principal diagonal of matrix C (i.e. the maximum variance value), and we indicate its square root (i.e. standard deviation) with Max_{SD} we have set T = 1.5* $\mathrm{Max}_{\mathrm{sd}}$ for Euclidean distance and T = $\mathrm{Max}_{\mathrm{sd}}$ / 4 for Mahalanobis distance. The process of region growing is iterated and ends when all the pixels have been assigned to a region.
2) The classical algorithms are assembled and applied to the input images in a proper sequence, in automatic or semi-automatic form, to perform the brightness and contrast adjustment, the segmentation and the word-association. The only choices the user has to make are the relevant points of the target and, for each of them, one of the two output results, obtained by the Euclidean and the Mahalanobis distance. The two outputs are presented on the screen with two options and the user just keeps the best one, only from a visual inspection of the corresponding output boxes (see figure 4). It is clear in the picture that the two distance options perform equally in identifying foots, but Mahalanobis distance is clearly outperforming in identifyng the rest of the man's body.

3) The introduction of a mechanism of *interactivity plus feedback* which allows the therapist of choose the objects of the scene which are more relevant for the patient rehabilitation, and the corresponding boxes. Each target may be identified by any number of mouse clicks (generally, from one to three are enough), and the boxes resulting form the clicks-related segmentations are fused in a standard OR–operation.

The second and third points are clear if we consider the IET flow-chart, from the user's point of view. In section 3, such description is provided, together with a complete global overview of the integration of the IET in E-Prime.

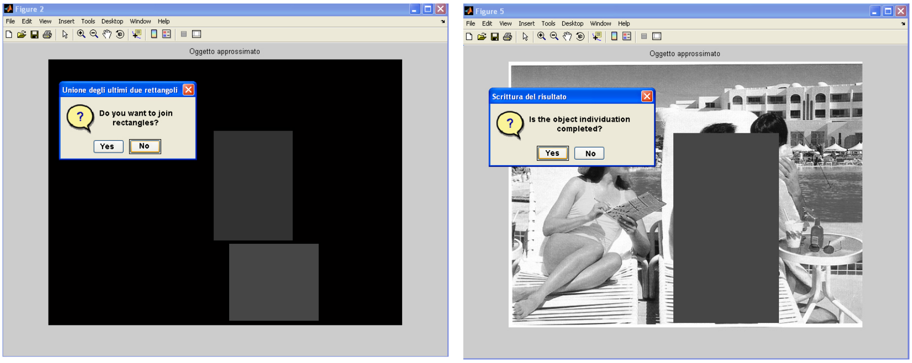

Scelta del risultato

Which result do you prefer?

Fig. 4. The user chooses the best output of the region-growing algorithm between two options (each of them corresponding to a different distance measure)

In all the cases we have tested, usually no more than one or two iteractions and feedbacks have been necessary to identify all the targets. One single segmentation takes few seconds (2-3 s) also in the intepreted implementation.

IET has been developed in Matlab (version 7.0), and a C-code executable file has been generated for the IET porting and integration in E-prime environment.


Unione degli ultimi due rettangoli

Do you want to join rectangles?

Scrittura del risultato

Is the object individuation completed?

Fig. 5. Two boxes may be fused to perform a real semantic segmentation. In this case, two different parts of target "man" are used.

4 The Integrated System

The functionality of the integrated system is the following: 1) when the therapist gets a new set of images (it typically happens when a new patient enters a therapeutic plan), he/she runs IET, which takes each image as an input, and acts accordingly to the algorithm explained above and summarised in the box below; 2) for each image, the IET output is a text file, having as many lines as the number of objects identified in the image, plus a line with the image file path; 3) the text file is further elaborated to make it fully E-prime compliant. In particular, if the word associated to the object is not found in the database, it is inserted into it. Each word is associated to a pre-defined category (Foods, Animals, Family, Geometric shapes, Clothes, Objects, Musical Instruments, Sports, Money, Human Body, Miscellanea), and the user is asked to associate the corresponding sound.

The above points are of the therapist's concern. Once they have been accomplished, the images are ready to be used within the therapeutic plan of the patient.

The IET algorithm

1. The image is properly acquired, adjusted and resized. The user chooses the image from a standard Windows interface and starts the segmentation phase.

2. *For each target:*

A) Click on a relevant point inside the target. The segmentation starts automatically and two results are presented, in terms of pixels associated to the identified regions and in term of corresponding rectangular boxes. The user is asked to choose one of the two results.

B) The user is asked if the result is fully satisfactory, or if he/she wants to discard the result, to keep it *and* to define another relevant point. In the last case, Step A is performed again and the feedback consists in using the previous obtained result in the phase C.

C) The user is asked if he/she wants to fuse the boxes corresponding to the two successive segmentations (see figure 5)

D) Step B and C are repeated until the user consider the corresponding box suitable for an efficient target identification.

E) The User is asked to enter the word associated to the target.

3. The coordinates of the boxes of all the targets in the image and the corresponding words are stored in a text file that will be further elaborated to produce a standardized format suitable for E-Prime interface.

4.1 Results

We test the performance of the IET algorithm in terms of its sensibility and specificity in associating objects with one or more suitable portions of the screen. In fact, due to the mentioned limitation of E-Prime, the algorithm approximates an object with a rectangular region. It may happen that the patient points to a correct screen position, but the system does not recognise this portion as part of the object. On the other hand, the patient could point to an incorrect screen position and, if it is close to the target object, the algorithm could recognise it as correct. We made an experiment with 10 healthy volunteers and 10 images, with an average of three "objects of interest" per image, for a total of 30 objects of interest. Volunteers have been enrolled from our laboratory personnel. Healthy people were needed because the evaluation goal was to test the algorithm performance, thus the subject is intended to point in the correct positions of the screen. A neutral observer was present during all the tests. For each volunteer:

for a generic object of interest o_i, we pronounce its name and ask the person to touch 5 different points of the screen that he considers suitable for indicating the object itself. The ratio "N. of times a click is recognised as incorrect by the system/total number of clicks" is the false positive rate (FPR);
then we pronounce the name of an object o_j close to o_i and again ask the person to touch 5 different points of the screen that he/she considers suitable for indicating o_j. The ratio "N. of times a click is recognised by the system as belonging to o_i /total number of clicks" is the false negative rate (FNR).

We obtained FPR=0.05 and FNR=0.15, a result that we consider very satisfying.

One possible bias in this evaluation procedure is represented by the use of the mouse, that allows a very precise pointing. Several patients will use a touch screen, much more easier for them than the mouse, but its lower precision could affect the results.

5 Conclusions and Future Works

A new procedure based on techniques of Image Processing has been developed as an added-value to a system for aphasic patient rehabilitation. The algorithms of pre-processing and region segmentation have been integrated into an exiting framework for rehabilitation based on a stimulus-response approach. The advantage is that now it is possible to generate patient-tailored exercises, inspired to his/her daily living. Testing with real patients has just started and we plan to have the first reliable results in few months. Future works from the technical point of view will concern (a) the full integration of the image processing tool and the sound database and (b) the exploitation of other image processing techniques of automatic object recognition to create a virtual desk populated by familiar objects in order to prepare new exercises for the patients.

References

1. The Stroke Prevention and Educational Awareness Diffusion (SPREAD) Collabora-tion. The Italian Guidelines for stroke prevention. Neurol Sci 2000, 21, 5–12, last ver-sion (2005) at www.spread.it
2. Bhogal, S.K., Teasell, R., Speechley, M., Albert, M.L.: Intensity of Aphasia Therapy, Impact on Recovery * Aphasia Therapy Works. Stroke 34(4), 987–993 (2003)
3. 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)
4. Waller, A., Dennis, F., Brodie, J., Cairns, A.Y.: Evaluating the use of TalksBac, a predictive communication device for non fluent adults with aphasia. Int J Lang Commun Disord. 33(1), 45–70 (1998)
5. Van de Sandt-Koenderman, M., Wiegers, J., Hardy, P.: A computerised communication aid for people with aphasia. Disabil Rehabil 27(9), 529–533 (2005)

6. 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)
7. Wade, J., Petheram, B., Cain, R.: Voice recognition and aphasia: can computers understand aphasic speech? Disabil Rehabil. 23(14), 604–613 (2001)
8. http://www.dr-hein.com/NEU/ECare/Presse/forum_logopaedie.pdf
9. http://www.beac.it/HTM/ITA/7tnp.htm
10. Schneider, W., Eschman, A., Zuccolotto, A.: E-Prime User's Guide. Psycology Software Tools Inc. (2002)
11. Gonzalez, R.C, Woods, R.E.: Digital Image Processing. Prentice-Hall, Englewood Cliffs (2002)

A Pattern Recognition Approach to Diagnose Foot Plant Pathologies: From Segmentation to Classification

Marco Mora[1], Mary Carmen Jarur[1], Leopoldo Pavesi[1],
Eduardo Achu[2], and Horacio Drut[1]

[1] Department of Computer Science, Catholic University of Maule, Talca, Chile
Casilla 617, Talca, Chile
{mora,mjarur,lpavesi}@spock.ucm.cl
http://ganimides.ucm.cl/mmora/
[2] Department of Kinesiology, Catholic University of Maule, Talca, Chile
eachu@ucm.cl

Abstract. Some foot plant diseases such as flat foot and cave foot are usually diagnosed by a human expert. In this paper we propose an original method to diagnose these diseases by using optical color foot plant images. A number of modern image processing and pattern recognition techniques have been employed to configure a system that can dramatically decrease the time in which such analysis are performed, besides delivering robust and reliable results to complement efficiently the specialist's task. Our results demonstrate the feasibility of building such automatic diagnosis systems that can be used as massive first screening methods for detecting foot plant pathologies.

1 Introduction

Cave foot and flat foot are common pathologies presented in children from the age of three. If these foot malformations are not detected and treated on time, they get worst during adulthood producing several disturbances such as pain and posture-related disorders [1].

The footprint is the surface of the sole in contact with the ground when the foot is planted. A simple method to obtain footprints is directly stepping the inked foot onto a paper on the floor. After obtaining the footprints, an expert analyzes them and assesses if they present pathologies.

In the diagnosis of foot plant pathologies instruments such as Pedobarograph [2] and the Podoscope [3] are used to capture the footprints. On the other hand, non automatic methods to segment footprints from digital images have been proposed in [4,5].

In previous work [6] we have presented the diagnosis of foot plant pathologies based on neuronal networks considering a nonautomatic methods to segment the foot plant pattern . In this work we propose a fully automatic methodology to segment footprints and to diagnose foot plant pathologies from optical

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 378–387, 2007.

color images. In order to acquire the images we have developed a flexible and low cost podoscope based on a digital camera [7]. With our digital podoscope, a large database containing currently more than 230 classified foot plant images was generated for analysis purposes. To automatically segment the footprint we employed advanced image processing techniques such as color models [8], complex diffusion [9] and active contours [10]. We defined an original representation for the segmented footprint and performed a principal component transform to efficiently reduce the dimension of the patterns [11]. Finally, we formulated the diagnosis of foot plant pathologies as a pattern recognition problem adopting a neural network approach [12]. The several innovations proposed in our method give as result a robust, inexpensive, and automatic system to achieve massive screening for the early detection of pathologies as flat foot and cave foot.

The paper is outlined as follows. Section 2 describes the pre-processing stage, where as section 3 shows the footprint segmentation. In section 4 the characteristics extraction stage is developed and section 5 shows the classification stage. Finally, section 6 shows some conclusions.

2 Acquisition and Preprocessing of Images

The image acquisition has been made by using our Podoscope. It consists of a robust metallic structure with adjustable height and transparent glass in its upper part, was developed for acquiring the footprints images. The patient must stand on the glass and the footprint image is obtained with a digital color camera in the interior of the structure. For adequate lighting, white bulbs are used. The system includes the regulation of the light intensity of the bulbs, which allows the amount of light to be adjusted for capturing images under different lighting conditions [7].

On the other hand, some non automatic method for the footprint segmentation from color images have been proposed. In [4] a method based in simple image processing operations can be found. In [5] a method based on neural networks has been proposed. Our work presents a proposal of an automatic method for the footprint segmentation based on the exploration of several color models.

Figure 1(a) shows a color image of the foot plant. Figure 1(b) shows a foot plant image with uniform background. The background color was selected to improve the contrast between the footprint and the rest of the image. The footprint image has been represented considering traditional color models (CIE XYZ, RGB and CMY), chromatic color models (HSV,YIQ, YCbCr) and uniform color models (CIE Lab, CIE Luv). For details on color models see [8].

After representing the footprint image in all of the above mentioned color models, we found that the most suitable model for segmentation purposes is the CIE LAb. Figures 1(c)-(e) show channels for the CIE Lab model. In this figure it can be seen that the channel A of Lab model fits very well with the footprint shape. In addition, we have also found that the influence of the toes on the footprint and on the footprint borders is almost completely eliminated for this channel.

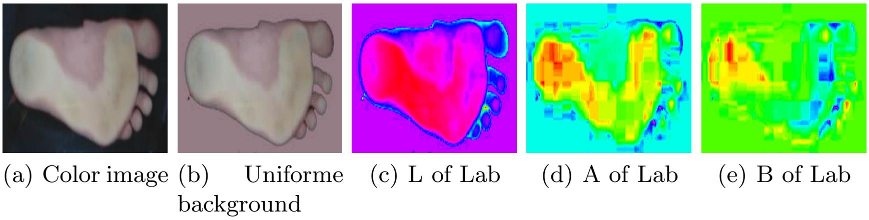


Fig. 1. Color foot plant image

In order to smooth the image we adopted a current diffusion model. Anisotropic diffusion has been introduced in [13]. This technique uses a variable diffusion coefficient in order to reduce the smoothing effects near the edges. The complex diffusion for image filtering is proposed in [9]. It has been reported as more aggressive than anisotropic diffusion at enhancing and preserving the edges. Due to these advantages, we adopted this method to enhance the footprint image. Figure 2 shows the results of smoothing iteratively channel A by using complex diffusion. With this procedure the discontinuities of the foot plant image are eliminated, preparing it for the contour detection.

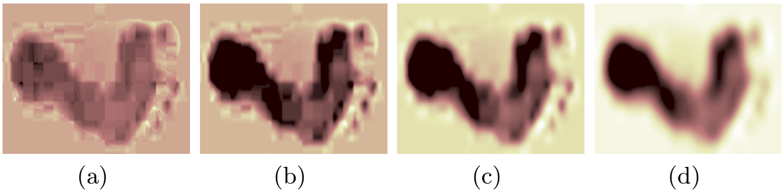


Fig. 2. Smoothing using complex diffusion

3 Footprint Segmentation Using Active Contours

Active contours are applied to detect objects in a given image using curve evolution techniques. The basic idea is to deform the curve until the boundary of the object under some constraints starting from an initial curve. A geometric active contour model based on the mean curvature is proposed in [14]. In order to cease the evolution of the curve at the edges, a function of the image gradient is normally used. The main problem of the traditional edge stopping term is that the edge stopping function is never exactly zero, and the moving curve may cross the boundaries of the object.

Recently, a method known as "active contour without edges" has been proposed in [10]. This active contours model does not use a stopping term based on the gradient. Instead, the stopping term is based on the Mumford-Shah segmentation techniques [15]. This model has significant advantages such as the ability to detect diffuse borders called cognitive contours [16], where the classical active contours methods based on the gradient are not applicable. Due to these advantages, this work uses the latter approach to achieve a robust footprint segmentation.

Figure 3 shows the contours detection by using active contours without edges. It clearly shows how the curve evolves until exactly stopping at the footprint border. It shows that the detected contour corresponds to the footprint contour.

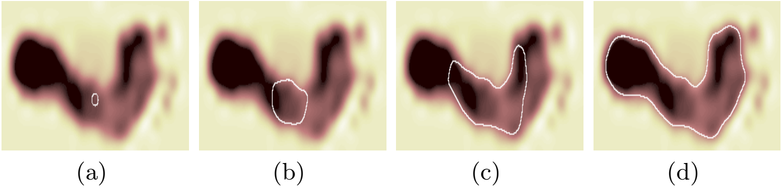


Fig. 3. Edge detection using actives contours

4 Footprint Representation and Characteristics Extraction

After performing the segmentation, the footprint is represented by a vector containing the width -in pixels- of the segmented footprint, without toes, by each column in the horizontal direction. Because every image has a width vector with different length, the vectors were normalized to have the same length. The value of each element was also normalized to a value in the range of 0 to 1. Figure 4(a)-(c) show the normalized width vectors (rugged red signal) of a flat, a normal and a cave foot.

As a way to reduce the dimensionality of the inputs to the classifier, a principal components analysis was used [11]. Given an eigenvalue λ_i associated to the covariance matrix of the width vector set, the percentage contribution γ_i (1) and the accumulated percentage contribution APC_i (2) are calculated by the following expressions:

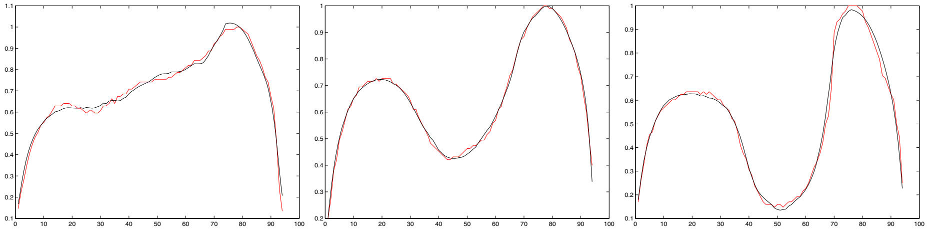
$$\gamma_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_i} \tag{1}$$

$$APC_i = \sum_{j=1}^{i} \gamma_i \tag{2}$$

Table 1 shows the value, the percentage contribution and the accumulated percentage contribution of the first nine eigenvalues λ_1 ... λ_9. It is possible to note that from the 8^{th} eigenvalue the percentage contribution is close to zero, so that it is enough to represent the width vector with the first seven principal components. Figure 4 shows the resulting approximation of the normalized width vectors using the seven first main components (smoothed black signal) for all three classes.

Table 1. Contribution of the first 9 eigenvalues

-	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9
Value	0.991	0.160	0.139	0.078	0.055	0.0352	0.020	0.014	0.010
Percentage contribution	63.44	10.25	8.95	5.01	3.57	2.25	1.31	0.94	0.65
Accumulated contribution	63.44	73.7	82.65	87.67	91.24	93.50	94.82	95.76	96.42



(a) Flat class (b) Normal class (c) Cave class

Fig. 4. Principal components approximation

5 Training and Validation of the Neural Network Classifier

5.1 Training of the Neuronal Network

The training set has a total of 199 images, 13% corresponding to a flat foot, 63% to a normal one and 24% to a cave foot. The first seven principal components were calculated for all the width vectors in order to build the training set. To classify the foot as flat, normal or cave, a Multilayer Perceptron (MLP) trained with Bayesian Regularization Backpropagation (BRBP) was used [17]. The Neural Network (NN) has 7 inputs, one for each main component, and 1 output. The output takes a value of 1 if the foot is flat, a value of 0 when the foot is normal and a value of -1 when the foot is cave.

To determine the amount of neurons at the hidden layer, the procedure described in [12] was followed. Batch learning was adopted and the initial network weights were generated using the Nguyen-Widrow method [18] since it increases the convergence speed of the training algorithm.

The details of this procedure are shown in table 2, where NNCO corresponds to the number of neurons at the hidden layer, SSE is the sum squared error and SSW are the sum squared weights. From the table 2 it can be seen that from 4 neurons in the hidden layer, the SSE, SSW and the effective parameters remain practically constant. As a result, 4 neurons are considered for the hidden layer.

Figure 5 shows the training process considering 4 neurons for the hidden layer. It can be observed that the SSE, SSW and the network effective parameters are relatively constant over several iterations, which means that the training process

Table 2. Determining the amount of neurons in the hidden layer

NNCO	Epochs	SSE	SSW	Effective parameters	Total parameters
1	114/1000	22.0396/0.001	23.38	08.49	10
2	51/1000	12.4639/0.001	9.854	16.30	19
3	83/1000	12.3316/0.001	9.661	19.70	28
4	**142/1000**	**11.3624/0.001**	**13.00**	**26.1**	**37**
5	406/1000	11.3263/0.001	13.39	28.7	46
6	227/1000	11.3672/0.001	12.92	26.2	55

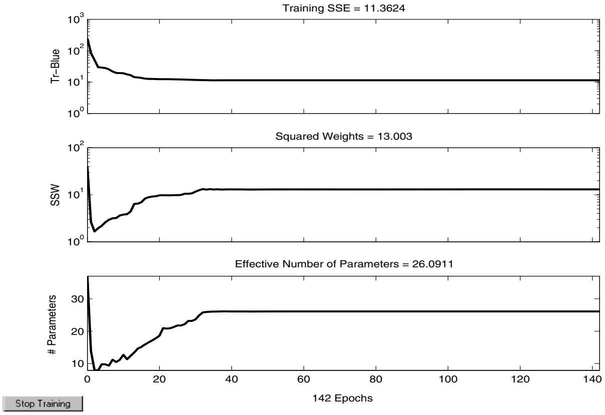
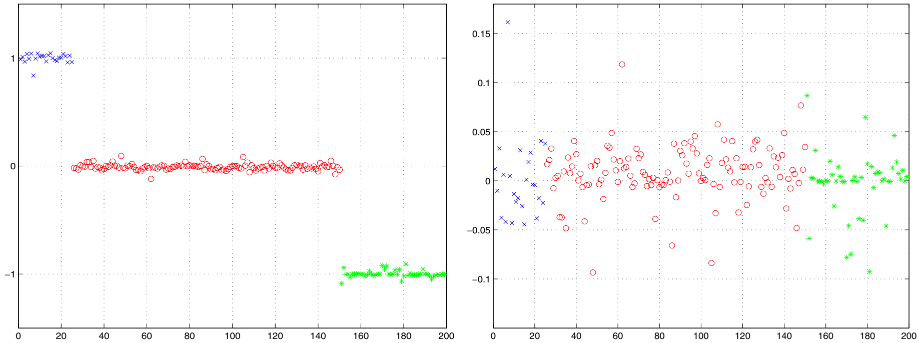


Fig. 5. Evolution of the training process for 4 neurons in the hidden layer. Top: SSE evolution. Center: SSW evolution. Down: Effective parameters evolution.

has been appropriately made [12]. From the figure it is important to emphasize that the classification errors are not very small values ($SSE = 11.3624$). This behavior assures that the network has not memorized the training set, and that will therefore generalize well.

5.2 Classification of the Training Set

The training set is classified with the weights found in the training phase. Figure 6 shows the classification for each pattern of the training set. This set was previously sorted in order to achieve a better visual representation of the classifier output. Figure 6(a) presents the classification results of the flat foot pattern using blue crosses, the normal foot pattern using red circles and the cave foot pattern in green asterisks. Figure 6(b) shows the classification error of the training set and provides a way to make a qualitative evaluation of the training set. It can be seen that the neural network has been able to classify the training set. The value obtained for SSE attests to the strength of the neural network classifier in the given conditions, which means that it will perform its task preserving a generalization behavior.



(a) Training set classification (b) Training set classification error

Fig. 6. Classification of the training set

The classification error of a network input $Input(i)$ is defined as:

$$Error(i) = |Target - Output(i)| \tag{3}$$

where $Target$ is the value assigned to the class, and $Output(i)$ the network output to the corresponding $Input(i)$.

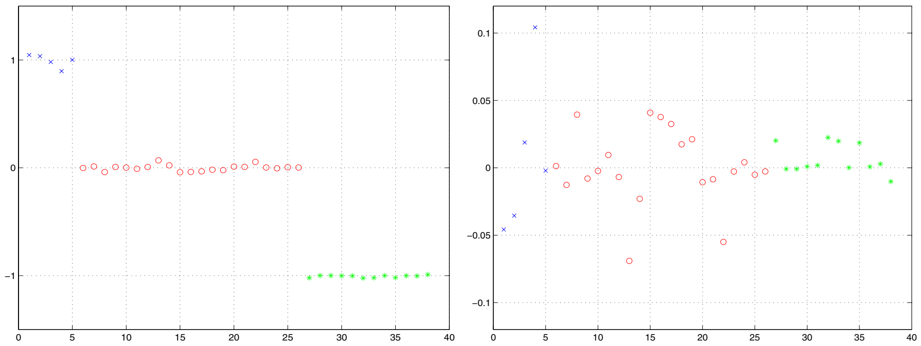
Table 3 presents several measures to quantify and characterize the classification of the training set, such as the Output Average Value, the Output Standard Deviation, the Maximum Error, the Minimum Error and the Error Average Value. Considering that the measures have been computed using the network weights obtained from a training process that allows a suitable generalization, the adopted values of the measures are a comparison reference.

Table 3. Statistics of the training set classification

Type	Number	Percentage %	Target	Average Output	Output Standard Deviation	Max Error	Min Error	Average Error
Flat	25	13%	1	0,9987	0,0426	0,1616	0,0010	0,03
Normal	125	63%	0	-0,0083	0,0276	0,1186	0,0001	0,02
Cave	49	24%	-1	-0,9970	0,0319	0,0926	0,0002	0,02

5.3 Classification of the Validation Set

The validation set contains 38 new real footprint images classified by an expert, where 13% corresponds to a flat foot, 55% to a normal foot and 32% to a cave foot. The corresponding normalized-width vector was calculated for each segmented footprint, and then, by performing principal component decomposition, only the first 7 axes were presented to the trained neural network.



(a) Validation set classification (b) Validation set classification error

Fig. 7. Classification of the validation set

Table 4. Classification statistics of the validation set

Type	Number	Percentage %	Target	Average Output	Output Standard Deviation	Max Error	Min Error	Average Error
Flat	5	13%	1	0,9921	0,0597	0,1042	0,0021	0,04
Normal	21	55%	0	0,0001	0,0277	0,069	0,0013	0,02
Cave	12	32%	-1	-1,0063	0,0108	0,0224	0,0001	0,01

Figure 7 presents the classification results of the validation set. From the figure it is possible to observe a similar behavior to the classification of the training set. The classification errors allow to correctly classify each pattern of the validation set.

Table 4 presents the classification indexes for the validation set. The values of the standard deviation and the average error for every class are in the same order of magnitude that the classification values of the training set. This fact constitutes a quantitative confirmation that the obtained network is performing an appropriate generalization. Therefore, the network will be able to classify other patterns that do not belong to the initial training set.

6 Conclusions

In this paper a computational method to diagnose foot plant diseases in an automatic and reliable way was presented. Our approach is based on the utilization of several modern digital image processing techniques and on the formulation of the objectives as a pattern recognition problem. The results obtained in this study demonstrate the feasibility of implementing a system for the early, reliable and massive diagnosis of foot plant pathologies and are encouraging to make further research in the area.

Additionally, during the research a database containing a large amount of classified images with different pathologies from several patients has been generated and made available to the research community.

References

1. Valenti, V.: Orthotic Treatment of Walk Alterations (in spanish), 1st edn. Panamerican Medicine (1979)
2. Patil, K., Bhat, V., Bhatia, M., Narayanamurthy, V., Parivalan, R.: New online methods for analysis of foot preesures in diabetic neuropathy. Frontiers Me. Biol. Engg. 9, 49–62 (1999)
3. Morsy, A., Hosny, A.: A new system for the assessment of diabetic foot planter pressure. In: Proceedings of 26th Annual International Conference of the IEEE EMBS, pp. 1376–1379 (2004)
4. Chu, W., Lee, S., Chu, W., Wang, T., Lee, M.: The use of arch index to characterized arch height: a digital image processing approach. IEEE Transaction on Biomedical Engineering 42, 1088–1093 (1995)
5. Mora, M., Sbarbaro, D.: A robust footprint detection using color images and neural networks. In: Sanfeliu, A., Cortés, M.L. (eds.) CIARP 2005. LNCS, vol. 3773, pp. 311–318. Springer, Heidelberg (2005)
6. Mora, M., Jarur, M.C., Sbarbaro, D.: Automatic diagnosis of the footprint pathologies based on neural networks. In: Proceedings of 8th International Conference on Adaptive and Natural Computing Algoritms (ICANNGA'07). LNCS, Springer, Heidelberg (to appear, 2007)
7. Mora, M.: Pattern recognition system based on neuronal networks to classify foot plant images as flat foot, normal foot and cave foot (in spanish). Master's thesis, Department of Electrical Engineering, University of Concepcion, Casilla 160-C, Concepcion, Chile, (2004)Available from http://ganimides.ucm.cl/mmora/
8. Cheng, H., Jiang, X., Sun, Y., Wang, J.: Color image segmentation: advances and prospects. Pattern Recognition 34, 2259–2281 (2001)
9. Gilboa, G., Sochen, N., Zeevi, Y.: Complex diffusion processes for image filtering. In: Kerckhove, M. (ed.) Scale-Space 2001. LNCS, vol. 2106, pp. 299–307. Springer, Heidelberg (2001)
10. Chan, T., Vese, L.: Active contours without edges. IEEE Transaction on Image Processing 10, 266–277 (2001)
11. Jollife, I.: Principal Component Analysis. Springer, Heidelberg (2002)
12. Foresee, D., Hagan, M.: Gauss-newton approximation to bayesian learning. In: Proceedings of the 1997 International Joint Conference on Neural Networks (IJCNN'97), pp. 1930–1935 (1997)
13. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 629–639 (1990)
14. Caselles, V., Catte, F., Coll, T., Dibos, F.: A geometric model for active contours in image processing. Numerische Mathematik 66, 1–31 (1993)
15. Mumford, D., Shah, J.: Optimal approximation by piecewise smooth functions and associated variational problems. Commun. Pure and Applied Mathematics 42, 577–685 (1989)

16. Kanizsa, G.: La grammaire du voir. Essais sur la perception: Diderot Edituer. Arts et Sciences (1997)
17. Mackay, D.: Bayesian interpolation. Neural Computation 4, 415–447 (1992)
18. Nguyen, D., Widrow, B.: Improving the learning speed of 2-layer neural networks by choossing initial values of the adaptive weights. In: Proceedings of 1990 International Joint Conference on Neural Networks (IJCNN'90), pp. 21–26 (1990)

A Novel Way of Incorporating Large-Scale Knowledge into MRF Prior Model

Yang Chen[1,2], Wufan Chen[1], Pengcheng Shi[1], Yanqiu Feng[1],
Qianjin Feng[1], Qingqi Wang[2], and Zhiyong Huang[2]

[1] Institute of Medical Information&Technology, School of Biomedical Engineering
Southern Medical University, Guangzhou, 510515, China
{kshzh,chenwf,shipch,foree,fqianjin}@fimmu.com
[2] The 113 Hospital of People's Liberation Army, Ningbo, 315040, China
{wangqingqi,jd21}@163.com

Abstract. Based on Markov Random Fields (MRF) theory, Bayesian methods have been accepted as an effective solution to overcome the ill-posed problems of image restoration and reconstruction. Traditionally, the knowledge in most of prior models is from a simply weighted differences between the pixel intensities within a small local neighborhood, so it can only provide limited prior information for regularization. Exploring the ways of incorporating more large-scale knowledge into prior model, this paper proposes an effective approach to incorporate large-scale image knowledge into MRF prior model. And a novel nonlocal prior is put forward. Relevant experiments in emission tomography prove that the proposed MRF nonlocal prior is capable of imposing more effective regularization on original reconstructions.

Keywords: Markov Random Fields (MRF), Bayesian reconstruction, emission tomography, nonlocal prior.

1 Introduction

The reconstruction of an unknown image f from measurement data g which suffer from low counts and noise contamination, is often an ill-posed problem [1-2]. How to effectively overcoming such ill-posedness for image reconstruction has been widely studied in the past twenty years. Among all the resolutions, Bayesian methods or maximum *a posteriori* (MAP) ways have already been accepted as an effective solution to above problem. Traditionally, based on Bayesian and MRF theory, a generic contextual constraints can be transformed into some kind of prior knowledge to regularize the solution to the original ill-posed reconstruction problem [2-4]. We term them local prior for simplicity.

Simply enlarging the neighborhood seems to be a straightforward method to use the information in a large region, and we will test its effectiveness in experiment1. Illuminated by the nonlocal idea in [5], we devise a novel way of incorporating large-scale knowledge into MRF Prior, from which a spatial-varying ***nonlocal*** prior model can be built. In experiment2, we perform emission tomography using the proposed

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 388–392, 2007.

nonlocal prior and some other local priors. Relevant comparisons and analyses show the proposed nonlocal prior's good properties in image reconstruction.

2 Theory

We devise a novel way of incorporating large-scale knowledge into MRF Prior. In the building of such nonlocal prior, a large neighborhood Ne is set to incorporate geometrical configuration knowledge of image. And, the nonlocal prior weight w_{bj_NL} is estimated through a similarity measure of the respective neighborhoods centered on pixel b and pixel j, other than just a simple inverse proportional calculation of the spatial distance between two individual pixels. Such neighborhood-based similarity is computed as a decreasing function of the Gaussian Euclidean distance between the all the pixel densities within above two neighborhoods. The corresponding building is generalized as follows:

$$U(f)=\sum_{j}U(f,j)=\sum_{j}\sum_{b\in Ne_j} w_{bj\_NL}\left(f_b-f_j\right)^2. \tag{6}$$

$$w_{bj\_NL}=\exp\left(-\left\|V(n_b)-V(n_j)\right\|^2 \Big/ h^2\right)\Big/ dis_{bj}. \tag{7}$$

$$dis_{bj}=\sqrt{\left(b_x-j_x\right)^2+\left(b_y-j_y\right)^2}. \tag{8}$$

Here $V\left(n_b\right)$ and $V\left(n_j\right)$ are the two pixel density vectors in the two square comparing neighborhoods n_b and n_j that center at pixel b and pixel j, respectively. $\|A-B\|$ denotes Euclidean distance between two pixel density vectors A and B. dis_{bj} denotes the 2-D space distance between pixel b and pixel j whose 2-D coordinates are $\left(b_x,b_y\right)$ and $\left(j_x,j_y\right)$, respectively. h controls the decay of the exponential function of the weights. Through the computation of the similarities of the comparing neighborhoods centered on the pair of pixels in the neighborhood Ne, the weights of each pair of pixels in Ne for the nonlocal prior are distributed across the more similar configuration.

3 Experimentation and Analyses

In experiments, a synthetic simulated phantom data with 128×128 square pixels is used for emission reconstruction. Fig.1 shows the synthetic simulated phantom. It is a Shepp-Logan head phantom with pixel intensities from 0 to 8, and the number of counts for the sinogram 3×10^5. This phantom has pixel intensities that range from 0 to 5, and the number of counts for the sinogram amounts to 4×10^{5}. The simulated data in the sinogram are Poisson distributed and the percentage of delayed coincidences is set to be 10%. The transition probability matrix used in the reconstructions corresponds to parallel strip-integral geometry with 128 radial samples and 128 angular samples distributed uniformly over 180 degrees.

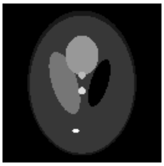


Fig. 1. Phantom image data used in experiments

In the first experiment, we perform the Bayesian reconstructions using the simple QM prior with neighborhoods of different sizes to see whether large-scale knowledge can be effectively exploited by just enlarging the neighborhood sizes. We choose the prior energy with form $v(t)=t^2$, in which three neighborhoods with sizes 3×3, 7×7 and 11×11 are used. For generalization reason, when neighborhoods of sizes 7×7 and 11×11 are used, weight w_{bj} in (4) is computed by following three weighting strategies:

$$(1)\ ,\ w_{bj}=1\Big/\sqrt{(b_x-j_x)^2+(b_y-j_y)^2}$$

$$(2)\ ,\ w_{bj}=1\Big/\sqrt{(b_x-j_x)^4+(b_y-j_y)^4}$$

$$(3)\ ,\ w_{bj}=1\Big/\sqrt{(b_x-j_x)^8+(b_y-j_y)^8}$$

And only the first weighting strategy is used for QM prior with 3×3 neighborhood.

In the study, the MAP-OSL reconstruction algorithm [5] is used, and the filtered FBP reconstruction is chosen as the initial image in iteration. The reconstruction quality is evaluated in terms of minimization of *SNR* (signal to noise rate):

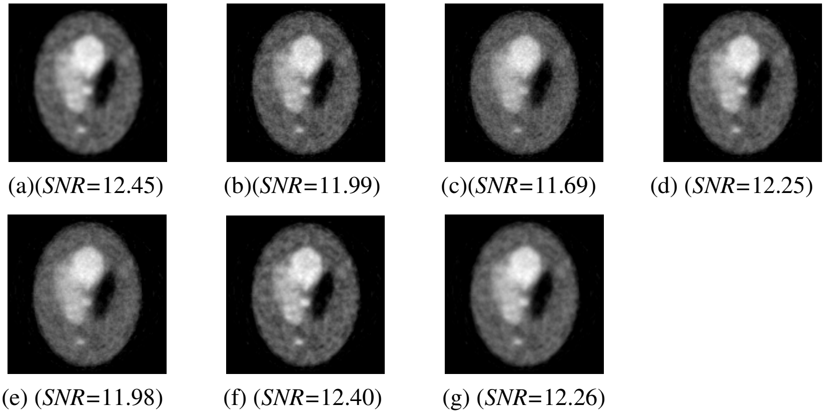
$$SNR=10\log_{10}\left(\sum_{x,y}(f(x,y)-\bar{f})^2\Big/\left(\sum_{x,y}(f(x,y)-f_{phantom}(x,y))^2\right)\right). \quad (9)$$

where f , $\bar{f}$ and $f_{phantom}$ denote the objective image , the mean of the objective image and the phantom image data in Fig.1, respectively.

In this experiment, we find that, with reasonable global hyperparameter tuning, the *SNR*s of the reconstructed images in the iteration process tend to decrease for reconstructions using priors with neighborhoods of sizes 7×7 and 11×11, and such decreases are especially prominent when the first weighting strategy is chosen. And it is found on the other hand that reconstructions using QM prior with local 3×3 neighborhoods can obtain reconstruction with stabler and higher *SNR*s.

Fig.2 showed the reconstructions using QM prior with different neighborhood sizes and weighting strategies. The 100th reconstructed images in iteration are chosen. The values of global hyperparameter β are also chosen manually to give the best reconstructed images in terms of maximization of *SNR*s.

From above, we can see that the connection between two pixels in the objective image becomes unclear as their distances increase. And, with above simply weighting strategies, just enlarging neighborhood can not improve the Bayesian reconstruction.



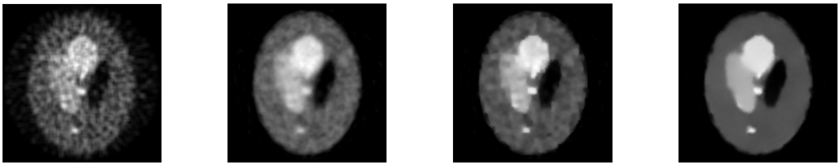
(a)(*SNR*=12.45) (b)(*SNR*=11.99) (c)(*SNR*=11.69) (d) (*SNR*=12.25)

(e) (*SNR*=11.98) (f) (*SNR*=12.40) (g) (*SNR*=12.26)

Fig. 2. Bayesian reconstructions using QM prior with different neighborhood sizes and weighting strategies(*SNR*s in the right parentheses below the respective reconstructed images): (a) 3×3 neighborhoods with the first weighting strategy, (b) 7×7 neighborhoods with the first weighting strategy, (c) 11×11 neighborhoods with the first weighting strategy, (d) 7×7 neighborhoods with the second weighting strategy, (e) 11×11 neighborhoods with the second weighting strategy, (f) 7×7 neighborhoods with the third weighting strategy, (g) 11×11 neighborhoods with the third weighting strategy

In the second experiment, we assess the proposed nonlocal prior by applying the nonlocal prior, the local QM prior (3x3 neighborhood) and local Huber prior (3x3 neighborhood) in the reconstructions of emission images. For reconstructions using QM prior and Huber prior, the values of parameter β and δ are also chosen by hand to give the best reconstructed images in terms of maximization of *SNR*s. The value of β is set to 1.5. The threshold Parameter δ of Huber prior is fixed to be 0.2. As to the Bayesian reconstructions using the proposed nonlocal prior, the values of parameters are also chosen manually to give the best reconstructed images in terms of maximization of *SNR*. The values of global parameter β are set to 0.4. And the values of parameter h of the nonlocal prior in (7) are set to 0.8. For all the Bayesian reconstructions using the nonlocal prior, the N_j in (6) is set to be 11x11 neighborhoods. n_b and n_j in (7) are both set to be 7x7 neighborhoods. The filtered FBP reconstructions are also chosen as the initial image in iteration.

Fig.3 shows the 100th iterated images (the iterations become stable after 100 iterations). We can see that the reconstructions using the proposed nonlocal prior yield much more appealing images than other methods. Reconstructions using the nonlocal prior exhibit excellent performances in both suppressing noise effect and preserving edges. And the corresponding reconstructed images are free from the unfavorable oversmoothing effect problem for quadratic QM prior and the staircase effect problem for nonquadratic Huber prior. Below the respective reconstructed images, we also list the computed *SNR*s for the corresponding reconstructions. Clearly, the *SNR*s comparisons indicate that the reconstructions using the proposed nonlocal prior are able to produce images with considerable higher *SNR*s than other reconstruction methods.



(a) (*SNR*=11.34) (b) (*SNR*=12.45) (c) (*SNR*=13.50) (d) (*SNR*=14.44)

Fig. 3. Emission FBP reconstruction and Bayesian reconstructions using different priors for the two phantom data (*SNR*s in the right parentheses below the respective reconstructed images): (a) FBP reconstruction, (b) QM prior reconstruction, (c) Huber prior reconstruction, (d) Proposed nonlocal prior reconstruction

4 Conclusions

From above analyses and experiments, we can see that just enlarging the neighborhoods of priors can not effectively incorporate more large-scale knowledge into Bayesian reconstruction. And on the other hand, our proposed nonlocal prior, which is devised to exploit the large-scale or global connectivity knowledge in objective image, is able to impose a much more effective and robust regularization on reconstruction than the local QM prior and Huber prior. In addition, stemming from MRF theory, the proposed approach is theoretically correct and can be easily implemented.

References

1. Bertero, M., Poggio, T., Torre, V.: Ill posed problems in early vision. Proc. IEEE 76, 869–889 (1988)
2. Lange, K.: Convergence of EM image reconstruction algorithms with Gibbs smoothness. IEEE Trans. Med. Imag. 9, 439–446 (1990)
3. Stan, Z.: Markov Random Field Modeling in image Analysis, pp. 1–40. Springer, Tokyo (2001)
4. Gindi, G., Rangarajan, A., Lee, M., Hong, P.J., Zubal, G.: Bayesian Reconstruction for Emission Tomography via Deterministic Annealing. In: Barrett, H., Gmitro, A. (eds.) Information Processing in Medical Imaging, pp. 322–338. Springer, Heidelberg (1993)
5. Green, P.J.: Bayesian reconstruction from emission tomography data using a modified EM algorithm. IEEE Trans. Med. Imag. 9, 84–93 (1999)

Predictive Modeling of fMRI Brain States Using Functional Canonical Correlation Analysis

S. Ghebreab[1,2], A.W.M. Smeulders[1], and P. Adriaans[2]

[1] ISLA lab, Informatics Institute, University of Amsterdam, The Netherlands
[2] HCS lab, Informatics Institute, University of Amsterdam, The Netherlands

Abstract. We present a novel method for predictive modeling of human brain states from functional neuroimaging (fMRI) data. Extending the traditional canonical correlation analysis of discrete data to the domain of stochastic functional measurements, the method explores the functional canonical correlation between stimuli and fMRI training data. Via an incrementally steered pattern searching technique, subspaces of voxel time courses are explored to arrive at (spatially distributed) voxel clusters that optimize the relationship between stimuli and fMRI in terms of redundancy. Application of the method for prediction of naturalistic stimuli from unknown fMRI data shows that the method finds highly predictive brain areas, i.e. brain areas relevant in processing the stimuli.

1 Introduction

Prediction of brain states directly from non-invasive measurements of brain activity has emerged as a powerful alternative to correlation of external stimuli with characteristic brain activity. The advantage of inverting the task from correlating external stimuli with brain activity to predicting stimuli from brain activity is that it facilitates evaluation of spatially distributed brain responses to complex uncontrolled stimuli [1]. Prediction of brain states from brain activity data, however, is a challenging task in its own right, requiring advanced data processing methods that go beyond conventional mass univariate data analysis of functional neuroimage data.

Various multivariate pattern classification approaches have recently been proposed for prediction of brain states directly from fMRI measurements. In these approaches, a classifier is trained on fMRI data to discriminate between different known brain states and then applied to predict brain states from unknown fMRI data. Several neuroimage studies (e.g., [2], [3]) successfully predicted complex stimuli from fMRI using multivariate pattern classification approaches, showing their ability to identify response patterns across the full spatial extent of the brain without attempting to localize function.

Here, we extend on the incremental functional multivariate regression method proposed in [4], which exploits the continuous nature of external stimuli and brain processes. Cast into an incremental pattern searching framework, this method performs functional principal component regression to find distributed voxel clusters that optimize a linear model in terms of F-statistic. In this work, we pursue canonical correlation analysis rather than principal component analysis in order to fully exploit correlated variation in stimuli and brain activity data. We show that in comparison to functional principal component analysis, functional canonical correlation analysis captures functional subspaces that are more appropriate for prediction of brain states.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 393–397, 2007.

2 Method

In the remainder, we consider stimuli and fMRI data as continuous functions of time, sampled at the scan interval and subject to observational noise. The functional form of the discrete data points is obtained by fitting a continuous curve through them.

2.1 Functional Data Representation

The four-dimensional fMRI data $I(\mathbf{x}, t)$, where $\mathbf{x} \in \Re^3$ denotes the spatial position of a voxel and t denotes its temporal position, defines the predictor set, i.e. the independent variable set. The image $I(\mathbf{x}, t)$ is preprocessed to arrive at a (spatially) normalized data set. We represent each voxel time course in functional form by $f(t)$, with t denoting the continuous path parameter. The vector $\mathbf{f} = [f_1, ..., f_S]^T$ of S functionalized voxel time courses contains the complete set of independent variables.

We represent the stimuli data by the functional $g(t)$, t being the continuous time parameter. We register $g(t)$ to each voxel time course $f_s(t)$ in order to be able to compare equivalent time points on stimulus and brain activity data, i.e. to capture subtle localized variations in Haemodynamic delays across brain regions and subjects. Curve registration here reduces to finding the small shift and nonlinear transformation that minimizes a global alignment criteria in least squares sense. Registration of $g(t)$ to all voxel time courses S results in the dependent variable set $\mathbf{g}(t) = [g_1(t), ..., g_S(t)]^T$.

2.2 Functional Canonical Correlation

We employ canonical correlation analysis to capture the relationships between the two sets of functions $\mathbf{f}(t)$ and $\mathbf{g}(t)$. Functional canonical correlation analysis [5] explores the dominant modes of correlation between each pair of functions $f_s(t)$and $g_s(t)$. Canonical weighting functions $\eta(t)$ and $\xi(t)$ are sought that maximize the sampled squared correlation of

$$\int \eta(t) f_s(t) dt \quad \text{and} \int \xi(t) g_s(t) dt \tag{1}$$

across all pair of functions. The correlation that results from the maximizing weight functions $\eta_1(t)$ and $\xi_1(t)$ is the first canonical correlation ρ_1. The corresponding first pair of canonical loadings are defined as

$$f_{s1} = \int \eta_1(t) f_s(t) dt \quad \text{and} \quad g_{s1} = \int \xi_1(t) g_s(t) dt. \tag{2}$$

The second pair of canonical weight functions $\eta_2(t)$ and $\xi_2(t)$ that maximize the correlation ρ_2 is found in the same manner. This second set is orthogonal to the first, i.e. satisfies the constraints

$$\int \eta_1(t)\eta_2(t) dt = 0 \quad \text{and} \int \xi_1(t)\xi_2(t) dt = 0. \tag{3}$$

This process in repeated until Q main modes of correlation have been found. In order to arrive a small number of meaningful modes of correlation, we regularize the weighting functions by penalizing their roughness (see [5] for more detail).

2.3 Overall Fit Function

To determine the amount of shared variance, we make use of the redundancy index [6], which is analogous to the R^2 statistic in multiple regression. This index provides a summary measure of the ability of the set of independent variables $\mathbf{f}(t)$ to explain variation in the dependent variables $\mathbf{g}(t)$. Here we compute

$$R = \frac{1}{Q}\sum_{q=1}^{Q}(\rho_q^2 \frac{1}{S}\sum_{s=1}^{S} g_{sq}^2) \tag{4}$$

where $\frac{1}{S}\sum_s g_{sq}^2$ is the amount of shared variance in $\mathbf{g}(t)$ explained by $\xi_q(t)$ and the squared canonical correlation ρ_q^2 is the amount of variance in $\xi_q(t)$ that can be explained by $\eta_q(t)$. We use this measure as the overall fit function that drives the search for voxel-time courses that are strongly related to the external stimuli.

2.4 Incremental Subspace Exploration

In order to efficiently find the subset of voxels that maximizes the overall fit R, we use the incremental search technique described in [7]. With help of this technique, at each increment a refined voxel subset is obtained and evaluated in terms of equation 4. At increment i, the subset of voxels time courses $S^i \subset S$ gives rise to canonical weight functions $\eta_q^i(t)$ and $\xi_q^i(t)$ and loadings

$$f_{sq}^i = \rho_q \int f_s(t)\eta_q^i(t) \quad \text{and} \quad g_q^i = \rho_q \int g(t)\xi_q^i(t). \tag{5}$$

Then, the set $\mathbf{F}^i = [\mathbf{f}_1^i, ..., \mathbf{f}_S^i]$ is explored using $\mathbf{g}^i = [g_1^i, ..., g_Q^i]$ as pilot. In short, elements are selected from $\mathbf{F}^i$ that have smallest Euclidean distance to $\mathbf{g}^i$ and form one or more spatially distributed clusters of a predefined size. These voxel elements are assumed to have some relationship with the stimulus and form the basis for computations at increment $i + 1$. This process is continued until convergence is reached with voxel subset $\mathcal{S} \subset S$, yielding scalar vector $\boldsymbol{\rho}_{\mathcal{S}}$ and vector of weight functions $\boldsymbol{\eta}_{\mathcal{S}}(t)$ and $\boldsymbol{\xi}_{\mathcal{S}}(t)$.

2.5 Brain State Prediction

We use $\boldsymbol{\rho}_{\mathcal{S}}$, $\boldsymbol{\eta}_{\mathcal{S}}(t)$ and $\boldsymbol{\xi}_{\mathcal{S}}(t)$ for prediction of brain states from new and spatially normalized fMRI data. The voxel time courses at spatial locations corresponding to those resulting from incremental exploration are extracted from this fMRI data and functionalized into $\tilde{\mathbf{f}}(t) = [\tilde{f}_1(t), ..., \tilde{f}_{\mathcal{S}}(t)]^T$. Then, following [8], prediction reduces to

$$\tilde{\mathbf{g}}(t) = \tilde{\mathbf{F}}\boldsymbol{\xi}_{\mathcal{S}}(t) \tag{6}$$

where the $\mathcal{S} \times Q$ canonical correlation loadings matrix $\tilde{\mathbf{F}}$ has elements

$$\tilde{f}_{sq} = \rho_q \int \tilde{f}_s(t)\eta_q(t) \tag{7}$$

and $\tilde{\mathbf{g}}(t)$ is the vector of predicted stimuli. We define the mean of $\tilde{\mathbf{g}}(t)$ as the stimulus that gave rise to the brain response represented by $\tilde{\mathbf{f}}(t) = [\tilde{f}_1(t), ..., \tilde{f}_{\mathcal{S}}(t)]^T$. Hence, we have brain locations that are likely involved in processing the external stimulus as well as characterizations of the relationship between activity at these areas and the stimulus.

3 Experiments and Results

3.1 Experiment

Evaluation of our method is done on a data subset from the brain activity interpretation competition [9,10], involving fMRI scans of three different subjects and two sessions. In each session, a subject viewed a new Home Improvement sitcom movie for approximately 20 minutes. All three subjects watched the same two movies. The scans produced volumes with approximately 35.000 brain voxels, each approximately 3.28mm by 3.28mm by 3.5mm, with one volume produced every 1.75 seconds. These scans were preprocessed (motion correction, slice time correction, linear trend removal) and spatially normalized to the Montreal Neurological Institute brain atlas.

After fMRI scanning, the three subjects watched the movie again to rate 30 movie features at time intervals corresponding to the fMRI scan rate. In our experiments, we focused on the 13 core movie features: Amusement, Attention, Arousal, Body Parts, Environmental Sounds, Faces, Food, Language, Laughter, Motion, Music, Sadness and Tools. The real-valued ratings were convolved with a standard hemodynamic response function, then subjected to voxel-wise non-linear registration as described in 2.1.

For training and testing our model, we divided each fMRI scan and each subject rating into 6 parts corresponding with movie on parts and functionalized these parts by fitting a 30 coefficient B-spline to their discrete data points. This resulted in 18 data sets for training (3 subjects × 6 movie parts) and another 18 for testing. We used movie 1 data for training and movie 2 data for prediction, and vice versa, with parameter values as in [4]. Functional cross correlation between manual feature rating functions and the automatically predicted feature functions was used as an evaluation measure.

3.2 Results

Average cross correlation results of 2 × 18 cross validations for all 13 movie features are shown in figure 1a. Also shown are previous results based on principal component regression. As can be seen, canonical correlation analysis produces higher cross correlation values for all features except for feature ”Motion”. Four features exceed the 0.5 threshold, indicating that there is a significant degree of match between the subject ratings and the predicted ratings. The average cross correlation across features for canonical correlation analysis is 3.7, against 3.2 for principal component regression. In almost all cross validation predictions, the number of voxels used were significantly smaller for canonical analysis than for principal component analysis.

The highest average cross correlation value of 6.9 is obtained for feature ”Faces”, with the best single result of 7.8 for prediction of subject 3 watching part 2 of movie 1. For this feature, first level analysis of each of the 18 training data sets associated with movie 2 produced a total number of 480 predictive voxels. In the second level analysis, these voxels were analyzed again to arrive at a reduced data set of 104 voxels for performing canonical correlation analysis and determining weight functions. Figure 1 shows gray level image with color overlay and surface rendering of a subset of the 104 voxels from second level analysis. The cross hair shows the voxel location in the occipital lobe that was found to be predictive across most subjects and movie parts.

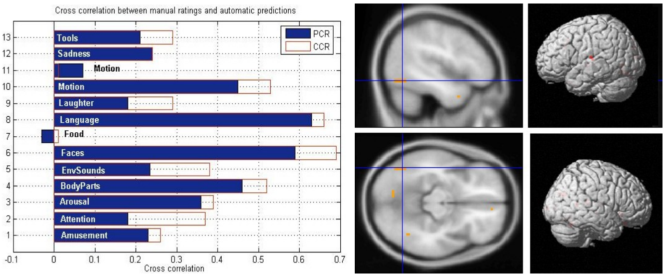


Fig. 1. Left: cross correlation values from cross-validation for 13 core movie features, using principal component analysis (PCR) and canonical corelation analysis (CCR) Right: gray level image with color overlay and surface rendering of a subset of predictive voxels from second level analysis. Color denotes predictive power and cross hair shows most predictive location.

4 Conclusion

We have proposed an incremental functional canonical correlation analysis method for prediction of brain states from fMRI. In comparison with the principal component regression method in [4], the proposed method produces better prediction results using a smaller amount of spatially distributed brain voxel clusters. We conclude that functional canonical correlation analysis captures important modes of correlation between fMRI and stimuli data that are very suited for prediction of stimuli based on new fMRI data. Given the high prediction results, we emphasize that our method is very promising for identifying and characterizing complex brain responses to intricate external stimuli.

References

1. Haynes, J., Rees, G.: Decoding mental states from brain activity in humans. Nature Neuroscience 7(8) (2006)
2. Kamitani, Y., Tong, F.: Decoding the visual and subjective contents of the human brain. Nature Neuroscience 8 (2005)
3. Polyn, S., Natu, V., Cohen, J., Norman, K.: Category-Specific Cortical Activity Precedes Retrieval During Memory Search. Science 310 (2005)
4. Ghebreab, S., Smeulders, A., Adriaans, P.: Predicting mental states from fmri data: Incremental multivariate functional regression (Submitted, 2007)
5. Ramsay, J., Silverman, B.: Functional data analysis (1997)
6. Stewart, D., Love, W.: A general canonical correlation index. Psychological Bulletin 70 (1968)
7. Ghebreab, S., Jaffe, C., Smeulders, A.: Population-based incremental interactive concept learning for image retrieval by stochastic string segmentations. IEEE transactions Medical Imaging 23-6, 676–689 (2004)
8. Glahn, H.: Canonical correlation and its relationship to discriminant analysis and multiple regression. Journal of the Atmospheric Sciences 25 (1968)
9. Schneider, W., et al.: Competition: Inferring experience based cognition from fmri. In: Proceedings Organization of Human Brain Mapping Florence Italy, June 15, 2006 (2006)
10. Editorial: What's on your mind. Nature Neuroscience 7(8) (2006)

Part VIII

Protocols and Guidelines

Formalizing 'Living Guidelines' Using LASSIE: A Multi-step Information Extraction Method

Katharina Kaiser[1] and Silvia Miksch[1,2]

[1] Institute of Software Technology & Interactive Systems
Vienna University of Technology, Vienna, Austria
[2] Department of Information and Knowledge Engineering
Danube University Krems, Krems, Austria
{kaiser,silvia}@ifs.tuwien.ac.at
http://ieg.ifs.tuwien.ac.at

Abstract. Living guidelines are documents presenting up-to-date and state-of-the-art knowledge to practitioners. To have guidelines implemented by computer-support they firstly have to be formalized in a computer-interpretable form. Due to the complexity of such formats the formalization process is challenging, but burdensome and time-consuming.

The LASSIE methodology supports this task by formalizing guidelines in several steps from the textual form to the guideline representation language Asbru using a document-centric approach. LASSIE uses Information Extraction technique to semi-automatically accomplish these steps.

We apply LASSIE to support the implementation of living guidelines. Based on a living guideline published by the Scottish Intercollegiate Guidelines Network (SIGN) we show that adaptations of previously formalized guidelines can be accomplished easily and fast. By using this new approach only new and changed text parts have to be modeled. Furthermore, models can be inherited from previously modeled guideline versions that were added by domain experts.

1 Introduction

The development process for a clinical practice guideline (CPG) takes at least two years. Thus, CPGs can be out of date as soon as they are produced, as new research findings are continuously published. To overcome this problem sometimes the shelf life for a guideline is identified; either by a date (e.g., this guideline will be reviewed in 2 years) or by a statement that the review date will be determined by the availability of new evidence (e.g., this guideline will be considered for review as new evidence becomes available). Alternatively, we can consider a new option – the living guideline. A living guideline is one that remains under review on an ongoing basis, with updates published at set intervals (e.g., annually).

The review of the guideline (i.e., a new article in the specified field is available) may have various characteristics. On the one hand it can add additional evidence and thus alter the evidence level of a recommendation. On the other hand it can lead to a new recommendation or it may change an existing one. However, in the majority of cases only small text parts are changed; often only the reference to the new article is added or to an obsolete article is removed.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 401–410, 2007.

Modeling CPGs in a computer-interpretable form is a prerequisite for various computer applications to support their application. However, transforming guidelines in a formal guideline representation is a difficult task. In [1] and [2] we have proposed a semi-automatic methodology called LASSIE to model treatment processes in multiple steps using Information Extraction (IE).

We will now show that we can use LASSIE to support the formalization of living guidelines. Applying this method, which traces both the general formalization steps and the changes to new versions has the potential to reduce the modeling effort. The Scottish Intercollegiate Guidelines Network (SIGN) has already published a living guideline [3]. Based on the documents provided we will show that adaptations of formalized guidelines can be accomplished easily and fast.

In the next section we will discuss some work on guideline formalization tools and guideline versioning methods. Afterwards we will give a short introduction in LASSIE. In Section 4 we describe the adaptation of LASSIE for supporting living guidelines followed by a case study illustrating our methodology. Section 6 summarizes our work and represents our conclusions.

2 Related Work

In this section, we present relevant work describing guideline formalization tools and approaches for guideline versioning.

For formalizing clinical guidelines into a guideline representation language (see [4] for an overview and comparison) various tools exist. We can classify such tools in document-centric and model-centric tools.

2.1 Document-Centric Approaches

Markup-based tools utilize a document-centric approach. Thereby, the original guideline document is systematically marked-up by the user in order to generate a semi-formal model of the marked text part.

The *GEM Cutter* [5] was one of the first exponents of this apporach transforming guideline information into the GEM format [6]. *Stepper* [7] is a tool that formalizes the initial text in multiple user-definable steps corresponding to interactive XML transformations. The *Document Exploration and Linking Tool / Addons (DELT/A)* [8] supports the translation of HTML documents into any XML language. It uses links between the text part in the original document and its corresponding XML model. To generate a specific model user-definable *macros* can be used. *Uruz*, part of the *Digital electronic Guideline Library (Degel)* framework [9], is a web-based markup tool that supports indexing and markup using any hierarchical guideline-representation format. It enables the user to embed in the guideline document terms originating from standard vocabularies.

2.2 Model-Centric Approaches

In model-centric approaches a conceptual model is formulated by domain experts. The relationship between the model and the original document is only indirect.

AsbruView [10] uses graphical metaphors to represent Asbru plans. *AREZZO* and *TALLIS* [11] support the translation into PROforma using graphical symbols representing the task types of the language. *Protégé* [12] is a knowledge-acquisition tool that supports the translation into guideline representation languages EON, GLIF, and PROforma. It uses specific ontologies for these languages, whereas parts of the formalization process can be accomplished with predefined graphical symbols. AREZZO, TALLIS, and Protégé offer a flowchart-based representation of the processes.

2.3 Guideline Versioning

Unfortunately, guideline versioning has not been adequately addressed by now. There are two approaches dealing with versioning:

Peleg and Kantor [13] propose a model-centric approach for GLIF. Thereby, the underlying GLIF ontology is extended by version information and a versioning tool was developed that supports the creation of a new CPG model or the modification of an existing one as well as the displaying of versions of a CPG model, highlighting the differences.

Seyfang et al. [14] describe the formalization of 'living guidelines' using a document-centric approach. They start with an HTML version of the guideline and use different intermediate representations to derive a formal model of the guideline. The first intermediate representation is MHB and the DELT/A tool is used to mark-up text chunks. The original marked-up guideline document is then manually updated to the new version by highlighting both newly added and removed text fragments. Using the DELT/A tool the highlighted text fragments are selected to visualize the corresponding MHB chunk in order to make the necessary changes.

But still, using the mentioned tools the modeling process is complex and labor intensive. Methods are needed to automate parts of the modeling task.

3 LASSIE – Modeling Treatment Processes Using Information Extraction

Most guideline representation languages are very powerful and thus very complex. They can present a multitude of different information and data. We apply a multi-step transformation process that facilitates the formalization process by various intermediate representations (IRs) obtained in stepwise procedures.

Our multi-step transformation methodology, called LASSIE[1], supports the document-centric approach by marking the original guideline document and generating the particular models for each marked text part. It is intended to be a semi-automatic approach. This enables the user not only to correct the transformations, but also to augment them by implicit knowledge necessary for a subsequent execution. After each step the user is able to view the results using the DELT/A tool [8].

The benefits of the multi-step approach and in the following of the IRs are that IRs (1) support a concise formalization process, (2) provide different formats and separate

[1] Modeling treAtment proceSSes using Information Extraction.

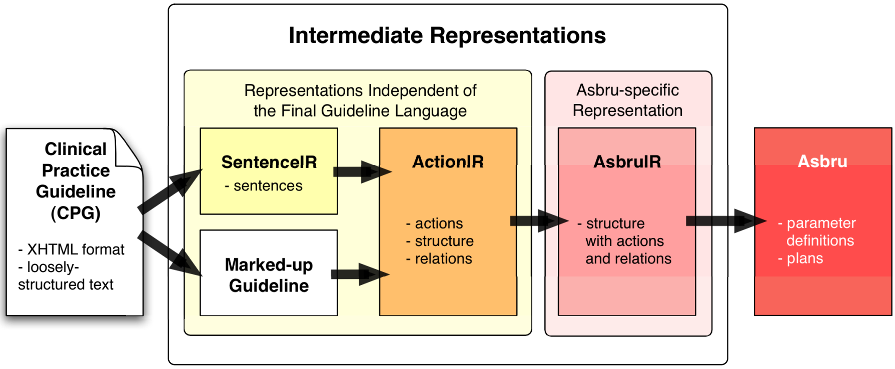

Intermediate Representations
Representations Independent of the Final Guideline Language
Asbru-specific Representation
Clinical Practice Guideline (CPG)
- XHTML format
- loosely-structured text
SentenceIR
- sentences
Marked-up Guideline
ActionIR
- actions
- structure
- relations
AsbruIR
- structure with actions and relations
Asbru
- parameter definitions
- plans

Fig. 1. Steps to (semi-)automatically gain an Asbru representation of CPGs. To gain process information from a CPG the first two steps are accomplished in order to have a representation independent of the final guideline language.

views and procedures for various kinds of information, (3) specific heuristics for each particular kind of information can be applied, and (4) a simpler and more concise evaluation and tracing of each process step is accomplishable. The IRs are specific templates used by IE methods to present the desired information. The IE methods use a terminology based on the Medical Subject Headings (MeSH)[2] [15] and manually generated extraction patterns.

CPGs present effective treatment processes. One challenge when authoring CPGs is the detection of individual processes and their relations and dependencies. We can generate simple representations of treatment instructions (i.e., actions), which are independent from the final guideline representation language. Based on this independent representation we can transform the information in further steps into the guideline languages. In [1] and [2] we have demonstrated that it is possible to formalize processes using IE for modeling guidelines in Asbru (see Fig. 1).

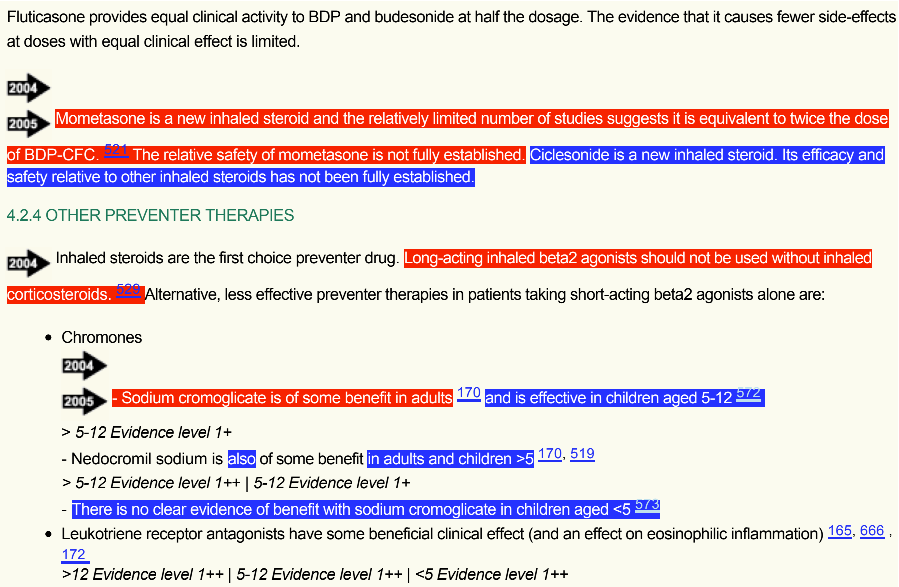
4 Adaptation of LASSIE for ’Living Guidelines’

Using LASSIE a unique identifier (i.e., the DELT/A link) marks information transformed from one step to the next. We now apply LASSIE to support the formalization of living guidelines. The document provide us the information that has changed: Adaptations of every new revision are marked by arrows and highlighted in terms of color (or in different gray scales) (see Fig. 2).

We now propose a new method utilizing this information. Thereby, the new guideline is not going to be modeled from scratch, but already modeled parts from previous versions are inherited. Thus, only new text parts have to be modeled (see Fig. 3).

As LASSIE is a multi-step methodology, we have to satisfy each step for the living guideline’s formalization.

[2] http://www.nlm.nih.gov/mesh/



Fluticasone provides equal clinical activity to BDP and budesonide at half the dosage. The evidence that it causes fewer side-effects at doses with equal clinical effect is limited.

2004

2005 Mometasone is a new inhaled steroid and the relatively limited number of studies suggests it is equivalent to twice the dose of BDP-CFC. [521] The relative safety of mometasone is not fully established. Ciclesonide is a new inhaled steroid. Its efficacy and safety relative to other inhaled steroids has not been fully established.

4.2.4 OTHER PREVENTER THERAPIES

2004 Inhaled steroids are the first choice preventer drug. Long-acting inhaled beta2 agonists should not be used without inhaled corticosteroids. [529] Alternative, less effective preventer therapies in patients taking short-acting beta2 agonists alone are:

- Chromones

  2004

  2005 - Sodium cromoglicate is of some benefit in adults [170] and is effective in children aged 5-12 [572]

  > 5-12 Evidence level 1+

  - Nedocromil sodium is also of some benefit in adults and children >5 [170], [519]

  > 5-12 Evidence level 1++ | 5-12 Evidence level 1+

  - There is no clear evidence of benefit with sodium cromoglicate in children aged <5 [573]
- Leukotriene receptor antagonists have some beneficial clinical effect (and an effect on eosinophilic inflammation) [165], [666], [172]

  >12 Evidence level 1++ | 5-12 Evidence level 1++ | <5 Evidence level 1++

Fig. 2. Excerpt of the 2005 version of the ”living guideline” [3]. Adaptations of every new revision are marked by arrows and highlighted in terms of color (or in different gray-scales).

4.1 Pre-processing

As the input of LASSIE’s first step is the XHTML-conform guideline document, we have to preprocess the document to get a unified document format. We accomplish this by XSLT scripts, HTML Tidy[3], and manual post-processing in order to obtain not only a well-formed but also a hierarchically well-structured XHTML document.

4.2 Marking-Up the New Guideline Version

LASSIE’s first step is to detect relevant sentences and text parts in the guideline document. Text parts are thereby list entries that may not be complete sentences, but are referred to as sentences in the remaining paper.

The output of LASSIE’s first step are two files: (1) the marked-up guideline document, where relevant sentences are marked and tagged by a DELT/A link, and (2) a file containing all relevant sentences and their corresponding DELT/A links.

We use these files of the previous guideline version to detect unchanged relevant sentences in the new guideline version. We parse the new guideline document and search for each sentence marked-up in the previous version. Thereby, we have to consider not only equal sentences but also equal contexts of them. This is necessary as a marked-up sentence can appear repeatedly in the document and we have to assign the correct DELT/A link in the new document. For each sentence in the new guideline that

[3] http://tidy.sourceforge.net

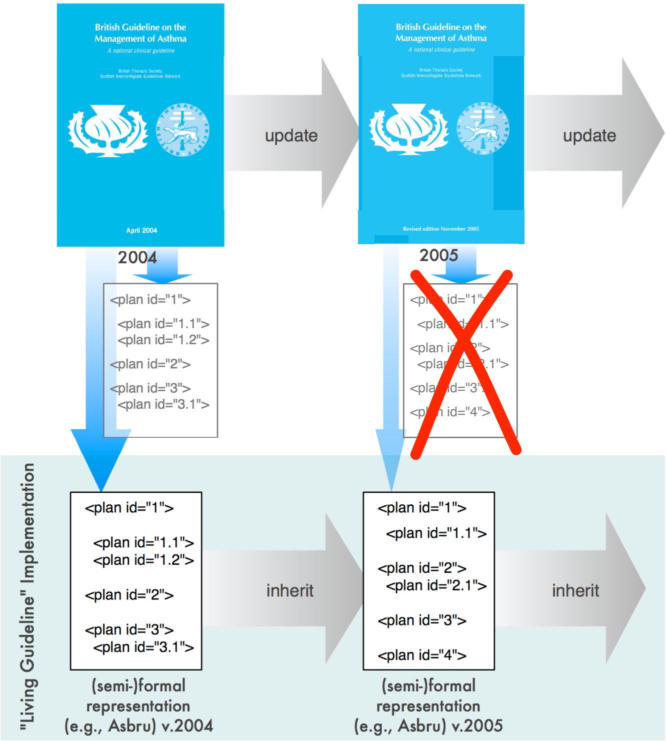

British Guideline on the Management of Asthma
A national clinical guideline
April 2004
British Guideline on the Management of Asthma
A national clinical guideline
Revised edition November 2005
update
update
2004
2005
<plan id="1">
<plan id="1.1">
<plan id="1.2">
<plan id="2">
<plan id="3">
<plan id="3.1">
"Living Guideline" Implementation
<plan id="1">
<plan id="1.1">
<plan id="1.2">
<plan id="2">
<plan id="3">
<plan id="3.1">
inherit
<plan id="1">
<plan id="1.1">
<plan id="2">
<plan id="2.1">
<plan id="3">
<plan id="4">
inherit
(semi-)formal representation (e.g., Asbru) v.2004
(semi-)formal representation (e.g., Asbru) v.2005

Fig. 3. Formalizing a living guideline using LASSIE. The documents provide us the information that has changed. After comparing the new documents with the previous ones we are able to adapt the former formalized documents using LASSIE.

is marked as updated as a part or whole we apply step 1 of the LASSIE methodology (see [2] for details) in order to detect relevant sentences for further processing. Relevant sentences of the old guideline version that are not found in the new document can be seen as removed.

For each sentence of the new guideline that has been marked as relevant by the LASSIE methodology we assign also a version id. Furthermore, we have to be aware to not assign an obsolete DELT/A link to a new sentence.

Thus, we obtain the new marked-up guideline version and are now able to extract the processes in order to gain a representation independent of the final guideline language. After this step the user is also able to view the resulting files with the DELT/A tool and make corrections.

4.3 Further Transformation of the Extracted Information

After obtaining the new marked-up guideline document we can proceed with the subsequent steps coming up with LASSIE. That means, we can inherit models of subsequent representations that correspond with text parts that were not changed in the new guideline version. For new or changed models the particular processing step of LASSIE is applied. For instance, to detect processes we proceed as following:

Within the next step of the LASSIE methodology relevant sentences are structured and relationships between sentences are found. The output of this step is a representation (*ActionIR*) containing actions, relations controlling the process flow between these actions, and the structure illustrating the hierarchy and nesting of groups of actions.

An *action* contains the action sentence, possible assigned annotation sentences, treatment instruments, information about the dosage, duration or iteration of a drug administration, and conditions. If the action is part of a selection, it is given a selection id. DELT/A links are inherited from the *SentenceIR* representation in order to provide the traceability of the process.

In order to obtain actions from our new version of the marked-up guideline, we can inherit action and annotation sentences from the previous *ActionIR* version. Furthermore, new relevant sentences of the current guideline version are classified in action and annotation sentences. When sentences are classified as annotation they must be assigned an *action* sentence. If an action and its assigned annotation sentences were not changed in the new version, the complete action node is inherited to the new *ActionIR* representation. Otherwise, the action node and its additional information has to be generated by LASSIE. Additionally, a version id is assigned for these new nodes. Likewise, we are able to inherit relations between actions nodes if none of the both action nodes has changed. Otherwise, we have to detect new relations using LASSIE. The third part of the *ActionIR* representation, the structure of the actions, is then generated by LASSIE.

The output of this step is then a new version of the *ActionIR* representation, which can be viewed with the DELT/A tool. Changed information is identifiable by the version id. The user may then make corrections or add new information to the representation.

5 Case Study

We tested the applicability of our method to a real *living guideline*. Based on the *British guideline on the management of asthma* [3] from SIGN in its version of 2005 we generated the previous guideline versions (i.e., from 2004) due to the non-availability of the old documents[4]. This was possible because SIGN offers a document which clearly describes every adaptation (i.e., change, adding, removal) of the text. For evaluating the method we only used Section 4 (*Pharmacological Management*) of the guideline. It describes an important part of the asthma treatment and contains also updated text parts.

5.1 Formalizing the Original Guideline Version

We preprocessed the old guideline document to comply our unified document format. Starting with the old guideline document we used LASSIE to generate the particular

[4] We were not able to receive the older guideline versions from SIGN.

models necessary for formalization in Asbru. We automatically generated the intermediate representations and adapted them according to our needs. The document consists of 509 sentences. 139 of them were classified as relevant for further processing.

5.2 Formalizing the New Guideline Version

The next step was to model the new guideline version using our new method. Therefore, we prototypically implemented our method to automate this task and adapted our implementation of LASSIE to enable the processing of *living guidelines*.

Preprocessing. We preprocessed the new guideline document in order to gain a unified document format complying the XHTML format.

Markup of new guideline version. Afterwards, we automatically searched for unchanged sentences that were marked in the previous guideline version and added the corresponding DELT/A links into the current document. Now, we were able to have LASSIE check the adapted sentences for relevancy. The new version of Section 4 consists of 515 sentences. We were able to inherit 133 sentences of the old version, which means that six relevant sentences were either changed or removed in the new version. 13 updated or new sentences were found and checked with LASSIE, which classified ten as relevant. The new relevant sentences were marked and assigned a new DELT/A link as well as a version id.

Action generation and further transformations. Within the next step the new sentences were classified in action or annotation sentences. The latter are then assigned to action sentences. We received five action sentences and five annotation sentences. Four of the annotation sentences were assigned to two previously available action sentences; one to a new action sentence. Thus, the remaining unchanged action models were inherited from the previous version.

The same procedure is done for all subsequent steps in an analogous manner.

5.3 Discussion

Our study shows that using a document-centric approach – LASSIE with the DELT/A tool – offers distinct benefits in modeling *living guidelines*. A fast adaptation of the new document is possible. As in *living guidelines* there will not be radical changes from one version to the succeeding version, inheriting of previous models is a simple, time-saving, but effective method for modeling computer-supported guidelines. Also, in the intermediate representations the new models are marked by their version ids to enable a prompt identification. Thus, the user is able to perform adaptations quickly and conveniently.

A limitation of our methodology is that minor changes in the text may result in applying a new relevance check, sentence classification, action generation, and so on, which will require an evaluation by a human afterwards. In methods described in Section 2.3 such minor changes may be checked and accomplished by a human user more efficiently.

Furthermore, we have to mention that the IRs do not contain the models of all versions, only the actual ones. Thus, it is not possible to have one file for all versions, but one file for each version of a representation.

6 Conclusion

Living guidelines are documents presenting up-to-date and state-of-the-art knowledge to practitioners. To support their application they have to be brought in a computer-interpretable form, which is a difficult task.

We propose a method applicable on documents previously being formalized using a document-centric approach. Thereby, the guideline document is marked-up and corresponding formal models are generated. Our method utilizes these links between the textual document and the formal models. It inherits formalized models of the previous guideline version by re-linking them to their corresponding text parts in the new guideline version. Only changed or added texts have to be analyzed and modeled. The formalization task is thereby done using the LASSIE methodology. It is a semi-automatic approach using IE and various intermediate representations to model different kinds of information in various granularities. Our case study showed that the modelling effort can be reduced considerably by applying our LASSIE methodology.

By re-using previously formalized models of guidelines we are able to quickly and effectively formalize new guideline versions.

Acknowledgements. This work is supported by "Fonds zur Förderung der wissenschaftlichen Forschung FWF" (Austrian Science Fund), grant L290-N04.

References

1. Kaiser, K., Akkaya, C., Miksch, S.: How can information extraction ease formalizing treatment processes in clinical practice guidelines? A method and its evaluation. Artificial Intelligence in Medicine 39(2), 151–163 (2007)
2. Kaiser, K., Miksch, S.: Modeling treatment processes using information extraction. In: Yoshida, H., Jain, A., Ichalkaranje, A., Jain, L.C., Ichalkaranje, N. (eds.) Advanced Computational Intelligence Paradigms in Healthcare – 1. Studies in Computational Intelligence (SCI), vol. 48, pp. 189–224. Springer, Heidelberg (2007)
3. Scottish Intercollegiate Guidelines Network (SIGN), British Thoracic Society: British guideline on the management of asthma. a clinical national guideline. Scottish Intercollegiate Guidelines Network (SIGN) (2005)
4. Peleg, M., Tu, S.W., 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. Journal of the American Medical Informatics Association (JAMIA) 10(1), 52–68 (2003)
5. Polvani, K.A., Agrawal, A., Karras, B., Deshpande, A., Shiffman, R.: GEM Cutter Manual. Yale Center for Medical Informatics, New Haven, CT (2000)
6. 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 (JAMIA) 7(5), 488–498 (2000)

7. Ružička, M., Svátek, V.: Mark-up based analysis of narrative guidelines with the Stepper tool. In: Kaiser, K., Miksch, S., Tu, S.W. (eds.) Computer-based Support for Clinical Guidelines and Protocols. Proceedings of the Symposium on Computerized Guidelines and Protocols (CGP 2004), Amsterdam, NL, vol. 101, pp. 132–136. IOS Press, Amsterdam (2004)
8. Votruba, P., Miksch, S., Kosara, R.: Facilitating knowledge maintenance of clinical guidelines and protocols. In: Fieschi, M., Coiera, E., Li, Y.C.J. (eds.) Proceedings from the Medinfo 2004 World Congress on Medical Informatics, AMIA, pp. 57–61. IOS Press, Amsterdam (2004)
9. Shahar, Y., Young, O., Shalom, E., Mayaffit, A., Moskovitch, R., Hessing, A., Galperin, M.: DEGEL: A hybrid, multiple-ontology framework for specification and retrieval of clinical guidelines. In: Dojat, M., Keravnou, E.T., Barahona, P. (eds.) AIME 2003. LNCS (LNAI), vol. 2780, pp. 122–131. Springer, Heidelberg (2003)
10. Kosara, R., Miksch, S.: Metaphors of Movement: A Visualization and User Interface for Time-Oriented, Skeletal Plans.. Artificial Intelligence in Medicine, Special Issue: Information Visualization in Medicine 22(2), 111–131 (2001)
11. Steele, R., Fox, J.: Tallis PROforma Primer – Introduction to PROforma Language and Software with Worked Examples. Technical report, Advanced Computation Laboratory, Cancer Research, London, UK (2002)
12. Gennari, J.H., Musen, M.A., Fergerson, R.W., Grosso, W.E., Crubézy, M., Eriksson, H., Noy, N.F., Tu, S.W.: The Evolution of Protégé: An Environment for Knowledge-based Systems Development. International Journal of Human Computer Studies 58(1), 89–123 (2003)
13. Peleg, M., Kantor, R.: Approaches for guideline versioning using GLIF. In: Musen, M.A. (ed.) Proceedings of the 2003 American Medical Informatics Association (AMIA) Annual Symposium, Washington, DC, pp. 509–513. American Medical Informatics Association (2003)
14. Seyfang, A., Martinez-Salvador, B., Serban, R., Wittenberg, J., Miksch, S., Marcos, M., ten Teije, A., Rosenbrand, K.: Maintaining formal models of living guidelines efficiently. In: Bellazzi, R., Abu-Hanna, A., Hunter, J. (eds.) Proc. of the 11th Conference on Artificial Intelligence in Medicine (AIME'07), Springer, Heidelberg (2007)
15. National Library of Medicine: Medical Subject Headings. The Library (updated annually)

The Role of Model Checking in Critiquing Based on Clinical Guidelines*

Perry Groot[1], Arjen Hommersom[1], Peter Lucas[1],
Radu Serban[2], Annette ten Teije[2], and Frank van Harmelen[2]

[1] Radboud Universiteit Nijmegen
[2] Vrije Universiteit Amsterdam

Abstract. Medical critiquing systems criticise clinical actions performed by a physician. In order to provide useful feedback, an important task is to find differences between the actual actions and a set of 'ideal' actions as described by a clinical guideline. In case differences exist, insight to which extent they are compatible is provided by the critiquing system. We propose a methodology for such critiquing, where the ideal actions are given by a formal model of a clinical guideline, and where the actual actions are derived from real world patient data. We employ model checking to investigate whether a part of the actual treatment is consistent with the guideline. Furthermore, it is shown how critiquing can be cast in terms of temporal logic, and what can be achieved by using model checking. The methodology has been applied to a clinical guideline of breast cancer in conjunction with breast cancer patient data.

1 Introduction

There is an increasing interest amongst researchers to develop computerised versions of clinical guidelines, which at the moment are still just documents, using one of the specialised guideline representation languages. The resulting computer-based guidelines can then act as a basis for the development of decision-support systems, which, thus, allow computer-based deployment of guidelines in a clinical setting. One possible application of such clinical decision-support systems is *critiquing*, i.e., to spot and analyse differences between the proposed actions taken by a medical doctor, and a set of 'ideal' actions as prescribed by the computerised guideline. As a computer-based clinical guideline is represented in a formal language, there is, also room for a formal underpinning of the various ways a guideline can be manipulated.

A natural way to formally describe the actions taken by a medical doctor in the management of the disease of a patient is offered by temporal logics. As a family of languages, logics make it possible to describe the meaning of the

* This work has been supported by the European Commission's IST program, under contract number IST-FP6-508794 Protocure-II. We acknowledge the help received from medical experts from the Dutch Institute for Healthcare Improvement, Utrecht, The Netherlands and from the Fundacio 'Josep Laporte', Barcelona, Spain.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 411–420, 2007.

various aspects of the disease and condition of the patient in a precise fashion. By the addition of temporal operators, *temporal* logic adds various notions of progress of the disease and sequencing of actions in time.

One way to look upon a patient and a patient's disease logically is as a concurrent system, i.e., as a system described in terms of states and state transitions in time. Model checking technology offers methods that allow one to analyse concurrent systems for their consistency. One can rely on an extensive collection of tools and techniques readily available. It is a well investigated technique for verification of systems that can be modelled by a finite transition system. However, model checking has been mainly applied to technical systems, such as hardware, software-based communication protocols, concurrent programs, etc. This raises the question when adopting this global view on the representation of diseases, patient conditions, and disease management actions, whether model checking can be used as a basis for critiquing. It is this question that is being explored in detail in this paper.

Model checking takes domain knowledge, called a system description, and sequences of actions as input. In this case, a formalised guideline is taken as a system description; the actions that have been performed on a specific patient are represented as a temporal formula. Model checking then involves investigating the consistency of the formalised guideline and actual treatment. The exploration of the use of model checking in the analysis of medical knowledge (guidelines and patient data) with the purpose of critiquing, is the innovative part of this work.

2 Approach

The common feature of a critiquing system is that the user of the system provides as input (1) a problem description (e.g., patient symptoms), and (2) a proposed solution (e.g., a treatment plan). This second input is what distinguishes critiquing systems from the more traditional expert systems, which only take a problem description as input [10,4]. The second input to a critiquing system, i.e., a proposed solution, is typically the output of an expert system.

In our approach of critiquing medical treatment plans using model checking, the input to the system consists of patient data and a treatment plan (cf. Fig. 1). Patient data consists of patient symptoms and test outcomes measured for the patient, whereas the treatment plan consists of all actions (to be) performed by the practitioner. As the critiquing process is difficult to accept by practitioners when they are continually interrupted to provide input to the system, both patient data and treatment plan are typically provided by electronic records. We will assume that these are given to the system as temporal logic formulas.

The critiquing system uses the patient data and treatment plan as specifications that need to be checked against a formal model of the guideline, i.e., a state transition system. When the specifications are consistent with the guideline model, no critique needs to be generated as the proposed treatment plan conforms with the guideline. In case an inconsistency is found between the specification and the guideline model, the specification is weakened to get insight to

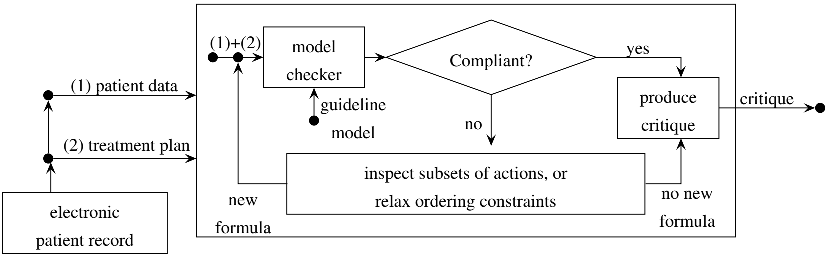


Fig. 1. Critiquing approach using model checking. Given patient data and a treatment plan as input (temporal specifications), the critiquing system uses a model checker to verify consistency w.r.t. to a guideline model (state transition system) to generate a critique (empty in case of compliance).

which extent the treatment plan is consistent with the guideline. There are two possible reasons for the incompatibility:

Non-compliant order: It is possible that each of the actions in the treatment plan can be applied to this patient, but only in a different order than the treatment plan proposes. This can be established by removing the order between some of the actions in the treatment plan.

Non-compliant actions: Another possibility is that, according to the guideline, some of the actions cannot be prescribed at all for the patient in question. This can be investigated by considering a subset of the actions in the treatment plan.

The approaches can be combined and lead to further insight into the nature of the detected inconsistency allowing the system to exploit these insights into a critique, which is then given to the practitioner.

3 Temporal Logic for Critiquing

In Subsection 3.1, the formal preliminaries of temporal logic are introduced. In Subsection 3.2, temporal logic is applied to critiquing and examples are provided.

3.1 Preliminaries

Temporal logic is a modal logic, where relationships between worlds in the usual possible-world semantics of modal logic is understood as time order. The logic that we use here for specifying properties of medical guidelines is a combination of Computation Tree Logic (CTL) and Linear Temporal Logic (LTL) [3].

In this paper we model a guideline as a Kripke structure M over a set of atomic propositions AP, which formally is defined as a four tuple $M = (S, S_0, R, L)$ where S is a finite set of states, $S_0 \subseteq S$ is the set of initial states, $R \subseteq S \times S$ is a total transition relation, and $L : S \rightarrow 2^{AP}$ is a function that labels each state with the set of atomic propositions true in that state. A *path* in the model M from a state s is an infinite sequence $\pi = s_0 s_1 s_2 \ldots$ such that $s_0 = s$ and $R(s_i, s_{i+1})$ holds for all $i \geq 0$. With π^i we denote the suffix of π starting at s_i, i.e., $\pi^i = s_i s_{i+1} s_{i+2} \ldots$.

CTL uses atomic propositions, propositional connectives, *path quantifiers* and *temporal operators* for describing properties of *computation trees*, i.e., the tree that is formed by designating a state in the Kripke structure as the initial state and then unwinding the structure into an infinite tree according to the transition relation R with the initial state as root. This leads to two types of formulas: *state formulas*, which are true in a specific state, and *path formulas*, which are true along a specific path. A path formula is build up by applying one of the temporal operators to one or two state formulas. In this paper, the temporal operators used are $\mathbf{X}$, $\mathbf{G}$, $\mathbf{F}$, and $\mathbf{U}$. With $\mathbf{X}\varphi$ being true if φ holds in the next state, $\mathbf{G}\varphi$ if φ holds in the current state and all future states, $\mathbf{F}\varphi$ if φ holds in the current state or some state in the future, and $\varphi\mathbf{U}\psi$ if φ holds until eventually ψ holds. A state formula can be built inductively from atomic propositions, propositional connectives, and if f and g are path formulas, then $\mathbf{E}f$ and $\mathbf{A}f$ are state formulas. The path quantifiers $\mathbf{A}$ and $\mathbf{E}$ are used to specify that all or some of the paths starting at a specific state have some property.

The semantics of CTL is defined with respect to a Kripke structure M. Given a state formula f, the notation $M, s \models f$ denotes that f holds in state s of the Kripke structure M. Assuming that f_1 and f_2 are state formulas and g_1 and g_2 are path formulas, the relation $\models$ is defined inductively as shown in Fig. 2. The remaining syntax consisting of $\vee, \rightarrow, \mathbf{G}, \mathbf{A}$ can be defined as usual, i.e., $f_1 \vee f_2 \equiv \neg(\neg f_1 \wedge \neg f_2)$, $f_1 \rightarrow f_2 \equiv \neg f_1 \vee f_2$, $\mathbf{G}g \equiv \neg\mathbf{F}\neg g$, and $\mathbf{A}f \equiv \neg\mathbf{E}\neg f$.

In contrast to CTL, LTL provides operators for describing events along a single computation path. Each formula is of the form $\mathbf{A}f$, with f being a path formula, which is either an atomic proposition or inductively defined as $\neg f$, $f \vee g$, $f \wedge g$, $\mathbf{X}f$, $\mathbf{F}f$, $\mathbf{G}f$, $f\mathbf{U}g$ with f, g path formulas. This language can be evaluated on Kripke structures presented in Fig. 2.

$$\begin{array}{ll}
M, s \models p & \Leftrightarrow p \in L(s) \\
M, s \models \neg f_1 & \Leftrightarrow M, s \not\models f_1 \\
M, s \models f_1 \wedge f_2 & \Leftrightarrow M, s \models f_1 \text{ and } M, s \models f_2 \\
M, s \models \mathbf{E}g_1 & \Leftrightarrow \text{there is a path } \pi \text{ from } s \text{ such that } M, \pi \models g_1 \\
M, \pi \models f_1 & \Leftrightarrow s \text{ is the first state of } \pi \text{ and } M, s \models f_1 \\
M, \pi \models \neg g_1 & \Leftrightarrow M, \pi \not\models g_1 \\
M, \pi \models g_1 \wedge g_2 & \Leftrightarrow M, \pi \models g_1 \text{ and } M, \pi \models g_2 \\
M, \pi \models \mathbf{X}g_1 & \Leftrightarrow M, \pi^1 \models g_1 \\
M, \pi \models \mathbf{F}g_1 & \Leftrightarrow \text{there exists a } k \geq 0 \text{ such that } M, \pi^k \models g_1 \\
M, \pi \models g_1\mathbf{U}g_2 & \Leftrightarrow \text{there exists a } k \geq 0 \text{ such that } M, \pi^k \models g_2 \text{ and} \\
 & \text{for all } 0 \leq j < k, M, \pi^j \models g_1
\end{array}$$

Fig. 2. Semantics of temporal logic with f_1 and f_2 representing state formulas and g_1 and g_2 representing path formulas

3.2 Critiquing Formulas

Each path in the state transition system can be considered a patient who is given a treatment that is consistent with the recommendation described by the guideline. Global properties of the guideline can be checked using LTL formulas or CTL formulas starting with **A**, for example, '**AF** radio-therapy', denotes that in each possible treatment, somewhere in the future radio-therapy is applied.

In the context of critiquing, CTL properties always start with an **E**, i.e., it is established that *some* treatment path exists in the guideline where the proposed treatment is described. For example, abstracting from the patient, a treatment given by a sequence of actions $\alpha_1, \alpha_2, \ldots$ can then be represented as:

$$\mathbf{EF}(\alpha_1 \wedge \mathbf{EX}\,\mathbf{EF}\,(\alpha_2 \wedge \mathbf{EX}\,\mathbf{EF}\,(\ldots))) \tag{1}$$

i.e., in some treatment α_1 is done, and after that α_2, etc. In general, CTL model checking is more efficient than LTL model checking, however, in case we do not know the order between the actions, a CTL formula consists of a disjunction of each possible order of actions and considers the existence of each order. In case of n actions, with all order unknown, this leads to formulas of size $O(n \times 2^n)$. Similarly, when global properties of the treatment path are introduced, for example the state of the patient or the fact that some action *never* occurs, such knowledge becomes difficult to express. Assume for example a global property described by β, then Formula (1) must be rephrased to the rather complicated formula:

$$\mathbf{E}(\beta\,\mathbf{U}\,(\alpha_1 \wedge \beta \wedge \mathbf{EX}\,\mathbf{E}(\beta\,\mathbf{U}\,(\alpha_2 \wedge \beta \wedge \mathbf{EX}\,\mathbf{E}(\ldots \wedge\ \mathbf{EG}\,\beta))))) \tag{2}$$

i.e., β holds until at some point α_1 and β (still) holds, after which β holds, etc.

Usually, the knowledge is reasonably complete and the global information is sparse, however, for a more succinct representation we can either use a more expressive logic such as CTL* [3] or consider LTL model checking. An approach using LTL is modular model checking [6], where the model is restricted using an LTL formula to those traces where the formula is valid. To prove the *existence* of a treatment in this approach, it is required to verify that the model restricted to a specification of a certain patient and treatment is not empty. Let φ be an LTL formula and $[\varphi]M\langle\bot\rangle$ denote that the set of LTL assertions φ leads to an empty model, i.e., φ describes a trace not present in the model. In contrast, if $[\varphi]M\langle\bot\rangle$ is shown to be false, then M can not be empty when restricted to φ proving that the trace described by φ exists in the model M. Formula (2) can thus be verified by showing that

$$[\mathbf{G}\beta \wedge \mathbf{F}(\alpha_1 \mathbf{XF}(\alpha_2 \wedge \ldots))]M\langle\bot\rangle \tag{3}$$

is *false*. An additional benefit of this presention is that when order information is absent, the property is typically more intuitivly specified. Nonetheless, when there are few actions involved and much of the order information is present, CTL formulas are expected to be more efficient to verify.

4 Application of the Methodology to Breast Cancer

4.1 Design and Choice of Case Studies

The clinical guideline used is the Dutch breast cancer guideline[1] and was represented as a state transition system in Cadence SMV using the techniques and representation described in [2]. The models used here were developed as part of the Protocure-II project.[2] Patient data were obtained from the Dutch Comprehensive Cancer Centre South (CCC), a registry in the Netherlands used for cancer research, planning of services, and evaluation and implementation of guidelines. The data collected concerns breast cancer patients treated in the period January 2003 - June 2004, when the guideline was applicable, and therefore suitable for compliance checks with the guideline. Each patient record consists of 269 variables, which includes information about the diagnosis and treatment.

The patient data from the registry could, in principle, directly be used for critiquing w.r.t. to the guideline. However, matching such data records to the terminology of the guideline is hard [7] and differs from the course commonly followed in medicine. In medical literature, specific patient cases, called *casuistics*, are frequenly discussed in detail to gain insight into the way the patient's disease was managed. These papers follow a long standing tradition and are seen as part of the 'education permanente' of the medical profession. Critiquing in this paper was therefore done casuistically by having the CCC patient data interpreted by medical experts who provided a direct mapping from the patient data in the registry to the guideline. Subsection 4.2 presents in more detail a case-study in critiquing using the casuistic interpretation of the CCC data.

A second case-study is presented in Subsection 4.3, which was obtained from the New South Wales Breast Cancer Institute, Australia.[3] These studies have been developed from the casuistic point of view to "allow clinicians, health professionals and members of the public to examine and understand some of the controversial and difficult aspects of breast cancer management". They are therefore more detailed than the patient data collected by the registry and are more suitable for an investigation of critiquing from a clinical point of view.

4.2 Case Study 1: Ductal Carcinoma in Situ

The steps of critiquing on one specific patient derived from the data, and subsequently interpreted by medical experts, is illustrated here. The diagnosis and treatment is summarised in Fig. 3. It can be said that this is a rather typical patient as it is a patient with one of the most frequent diagnoses in the data records. The following property describes the treatment sequence that our example patient has undergone. *"For a patient with diagnosis Ductal Carcinoma In Situ (DCIS), the following sequence of states is possible: the treatment starts, then axillary staging by sentinel node is activated, after which breast conserving*

[1] CBO: Richtlijn Behandeling van het mammacarcinoom, van Zuiden, 2002

[2] Breast cancer model can be obtained from `http://www.protocure.org`

[3] `http://www.bci.org.au/medical/caseindex.htm`

Medical condition: 79 years-old woman. Lesion of right breast: carcinoma in-situ with size between 1 and 2 cm. Two lymph nodes investigated and none positive.
Treatment: sentinel node biopsy + breast-conserving surgery without axillary clearance.

Fig. 3. Description of patient in conjunction with the prescribed treatment

therapy is activated". To specify and then verify that breast conserving therapy (denoted bct) can take place after axillary staging by sentinel node procedure (denoted asbSN), the following CTL formula is used:

$$\mathbf{EF}(\mathrm{DCIS} \wedge \mathbf{EX}\,\mathbf{EF}(\mathrm{asbSN} \wedge \mathbf{EX}\,\mathbf{EF}\,\mathrm{bct}))$$

A more strict formula could be obtained by assuming that the diagnosis DCIS, holds up to the moment of breast conserving therapy. However, this property stated above turns out to be false as it is, i.e., this treatment is non-compliant with respect to the guideline. In other words, according to the model of the guideline describing the treatment of DCIS, the sequence of actions performed by the doctor is incorrect for this patient. It could also be explained by the fact that, according to the model, at least one of the two actions in patient treatment should not be started, or they should be started in a different sequence. To identify this inconsistency, we reduce the actions that are being performed. If we reduce the sequence to only one action, then both actions are found possible, as shown by the following property (corresponding to the case when only 'bct' is activated as part of the DCIS treatment):

$$\mathbf{EF}(\mathrm{DCIS} \wedge \mathbf{EX}\,\mathbf{EF}\,\mathrm{bct})$$

The new conclusion is that under these circumstances the two actions cannot be activated in this sequence, or the ordering should be reversed.

In the experiment on the seven fairly prototypical patient-cases that can be found in the Dutch CCC data-set, some deviation was found between the guideline and each of the seven prototypical cases. Interestingly, for three of these, some differences could indeed be explained by looking at the new 2004 revision of the guideline.

4.3 Case Study 2: Infiltrating Ductal Carcinoma

For the second case study we have more elaborate information available. It concerns a patient who is a female with a lump in the 3 o'clock position of the right breast and a second lump just above this. No palpable axillary nodes or other abnormalities were found. The mammography revealed no focal mass, grouped microcalcifications, or anatomic distortion. Finally, the histopathology showed two lesions: both infiltrating duct carcinoma, 20mm in size, and with similar morphology. The sentinel nodes were mapped using lymphoscintigraphy and a biopsy was taken of a right axillary lymph node and an internal mammary node (the sentinel node procedure). In the right axillary lymph node, no malignancy was found. However, in the internal mammary node, metastatic carcinoma was identified. The treatment consisted of a total mastectomy of the right breast with immediate reconstruction. The axilla was treated by means of an axillary clearance and re-section of two further internal mammary nodes at higher levels (these were sampled partly because of the original pathology finding and partly because of ready access to the IMC).

The vocabulary of the guideline does not include the term 'infiltrating ductal carcinoma', but rather discusses 'operable invasive breast cancer' (OIBC). According to the guideline, operable invasive breast cancer is defined as T1-2 N0-1 M0, i.e., a tumour smaller than 5cm, with maximally one lymph node positive, and no distant metastasis. On basis of information provided by the diagnostic tests, the patient can be considered part of this patient group. Each of the three interventions (sentinel node procedure, mastectomy, and axillary clearance) can be mapped to terms found in the guideline. This can be done with reasonable confidence, however, some details have to be ignored such as the re-section of the internal mammary nodes as part of the axillary clearance, as this part of the treatment is not mentioned in the guideline. With respect to the order between interventions, it is only clear that the sentinel node procedure (asbSN) is performed before the other two interventions.

The treatment can again be critiqued using a CTL proof obligation, but as some of the information is missing here, we illustrate critiquing using modular model checking. The proof obligation is then described by $[\varphi]M\langle\bot\rangle$ where

$$\varphi = \mathbf{G}\,\mathrm{OIBC} \wedge \mathbf{F}\,(\mathrm{asbSN} \wedge \mathbf{X}(\mathbf{F}\,\text{axillary-clearance} \wedge \mathbf{F}\,\text{mastectomy}))$$

The proof obligation $[\varphi]M\langle\bot\rangle$ is true, showing that this combination of interventions is *not* possible (cf. Subsection 3.2). The reason for this can be further analysed by removing one of the order constraints yielding

$$\varphi' = \mathbf{G}\,\mathrm{OIBC} \wedge \mathbf{F}\,\mathrm{asbSN} \wedge \mathbf{F}\,\text{axillary-clearance} \wedge \mathbf{F}\,\text{mastectomy}$$

As $[\varphi']M\langle\bot\rangle$ is true, the formula φ' is further weakened by removing one of the interventions from the conjunct. This results in three new proof obligations, showing that the guideline model does not contain a trace with both a sentinel node procedure and axillary clearance for this patient, while all other combinations appear to be possible. Thus, the conclusion is that the combination of actions that are being prescribed is non-compliant with respect to the guideline.

5 Related Work

The use of the term *critiquing* to describe a system that criticises the solution provided by a human can be attributed to Miller [8], who developed his ATTENDING system for critiquing anaesthesia management. Although critiquing has first been used for evaluating medical treatment plans, since then it has been applied to a wide variety of problems such as engineering design, decision making, word processing, knowledge base acquisition, and software engineering [10]. At the end of the 1990s, when several guideline representation languages were introduced, critiquing using guidelines became a topic of interest, e.g., the approach of Shahar *et al.* [9]. In contrast with previous work, in this approach the patient states are considered for critiquing, besides the physician's actions. Advani *et al.* [1] argued that a critiquing system should adjust its critique for cases when the physician's actions are following the spirit and overall goals or intentions of the guideline designers, even though the actions deviate from the guideline. However, in [7], a case study showed that intentions of the protocol are often implicit and moreover, the intentions reported by experts almost always differ, which makes it hard to model. Recently, there was some progress to overcome this difficulty [11], which might be interesting to integrate in our proposed methodology. Using model checking for verifying properties of formalised medical guidelines is very recent [12,2].

6 Conclusions

The main conclusion of this work is that it is, in principle, possible to use model checking on formalised models in order to critique medical guidelines against patient data. We have shown how critiquing can be characterised in temporal logic and have applied this to a case study on the treatment of breast cancer. The strong aspect of this technology is the high degree of automation as compared to theorem proving, making it suitable for deployment in a critiquing system.

Model checking provides additional value to a simulation-based critiquing of an operational version of the guideline. Such critiquing based on running the operational guideline model through an interpreter only checks the consistency of a patient record against a single trace through the guideline (namely, the one chosen by the interpreter), while model checking compares the patient record against all possible traces through the guideline. This difference is crucial when the guideline is under-specified [5], which is usually the case, and therefore contains non-deterministic choices between treatments.

The fully automated nature of model checking also brings a weakness with it: model checking only *detects* inconsistencies, but does not contribute to the interpretation of the inconsistency. In general, model checking can construct a counter-example illustrating the inconsistency, which is often a very good guide towards tracing its source. However, this only works when model checking global properties, i.e., properties dealing with all possible treatment paths, while in Section 3 we argue that critiquing deals with formulas that establish the existence of a single treatment, thereby making it impossible for the model checker to generate a counter-example. In this paper, we have proposed some general strategies to deal with this (repeated experiments with weaker specifications by relaxing order constraints and by removing actions).

A general conclusion with respect to the breast cancer case study that can be drawn is that a closer correspondence is needed between the processes of guideline construction and data-collection. In fact, this is currently already being partially implemented by the Dutch Institute of Healthcare Improvement: newly constructed guidelines are currently being equipped with a data-collection dictionary, which will ensure the correspondence between collected data and guideline terminology.

Even though the steps in the analysis of the case studies was done manually, it is not difficult to see how to automate this process since the temporal formulas could be generated mechanically. A more challenging question is how to use the result of this process for the construction of a human readable critique. In evidence-based guidelines, explanation and references are often provided, however, formal models of guidelines often abstract from this information making it difficult to provide elaborate information to the practitioner. This is an interesting topic for future research.

References

1. Advani, A., Lo, K., Shahar, Y.: Intention-based critiquing of guideline-oriented medical care. In: Proceedings of AMIA Annual Symposium, pp. 483–487 (1998)
2. Bäumler, S., Balser, M., Dunets, A., Reif, W., Schmitt, J.: Verification of medical guidelines by model checking – a case study. In: Valmari, A. (ed.) Model Checking Software. LNCS, vol. 3925, pp. 219–233. Springer, Heidelberg (2006)
3. Clarke, E.M., Grumberg, O., Peled, A.D.: Model Checking. MIT Press, Cambridge, Massachusetts, London, England (2001)
4. Gertner, A.S.: Critiquing: effective decision support in time-critical domains. PhD thesis, Dept. of Computer & Information Science, University of Pennsylvania (1995)
5. Hommersom, A.J., Groot, P., Lucas, P.J.F., Marcos, M., Martinez-Salvador, B.: A constraint-based approach to medical guidelines and protocols. In: ECAI 2006 WS – AI techniques in healthcare: evidence based guidelines and protocols (2006)
6. Kupferman, O., Vardi, M.Y.: Modular model checking. In: de Roever, W.-P., Langmaack, H., Pnueli, A. (eds.) COMPOS 1997. LNCS, vol. 1536, pp. 381–401. Springer, Heidelberg (1998)
7. Marcos, M., Berger, G., van Harmelen, F., ten Teije, A., Roomans, H., Miksch, S.: Using critiquing for improving medical protocols: Harder than it seems. In: 8th European Conference on Artificial Intelligence in Medicine, pp. 431–441 (2001)
8. Miller, P.: A critiquing approach to Expert Computer Advice: ATTENDING. Pittman Press, London (1984)
9. Shahar, Y., Miksch, S., Johnson, P.: A task-specific ontology for the application and critiquing of time oriented clinical guidelines. In: Proceedings of the sixth Conference on Artificial Intelligence in Medicine in Europe, pp. 51–61 (1997)
10. Silverman, B.G.: Survey of expert critiquing systems: Practical and theoretical frontiers. Communications of the ACM 35(4), 106–127 (1992)
11. Sips, R., Braun, L., Roos, N.: Applying formal medical guidelines for critiquing. In: ECAI 2006 WS – AI techniques in healthcare: evidence based guidelines and protocols (2006)
12. Terenziani, P., Giodano, L., Bottrighi, A., Montani, S., Donzella, L.: SPIN model checking for the verifcation of clinical guideline. In: ECAI 2006 WS – AI techniques in healthcare: evidence-baded guidelines and protocols (2006)

Integrating Document-Based and Knowledge-Based Models for Clinical Guidelines Analysis

Gersende Georg[1,2,3] and Marc Cavazza[4]

[1] INSERM, U 872, Eq. 20, Paris, F-75006 France
[2] Université Paris Descartes, UMR S 872, Paris, F-75006 France
[3] Centre de Recherche des Cordeliers, Université Pierre et Marie Curie - Paris 6,
UMR S 872, Paris, F-75006 France
Gersende.Georg@spim.jussieu.fr
[4] School of Computing, University of Teesside
TS1 3BA Middlesbrough, United Kingdom
m.o.cavazza@tees.ac.uk

Abstract. Research in the computerization of Clinical Guidelines (CG) has often opposed document-based approaches to knowledge-based ones. In this paper, we suggest that both approaches can be used simultaneously to assess the contents of textual Clinical Guidelines. In this first experiment, we investigate the mapping between a document model, which has been marked-up to structure its recommendations, and a knowledge structure representing the management of specific disease. This knowledge representation is based on planning formalisms, more specifically Hierarchical Task Networks (HTN). Our system operates by first automatically encoding the textual guideline through the identification of specific expressions with surface natural language processing, as described in previous work. In a subsequent step, the HTN, constructed manually and independently, and represented as an explicit AND/OR graph, is searched for a solution sub-graph using an algorithm derived from AO*. Whilst the HTN is being traversed, corresponding information is accessed in the encoded textual CG, to guide the solution extraction process. We illustrate this through a case study developed around French guidelines for the management of hypertension. Recommendations included in the textual guideline provide complementary information for the instantiation of an HTN on specific patient data. The mapping takes place at different levels, from the pre-condition of operators to the rules playing a role as selection heuristics when extracting a solution sub-graph. Such a process, which explores the textual document from the prospective of a task model, can help analyzing the overall structure of clinical guidelines and ultimately improving its applicability.

Keywords: Clinical Guidelines, GEM, Planning, HTN, deontic operators.

1 Introduction and Rationale

The computerization of clinical guidelines has followed two main approaches, known as document-based and knowledge-based [1], which have so far largely remained separate. Document-based approaches are more closely connected to the actual

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 421–430, 2007.

guideline production process. On the other hand, knowledge-based models can directly be integrated into Decision Support Systems (DSS), whilst document-based models require an additional level of interpretation to extract rules or decision trees, whose automation has recently attracted interest from several researchers [2] [3] [4]. The direct extraction of complete knowledge structures from free text remains a long-term research objective still beyond the state-of-the-art (in particular in terms of Natural Language Processing techniques). As a first step, Hagerty [5] has proposed the use of Information Extraction techniques for the automatic identification of conditional expressions within recommendations (as part of his "Hypertext Guideline Markup Language (HGML)" markup annotation).

However, there could be benefits in the joint use of document-based and knowledge-based approaches for studying the structure of clinical guidelines and assessing their consistency and completeness. This is what we explore in this paper, as we describe a software environment for the analysis of textual guidelines based on the joint use of guideline document encoding and a knowledge-based formalization of the underlying clinical protocol.

2 Modeling Guidelines with Planning Formalisms

Most knowledge-based approaches, such as PRO*forma* [6] and Prodigy [7] are based on knowledge structures centered on clinical actions. These structures are related to action representations encountered as part of planning formalisms such as STRIPS or PDDL [8]. This has been systematized only recently by [8] who have analyzed the role of planning formalisms and discussed the applicability of planning approaches (i.e. not limited to the representation of elementary actions) to the modeling of clinical guidelines. In their review, Bradbrook et al. [8] mention the possible use of Hierarchical Task Network (HTN) planning to represent guidelines protocols, although their own approach is based on other planning formalisms. There are however many benefits in using an HTN approach to model the overall guideline behavior. HTN are one of the most successful planning formalisms and are used in a variety of implemented systems, from robotics [9], game playing (bridge) [10], and virtual characters animation [11]. HTN are considered appropriate to knowledge-intensive Planning problems and their top-down descriptions are well suited to the description of clinical protocols. Other knowledge-based approaches bear similarity to Planning. GUIDE, based on Petri Nets, represents the workflow of clinical guidelines [12] generally composed of a nested sequence of actions. Asbru [13] is a time-oriented, intention-based, skeletal-plan specification language that is used to represent clinical protocols. Its plans attributes are characterized by intentions, conditions, and effects which can be structured via a temporal representation. All these approaches make use of planning concepts without however implementing any of the traditional planning techniques: this is largely because their main focus is on making knowledge procedural to facilitate its integration in DSS.

Although Planning has not been one of the main AI technologies used in medical knowledge-based systems, there has been interest in Planning formalisms for the formalization of clinical protocols. Haddawy et al. [14] have shown as early as 1995 that Planning formalisms represented a benefit over traditional decision trees.

Spyropoulos [15] has reported the use of planning and scheduling techniques to model both therapy planning and hospital management procedures.

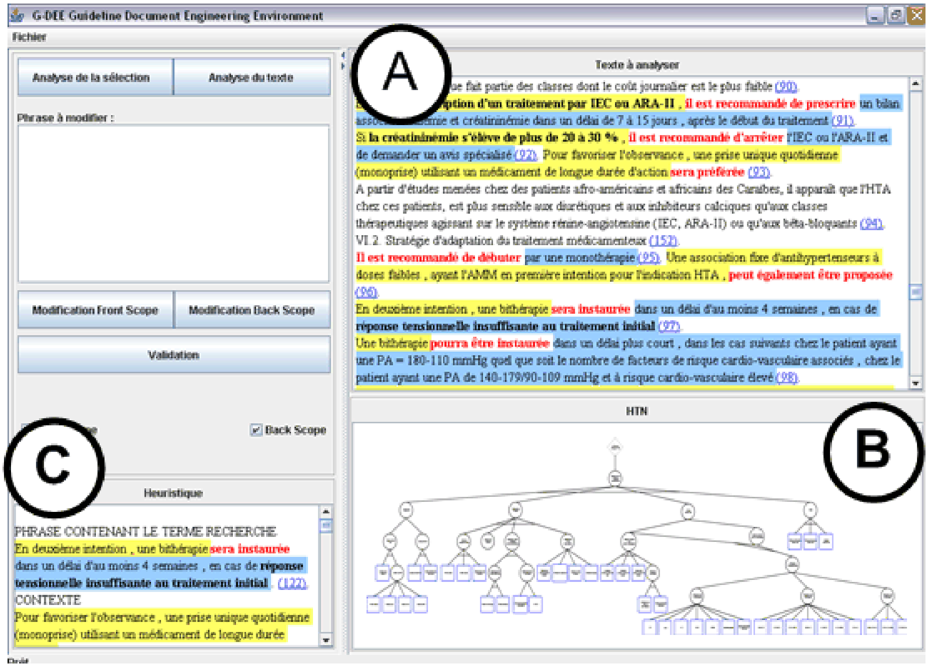
2.1 HTN Formalization of Clinical Guidelines

HTN are appropriate to represent multi-step decomposable processes, and this applies naturally to clinical protocols, as long as they can be decomposed into independent sub-tasks. This is why the various steps of clinical care can be represented as an AND/OR graph formalizing an explicit HTN (i.e. one in which the main task has been decomposed *a priori* and entirely, down to the level of grounded actions, rather than being dynamically refined using decomposition methods [16]).

Yet, there are other important representational elements to be used in conjunction with AND/OR graphs for the instantiation of a solution plan on a specific set of patient data. Each sub-task should be associated pre-conditions as well as post-conditions. Another element of representation is constituted by the heuristics guiding node selection at the level of OR nodes, and costs (in the algorithmic sense) associated to the actual clinical actions. The main difference with traditional HTN planners, such as SHOP [17], will consist in using an explicit and finite HTN. In this way, the overall guideline can also be represented visually, rather than as a collection of refinement methods. The explicit nature of AND/OR graphs thus allows a direct visualization, which in turn facilitates knowledge elicitation. The guideline contents are represented as an AND/OR graph with the highest-level task (the overall goal of the clinical protocol) as the top node. An instantiation of the guideline recommendations can be obtained by extracting a solution from this explicit HTN. Provided certain limitations are properly taken into account, such as task decomposability, absence of long-distance dependencies and stability of data over time (or at least within each sub-task covered by the guideline), a solution sub-graph can be extracted from the HTN with a simple variant of the AO* algorithm [18], which provides solutions to decomposable problems. Primary heuristics will be used in the selection of one out of several alternative sub-tasks subsumed by an OR node. Examples of such sub-tasks are constituted by alternative therapeutics (see below). The solution graph will then constitute an instance of the guideline, to be applied to the specific data for the patient being considered.

2.2 System Architecture and Overview

The approach described in this paper is based on an experimental software platform integrating a document engineering environment [19] (Fig. 1 – A) and an HTN-based module (Fig. 1 – B). In essence, this software aims at synchronizing the traversal of the explicit HTN and the consultation of a guideline document, previously processed to mark-up its elementary recommendations. This synchronization is based on the automatic extraction of information from the marked-up guideline corresponding to the HTN sub-task under consideration. In other words, once the protocol has been modeled as an HTN, the system-driven exploration of its AND/OR graph drives the interactive consultation of the guideline document model to assist in the instantiation of the solution sub-graph (Fig. 1 – C).



…ue fait partie des classes dont le coût journalier est le plus faible (90).
… …ption d'un traitement par IEC ou ARA-II , il est recommandé de prescrire un bilan asso… …mie et créatininémie dans un délai de 7 à 15 jours , après le début du traitement (91).
Si la créatininémie s'élève de plus de 20 à 30 % , il est recommandé d'arrêter l'IEC ou l'ARA-II et de demander un avis spécialisé (92). Pour favoriser l'observance , une prise unique quotidienne (monoprise) utilisant un médicament de longue durée d'action sera préférée (93).
A partir d'études menées chez des patients afro-américains et africains des Caraïbes, il apparaît que l'HTA chez ces patients, est plus sensible aux diurétiques et aux inhibiteurs calciques qu'aux classes thérapeutiques agissant sur le système rénine-angiotensine (IEC, ARA-II) ou qu'aux bêta-bloquants (94).
VI.2. Stratégie d'adaptation du traitement médicamenteux (152).
Il est recommandé de débuter par une monothérapie (95). Une association fixe d'antihypertenseurs à doses faibles , ayant l'AMM en première intention pour l'indication HTA , peut également être proposée (96).
En deuxième intention , une bithérapie sera instaurée dans un délai d'au moins 4 semaines , en cas de réponse tensionnelle insuffisante au traitement initial (97).
Une bithérapie pourra être instaurée dans un délai plus court , dans les cas suivants chez le patient ayant une PA = 180-110 mmHg quel que soit le nombre de facteurs de risque cardio-vasculaire associés , chez le patient ayant une PA de 140-179/90-109 mmHg et à risque cardio-vasculaire élevé (98).

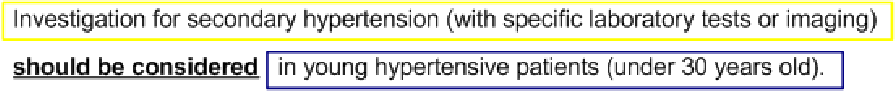
PHRASE CONTENANT LE TERME RECHERCHE
En deuxième intention , une bithérapie sera instaurée dans un délai d'au moins 4 semaines , en cas de réponse tensionnelle insuffisante au traitement initial (122).
CONTEXTE
Pour favoriser l'observance , une prise unique quotidienne (monoprise) utilisant un médicament de longue durée

Fig. 1. The *G-DEE* interface *(see text)*

2.3 Document Processing Applied to Clinical Guidelines

The central element of these experiments is a document engineering platform dedicated to the study of clinical guidelines, developed by one of the authors in previous work [19]. The *G-DEE* system (for Guidelines Document Engineering Environment) automatically performs XML encoding of guidelines based on the recognition of the guideline's linguistic content. The system uses a set of approximately 1200 Finite State Automata (FSA), which correspond to 70 syntactic patterns with their morphological variations, to recognize specific natural language expressions corresponding to the linguistic formulation of elementary recommendations (such as "recommended, advised, one should / ought to"… these expressions are known as *deontic operators* [19] [20]). In a subsequent step, the textual occurrence of these recommendations is structured by marking-up the deontic operator and the textual expression to which it applies. These expressions are named *front-scope* and *back-scope* [20], and correspond to the operands of the deontic operator, from which conditions and actions can be extracted. Figure 2 illustrates the *scopes* of the deontic operator "should be considered".

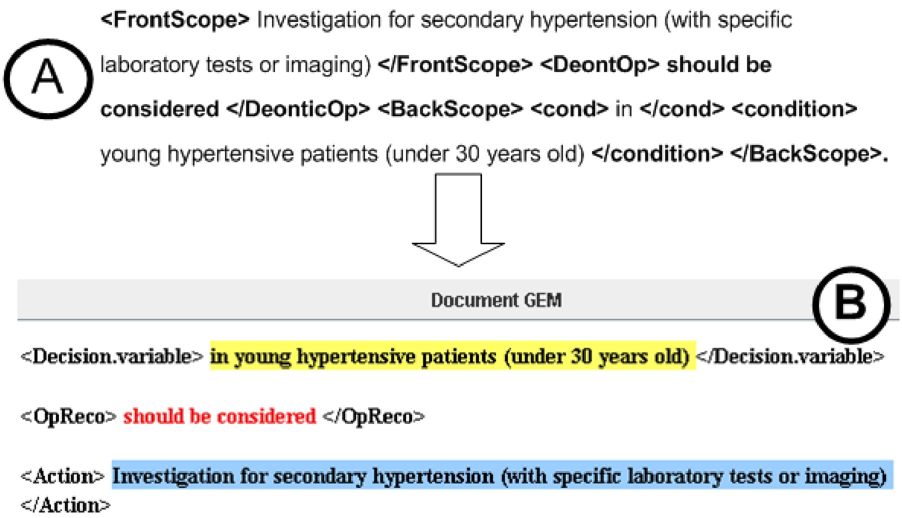
The encoding obtained (Fig. 3) can serve as a basis for further processing, for instance the extraction of decision rules under textual format, or the encoding using GEM [21] categories, for instance through the identification of decision variables [19].



Investigation for secondary hypertension (with specific laboratory tests or imaging)

should be considered in young hypertensive patients (under 30 years old).

Fig. 2. *Front-* and *back-scope* for the deontic operator "should be considered" (translated from the original French Guidelines for the management of hypertension)



<FrontScope> Investigation for secondary hypertension (with specific laboratory tests or imaging) **</FrontScope> <DeontOp> should be considered </DeonticOp> <BackScope> <cond>** in **</cond> <condition>** young hypertensive patients (under 30 years old) **</condition> </BackScope>.**

Document GEM

<Decision.variable> **in young hypertensive patients (under 30 years old)** </Decision.variable>

<OpReco> **should be considered** </OpReco>

<Action> **Investigation for secondary hypertension (with specific laboratory tests or imaging)** </Action>

Fig. 3. The marking-up of a recommendation (part A) and its automatic structuration in the GEM format using dedicated XSL style sheets (part B)

From a content perspective, the recommendations so identified correspond to decision steps and, when they relate to possible alternatives in the therapeutic plan, to actual heuristics that can be used to select the most appropriate alternative in the extraction of a solution task graph. This is the type of content which is central to the mapping of HTN traversal to the guideline document. One key problem of knowledge representation is actually to properly interpret textual recommendations in terms of selection heuristics or grounded action costs.

2.4 Synchronization of HTN Traversal and Document Exploration

The most important aspect of these experiments is the synchronization of HTN traversal with text consultation, which will support the interactive features of the environment. Considering that the HTN has been developed independently of the textual guideline[1] (Fig. 4), this synchronization should relate the contents of HTN nodes traversed to the contents of the textual guidelines.

The first step consists in an "offline heuristic calculation mode" that determines heuristic values of nodes by rolling back estimated costs of grounded actions (the actions associated to the bottom nodes of the HTN such as drug prescription). This

[1] "Management of adults with essential hypertension – 2005 update" (http://www.has-sante.fr/).

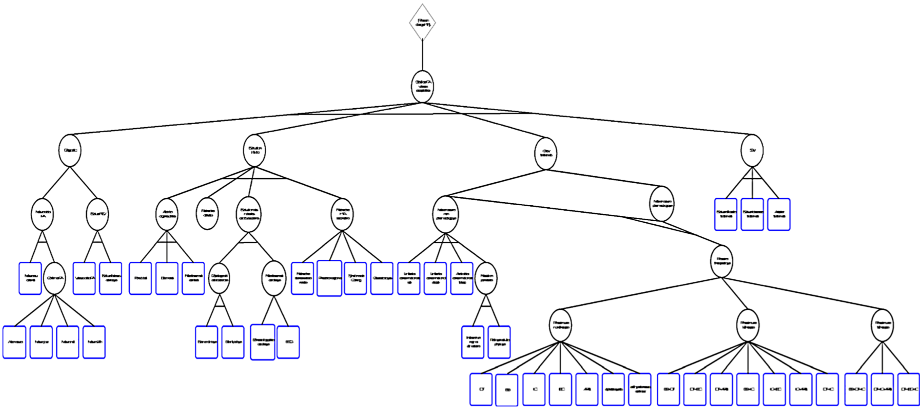


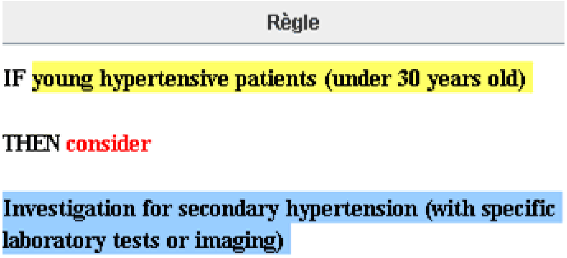
Fig. 4. An overview of the HTN that represents the management of hypertension (pre- and post-conditions attached to the HTN nodes are not represented on the figure)

determines the heuristic value of a node when it can be evaluated from its sub-tasks, e.g. the evaluation of cardiovascular risk which can have an impact on the choice of therapy. The interactive nature of the HTN exploration actually leads us to dissociate the two aspects of heuristic calculation in algorithms of the AO* family [16], which comprise a primary heuristic determining the selection of a solution basis as well as a "rollback" mechanism propagating the cost of grounded actions (for finite graphs).

The second step ("online heuristic mode") determines possible heuristics from the text contents, to be validated interactively by the user. The underlying principle is that certain recommendations actually take the form of a heuristic rule selecting a course of action, which is equivalent (in a non-numerical form) to a heuristic for sub-task selection / task decomposition (see section 3).

During HTN traversal (automatically driven by AO*), the marked-up document is searched for occurrences of linguistic terms associated to the node under consideration – this may require to associate to each node a short list of key synonym terms corresponding to the most frequent formulations of the node's concept (in a subsequent step concept recognition from NL expressions could be envisioned, although the current approach seems to account for the vast majority of actual occurrences). When reaching an OR node, the heuristic to select the solution basis can be derived from the textual recommendations. The overall process could be described as an interactive AO* accessing textual decision elements within the document. When reaching an OR node, the algorithm would access those sections of the textual guideline referring to its key concept and will identify the closest or embedding recommendation. The marked-up recommendation will be presented together with its decision rule format making its role as a heuristic more visible to the user (Fig. 5).

The user then interprets the different recommendations highlighted by *G-DEE* in the dedicated interactive window and she can specify the corresponding heuristic value for the node considered. The interactive exploration of the HTN resumes after validation of a heuristic value by the user. The final output of an interactive session is a candidate explicit task decomposition (only containing AND nodes) which is ready for instantiation on the data specific to a patient profile (Fig. 6).



Règle

IF young hypertensive patients (under 30 years old)

THEN consider

Investigation for secondary hypertension (with specific laboratory tests or imaging)

Fig. 5. Decision rule automatically derived from the recommendation described in Fig. 2 using dedicated XSL style sheets

3 Example Results: The French Hypertension Guidelines[1]

The overall system behavior consists in traversing the HTN from the top node and extracting a solution graph. In order to achieve this, data are accessed for an example patient, which will drive the instantiation of the various pre-condition of the plan operators. This is where it is important to determine which textual data is instantiating operators and which one is used by heuristic rules (e.g. such as age in the exploration of secondary hypertension).

For each node traversed, the terms attached to the node are first localized in the textual guideline, and the embedding recommendation (if any) is highlighted (Fig. 1 – C). As an example, we can consider a patient with: 1) a high level of cardiovascular risk and an antihypertensive monotherapy based on diuretics; 2) an inefficacy of this treatment. When the node "prescribe drug treatment" (Fig. 6) is selected in the HTN, three recommendations are highlighted in the clinical guidelines. The first recommendation reads "If the BP target is not achieved with first-line therapy, a combination of two drugs may be started as second-line therapy after at least 4 weeks", the second is "However, it may be started earlier in patients with BP • 180/110 mmHg regardless of the number of CVR factors; in patients with BP of 140-179/90-109 mmHg and a high CVR.", and the last: "If the patient does not respond to the initial therapy after 4 weeks or experiences side effects, a drug from a different therapeutic class should be prescribed". The heuristic that can be derived from these recommendations corresponds to the choice of therapy, i.e. bitherapy versus a change in therapeutic class or the prescription of a tritherapy. This node will typically be a successor of the "prescribe drug treatment" node in the HTN (Fig. 6). Some functions have been automated, such as the "offline heuristic mode" (and consequently the derivation of the sub-graph[2]).

In future versions of our system, the selection heuristics will also be automatically derived from a post-processing of the decision variable associated to the recommendation, which will be extracted using the same content extraction techniques than those identifying recommendations (with any remaining ambiguities interactively solved by the user).

[2] We used GraphViz (http://www.graphviz.org/) to interpret the file generated by *G-DEE* and enable to build png file of the HTN.

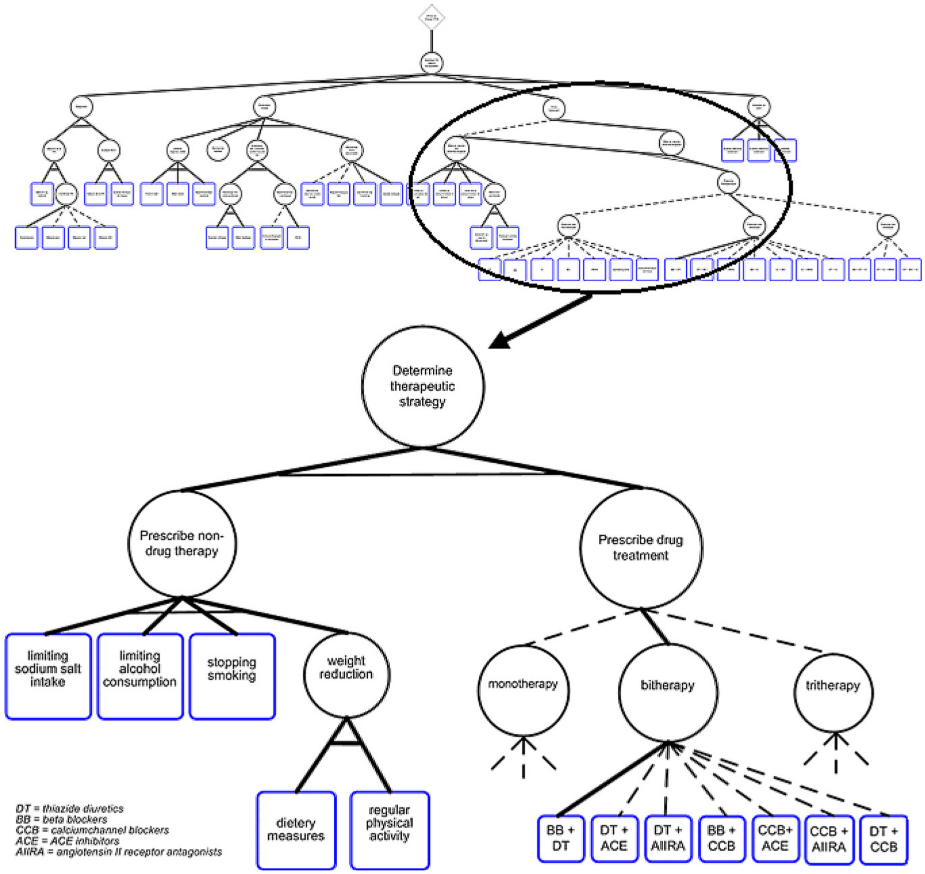


Fig. 6. A solution sub-graph (therapeutics) for initial patient data *(see text)*

4 Discussion

Our first experiments have shown that the structure of the document can be significantly disconnected from the logical flow of the clinical protocol. There are several possible explanations to this situation. Some are related to the social dynamics of guideline writing by committees, where, due to the necessity of achieving consensus, the inclusion and position in the text of some recommendations may not entirely reflect the logical flow of actions, at least within any given sections of the document (corresponding to a single phase of the protocol) or a protocol complexity which results in difficulties to present actions in a linear format in the document. Another aspect consists, when documents grow more complex, in various interim summarizations or the description of high-level strategies, which are generally redundant. Other descriptions encountered in the text are examples of plan outcomes, which illustrate the use of recommendations in context. The inclusion of example plan outcomes probably plays a useful explanatory role, as textual presentations do not offer the same kind of overview as more graphical ones such as HTN. On the other hand, textual guidelines seem to complement more formal representations such as HTN for which they constitute a rich source of contextual information (including heuristics), meta-knowledge and explanations/justifications.

5 Conclusions

The difference between document-based and knowledge-based approaches to clinical guidelines formalization is not simply a difference in the encoding of clinical knowledge or a difference in the semantics of the knowledge representation [22]. The two approaches seem to differ in the type on interpretation required and the meta-knowledge they contain. Because knowledge-based approaches often formalize protocols in an optimal way, producing a minimal and fully ordered structure, they can be useful to guide the exploration of textual guidelines, which often contain many additional data requiring interpretation and are sometimes structured in a less ordered fashion than the one allowed by hierarchical tasks descriptions.

Our current work on the system is dedicated to the progressive automation of HTN plan instantiation from textual guideline. Central to this is the automatic identification of decision variables from document content and their encoding in a rule format supporting automatic derivation of heuristics. Even when this step is achieved, the primary objective of our system will remain, in the first instance, the study of completeness and consistency of clinical guidelines.

Acknowledgments. Gersende Georg is funded through a post-doctoral fellowship from “Region Ile-de-France”. The HTN editor and Visio™ generator used in this paper has been developed by Fred Charles, who is thanked for his assistance.

References

1. Georg, G.: Computerization of Clinical Guidelines: an Application of Medical Document Processing. In: Silverman, B.G., Jain, A., Ichalkaranje, A., Jain, L.C. (eds.) Intelligent Paradigms for Healthcare Enterprises Systems Thinking, vol. 184, pp. 1–30. Springer, Heidelberg (2005)
2. Shiffman, R., Michel, G., Essaihi, A.: Bridging the guideline implementation gap: a systematic approach to document-centered guideline implementation. J Am Med Inform Assoc 11, 418–426 (2004)
3. Marcos, M., Balser, M., ten Teije, A., van Harmelen, F., Duelli, C.: Experiences in the formalisation and verification of medical protocols. In: Artificial Intelligence in Medicine, pp. 132–141. Springer, Heidelberg (2003)
4. Georg, G., Séroussi, B., Bouaud, J.: Does GEM-encoding clinical practice guidelines improve the quality of knowledge bases? A study with the rule-based formalism. In: Proceedings AMIA Symp, pp. 254–258 (2003)
5. Hagerty, C., Pickens, D., Chang, J., Kulikowski, C., Sonnenberg, F.: Prediction in Annotation Based Guideline Encoding. In: Proceedings AMIA Symp, pp. 314–318 (2006)

6. Fox, J., Johns, N., Rahmanzadeh, A., Thomson, R.: PROforma: A method and language for specifying clinical guidelines and protocols. In: Proc Medical Informatics Europe, pp. 516–520. IOS Press, Amsterdam (1996)
7. Johnson, P., Tu, S., Booth, N., Sugden, B., Purves, I.: Using scenarios in chronic disease management guidelines for primary care. In: Proceedings AMIA Symp, pp. 389–393 (2000)
8. Bradbrook, K., Winstanley, G., Glasspool, D., Fox, J., Griffiths, R.A.: Planning Technology as a Component of Computerised Clinical Practice Guidelines. In: Miksch, S., Hunter, J., Keravnou, E.T. (eds.) AIME 2005. LNCS (LNAI), vol. 3581, pp. 171–180. Springer, Heidelberg (2005)
9. Belker, T., Hammel, M., Hertzberg, J.: Learning to optimize mobile robot navigation based on HTN plans. In: Proceedings of International Conference on Robotics and Automation IEEE, pp. 4136–4141 (2003)
10. Smith, S., Nau, D., Throop, T.: Planning Approach to Declarer Play in Contract Bridge. Computational Intelligence 12, 106–130 (1996)
11. Cavazza, M., Charles, F., Mead, S.: Character-based Interactive Storytelling. IEEE Intelligent Systems 17, 17–24 (2002)
12. Quaglini, S., Stefanelli, M., Cavallini, A., Micieli, G., Fassino, C., Mossa, C.: Guideline-based careflow systems. Artif Intell Med 20, 5–22 (2000)
13. 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)
14. Haddawy, P., Doan, A., Goodwin, R.: Efficient Decision-Theoretic Planning: Techniques and Empirical Analysis. In: UAI '95 Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, pp. 229–236. Morgan Kaufmann, San Francisco (1995)
15. Spyropoulos, C.: AI planning and scheduling in the medical hospital environment. Artificial Intelligence in Medicine 20, 101–111 (2000)
16. Ghallab, M., Nau, D., Traverso, P.: Automated Planning - Theory and Practice. Morgan Kaufmann, San Francisco (2004)
17. Nau, D., Cao, Y., Lotem, A., Muñoz-Avila, H.: SHOP: Simple Hierarchical Ordered Planner. In: Proceedings IJCAI-99 968–973 (1999)
18. Heuristics, P.J.: Intelligent Search Strategies for Computer Problem Solving. Addison-Wesley Publishing Company Inc, Reading (1985)
19. Georg, G., Jaulent, M.-C.: An Environment for Document Engineering of Clinical Guidelines. In: Proceedings AMIA Symp, pp. 276–280 (2005)
20. Moulin, B., Rousseau, D.: Knowledge acquisition from prescriptive texts. In: International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. IEA/AIE, pp. 1112–1121 (1990)
21. 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)
22. Eriksson, H., Tu, S.W., Musen, M.: Semantic clinical guideline documents. In: Proceedings AMIA Symp, pp. 236–240 (2005)

Document-Oriented Views of Guideline Knowledge Bases

Samson W. Tu[1], Shantha Condamoor[1], Tim Mather[2],
Richard Hall[2], Neill Jones[3], and Mark A. Musen[1],

[1] Stanford University School of Medicine
Stanford, CA 94305-5479, USA
[2] SCHIN Lt. Newcastle upon Tyne, UK
[3] University of Newcastle, Newcastle upon Tyne, UK

Abstract. A computer-interpretable guideline knowledge base can be a very large network whose information content is difficult for developers and clinicians to comprehend and review. We created a method to annotate a guideline model and use the annotations to export the guideline knowledge base in an XML format that can be transformed into a readable document. We applied this method to knowledge bases developed in three different guideline modeling projects to analyze uses and limitations of this approach. We demonstrate the promise of creating such document-oriented views, but conclude that guideline models and knowledge bases should be constructed with the goal of creating such human-comprehensible views from the beginning.

1 Introduction

In recent years, professional societies, health-maintenance organizations, medical publishers, and government agencies have produced a flood of clinical practice guidelines (CPGs) for the purpose of disseminating evidence-based best practices. Computer-based clinical decision-support systems that provide assistance in clinical settings have been shown to be effective in improving the performance of care providers [1]. To provide computer-based decision support based on CPGs, the medical knowledge in largely narrative CPG documents must be formalized in computer-interpretable models that can be applied to coded patient data to generate patient-specific recommendations or critiques.

Recent literature on the relationship between narrative text and guideline-based decision support focuses on the translation of such documents to computer-interpretable knowledge bases [2-4].

Guideline knowledge bases, however, should be *human-comprehensible* as well as computer-interpretable. Usually, developing such knowledge bases involves collaborations between guideline encoders and subject-matter experts. Without a review format that makes the content of knowledge bases intelligible to those who are not users of sophisticated knowledge-engineering tools, knowledge bases become black boxes that are not subject to human inspection and review. Having such a review format not only benefits collaboration with subject matter experts, but also provides guideline encoders with an efficient method for reviewing a knowledge base

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 431–440, 2007.

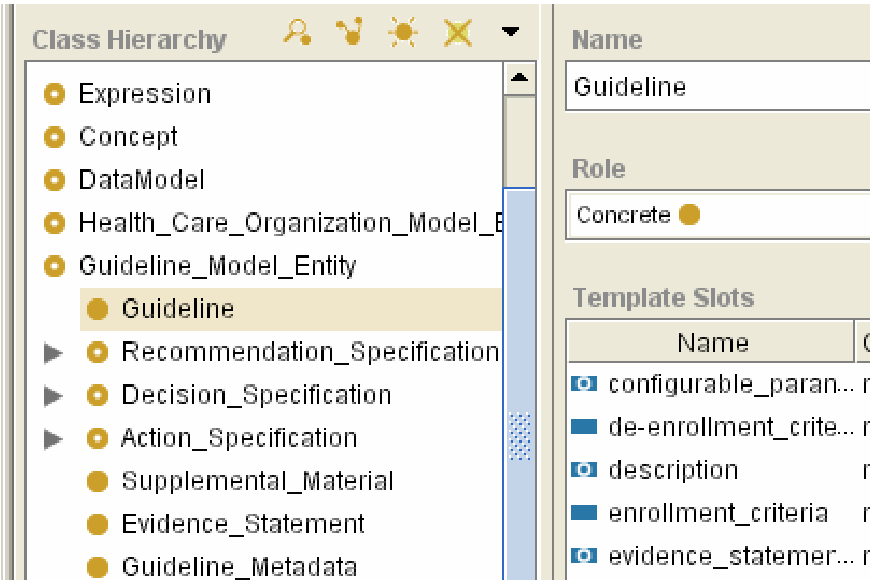
systematically, just as code inspection and walkthroughs are part of the software testing process [5]. Furthermore, any sharing of computer-interpretable guidelines across multiple institutions requires that they be reviewable by clinicians.

This paper describes experiments we conducted to create and evaluate document-oriented views of guideline knowledge bases. We developed a method for transforming frame-based knowledge bases to XML and then to HTML. We applied the method to guideline knowledge bases developed in the SAGE [6], ATHENA [7], and PRODIGY[8] projects. We conducted formative evaluation by obtaining feedback from clinical collaborators and by analyzing the document-oriented views in relation to properties of the guideline models used in these projects.

2 Problem Description

The SAGE (Standards-Based **S**harable **A**ctive **G**uideline **E**nvironment) project [6] was a collaboration among research groups at GE Healthcare Integrated IT Solutions, the University of Nebraska Medical Center, Intermountain Health Care, Apelon, Inc., Stanford University, and the Mayo Clinic. The project sought to create the technology for integrating guideline-based decision support into enterprise clinical information systems. The PRODIGY project [8] in the United Kingdom was similarly funded to develop a guideline-based decision-support system that can assist general practitioners in the task of choosing rational therapeutic actions for their patients. The ATHENA Hypertension Guideline Decision Support System [7] was a project that used EON technology [9] to develop and evaluate the implementation of a hypertension guideline decision-support system at Department of Veteran Affairs clinics. All three projects used Protégé-frame [10], a knowledge-engineering environment developed at Stanford University, as the tool for modeling and encoding CPGs. Each project defined its computer-interpretable guideline model as a *guideline ontology* consisting of Protégé classes (Figure 1). Individual guidelines (e.g., a guideline for managing hypertension) were encoded as instances of these classes. Generally, clinicians who were trained in the use of Protégé did the bulk of the encoding work. Each of the CPG knowledge bases typically included several thousand frames. Even expert users of Protégé might have difficulty drilling down to the depth of the knowledge base and understanding the modeling decisions made by the encoders.

The goal of creating a document-oriented view of a knowledge base was to allow subject-matter experts and guideline encoders to read and review the content of the knowledge base systematically. The format should expose large amount of information and should present an accurate view of the computer-interpretable knowledge content. The view should be "readable" on the web or as printed document, where "readable" meant that the generated text was comprehensible to someone not trained in the knowledge representation formalism, and that the document organized significant portions of the content as linear narrative. The document-generation capability should be generic, as we wanted to use the capability to generate views of guideline knowledge bases from different projects. Furthermore, the views should be highly configurable, as custom-tailored presentations might be required for different classes of readers and for different types of content.



- Expression
- Concept
- DataModel
- Health_Care_Organization_Model_E
- Guideline_Model_Entity
  - Guideline
  - Recommendation_Specification
  - Decision_Specification
  - Action_Specification
  - Supplemental_Material
  - Evidence_Statement
  - Guideline_Metadata

Name: Guideline

Role: Concrete ●

Template Slots:

- configurable_paran... r
- de-enrollment_crite... r
- description r
- enrollment_criteria r
- evidence_statemer... r

Fig. 1. Partial view of the SAGE guideline ontology. The *Guideline* class has properties such as *configurable parameters*, *de-enrollment criteria*, *description*, and *enrollment criteria*.

3 Method

Natural language generation (NLG), a subfield of artificial intelligence, is concerned the generation of text from structured information. In a survey of NLG in healthcare, Cawsey et al. outlined the architecture of NLG as consisting of three stages: (1) text planning (the large-scale organization of the text into coherent sections), (2) sentence planning (the division of information into paragraphs and sentences), and (3) realization (the generation of grammatically correct sentences) [11].

For text planning, we had to determine the scope of information to be presented. A fully developed CPG knowledge base contained not only the content of a computable guideline, but also formal terminologies and a patient data model that were need for the guideline to be implemented in the electronic medical record. We decided that, for our experiments, the scope of the document-oriented view should include frames that were reachable directly and indirectly from instances of the top-level *Guideline* class. In terms of Protégé, it meant the content of the document consisted of trees of instances, where the root of a tree was an instance of the *Guideline* class and branches were the relationships (e.g. *enrollment criteria and recommendation sets*) that defined the structure of a guideline. However, because of interconnectivity of the frames, a simple exhaustive tree-based expansion would result in tremendous repetitions. The content should therefore be partitioned into sections, with hyperlinks connecting references in one section to the content in another. Furthermore, to help navigating the document, we should exploit Protégé graphs that all three projects used to represent task networks of guideline scenarios, decisions, and actions. Thus, the tool should create clickable images of these graphs that allowed a user to navigate to different parts of the document.

For sentence planning, we used an outline format because it corresponded to the underlying nested frame structure and also because scanning an outline was often easier than reading paragraphs. The bottom level of the outline consisted of either text associated with textual properties of a frame or text generated from structured information in one or more frames.

For sentence generation, we decided that, for our initial experiments, instead of finding and using a sophisticated text generator that might add complexity to the system and might not be appropriate for our needs, we would use configurable templates that allowed us to generate the documents quickly.

To support the steps in document generation, we created an *annotation knowledge base* that, for a guideline ontology, specified the large-scale structure of the document and provided context-sensitive templates used by a Java program to convert Protégé instances and graphics into XML fragments and jpeg files (Figure 2). Next, we used Extensible Stylesheet Language Transformation (XSLT), a World Wide Web Consortium standard for rewriting XML documents, to transform the image files and XML file into an HTML file.

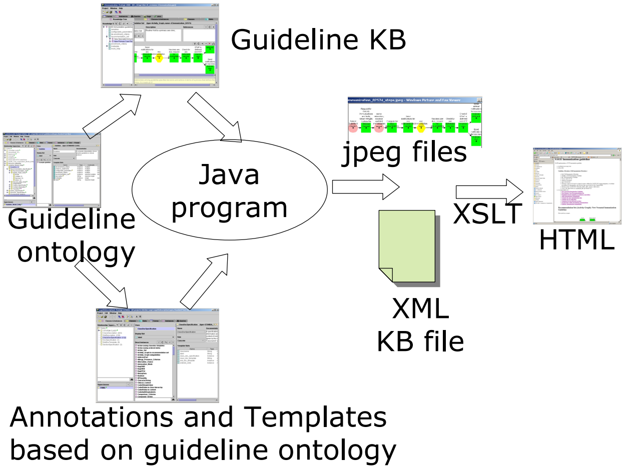


Fig. 2. The process to create a document-oriented view from a guideline knowledge base

The annotation knowledge base for the SAGE guideline ontology specified, for example, the *Guideline, Recommendation,* and *Variables* classes as classes whose instances constituted the main sections of the document to be generated. The Java program then traversed the instance trees anchored by these instances. For each node in the tree, it generated XML output that could be converted to readable text.

To generate the XML output, we created templates for specifying how Protégé instances should be translated. For each Protégé class whose instances we wanted to include in the XML output, we enumerated, as part of the template associated with the class, an ordered list of slots (i.e., properties) whose values should be included in the output. The default XML output used class and slot names as XML tags. Thus, for an instance of the class *Presence_Criterion* in the SAGE guideline model, which had slots *code, presence, valid_window, and vmr_class*, and which allowed a clinician to enter a decision criterion, such as “presence of MMR vaccine administration within last 28 days,” in a fill-in-blank GUI form, the XML fragment looks like the following:

```
  <Presence_Criterion p_id= "sageimmunization_02486">
     <code>MMR vaccine</code>
     <presence>true</presence>
     <valid_window>
       <RelativeTimeInteval> …
         </RelativeTimeInterval>
   </valid_window>
    <vmr_class> SubstanceAdministration
   </vmr_class>
</Presence_Criterion>
```

The *RelativeTimeInterval* element expands into an XML fragment that represents an instance of the *RelativeTimeInterval* class (e.g., the interval between 28 days ago and NOW).

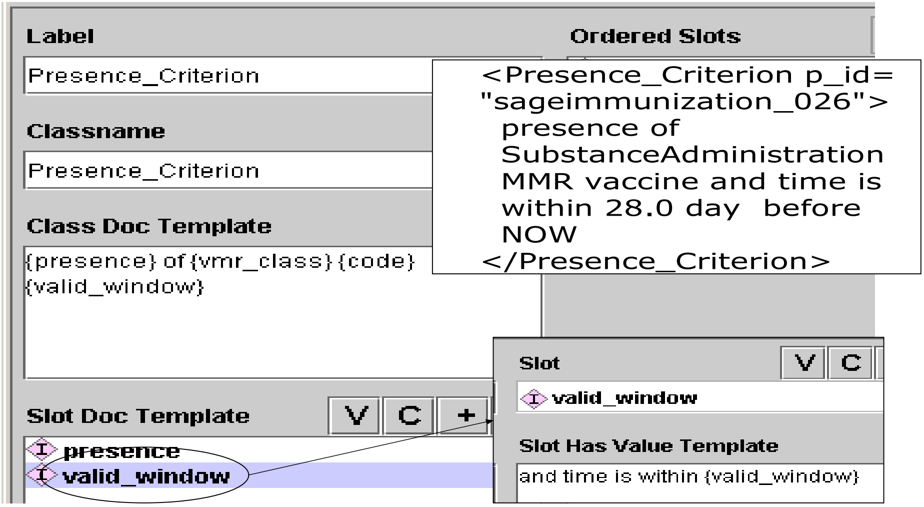
For selected classes in the guideline ontology, we provided alternative templates for specifying textual patterns used to write instances of those classes. The selection of template for an instance was based on the usage context of the instance. Thus, for example, when instances of *Recommendation* referenced instances of *Action* as slot values, one template was used, whereas when instances of class *Decision* referenced instances of *Action*, an alternative template for instances of *Action* was used.

The use of alternative templates for a class allowed us to specify when to expand the content of an instance, and when to reference that instance. In a Protégé knowledge base, frames, such as classes and instances, can be referenced from several other frames. In fact, the reference relationship can be circular. Determining when to expand the content of an instance and when to make a reference to the same content were design decisions that affected the readability of the generated document.

Another reason for providing alternative templates for a class was that we wanted to provide greater user control of the output format. Figure 3 shows an alternative template for generating text for instances of the *Presence_Criterion* class. The pattern “`{presence} of {vmr_class} {code}{valid_window}`” specifies how values of slots should be substituted into the pattern to generate a string. For selected slots, we specified how text could be generated based on presence or absence of slot values. For the slot *valid_window*, the slot template indicates that, the slot value, if it is available, should be preceded by “and time is within ” and followed by the expansion of the value of the *valid_time* slot, a *RelativeTimeInterval* instance. The previous XML fragment, using this template, would result in the text “`presence of SubstanceAdministration MMR vaccine and time is within` …” where the elided relative time interval could be “`28 days before NOW`.”

4 Results

We wrote scripts to generate default annotation knowledge bases for each of the three guideline ontologies to be tested. Similarly, we developed XSL transforms for each guideline ontology.



Label: Presence_Criterion

Classname: Presence_Criterion

Class Doc Template: {presence} of {vmr_class} {code} {valid_window}

Slot Doc Template: presence, valid_window

Slot: valid_window

Slot Has Value Template: and time is within {valid_window}

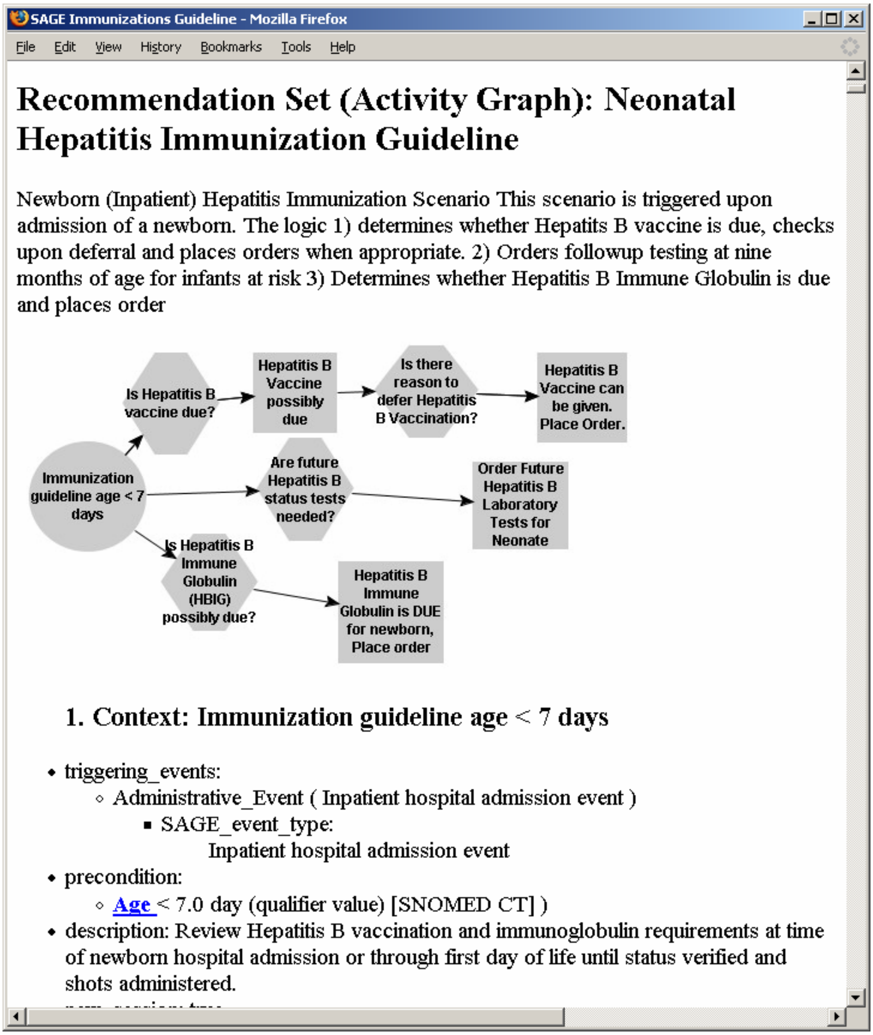
```
<Presence_Criterion p_id=
"sageimmunization_026">
  presence of
  SubstanceAdministration
  MMR vaccine and time is
  within 28.0 day  before
  NOW
</Presence_Criterion>
```

Fig. 3. Alternative template for specifying output format of an instance of Presence_Criterion. The inset box shows an example of the text generated from this template.

Figures 4 and 5 show parts of the HTML pages generated from a SAGE immunization guideline and a PRODIGY post-myocardial infarction guideline. The SAGE HTML page shows the use of graphics and hyperlinks to structure the document. The presentation of the *precondition* of the Context node (an oval in the task graph) shows a formal criterion being displayed as text, with the *Age* variable linked to the section in the document where it is defined.

Because constraints on the resources of the projects involved in this study, a formal evaluation of the document-oriented view of the guideline knowledge base was not feasible. Instead, we performed formative evaluation to explore the text-generation technology's possible uses and limitations, alternative presentations, and properties of guideline knowledge bases that allow better generation of readable text.

The document-oriented view was well-received by the SAGE team. Knowledge-base developers in the project were able to use the HTML document to identify dozens of errors and missing data in the knowledge base. Clinicians found the documents much more accessible than the Protégé knowledge bases from which the documents were generated. However, for the purpose of reviewing the content of the knowledge bases, the clinicians asked for more contextual information about the encoded guideline recommendations. The operationalization of SAGE guidelines involved developing usage scenarios, distillation and interpretation of guideline text, and formalization of decision logic in terms of standard terminologies and a patient information model [6]. Understanding the encoded recommendations required more than having access to the formalized knowledge base. Suggested enhancements to the document included (1) overview documentation to orient a reader, (2) clear indication of relationships between recommendations and sub-recommendations, (3) links to source documents and abstracts of guideline content selected for encoding, and (4) links to example scenarios. Despite these limitations, the document-generation capability proved sufficiently useful for it to be incorporated into the SAGE guideline


SAGE Immunizations Guideline - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Immunization guideline age < 7 days
Is Hepatitis B vaccine due?
Hepatitis B Vaccine possibly due
Is there reason to defer Hepatitis B Vaccination?
Hepatitis B Vaccine can be given. Place Order.
Are future Hepatitis B status tests needed?
Order Future Hepatitis B Laboratory Tests for Neonate
Is Hepatitis B Immune Globulin (HBIG) possibly due?
Hepatitis B Immune Globulin is DUE for newborn, Place order

Recommendation Set (Activity Graph): Neonatal Hepatitis Immunization Guideline

Newborn (Inpatient) Hepatitis Immunization Scenario This scenario is triggered upon admission of a newborn. The logic 1) determines whether Hepatitis B vaccine is due, checks upon deferral and places orders when appropriate. 2) Orders followup testing at nine months of age for infants at risk 3) Determines whether Hepatitis B Immune Globulin is due and places order

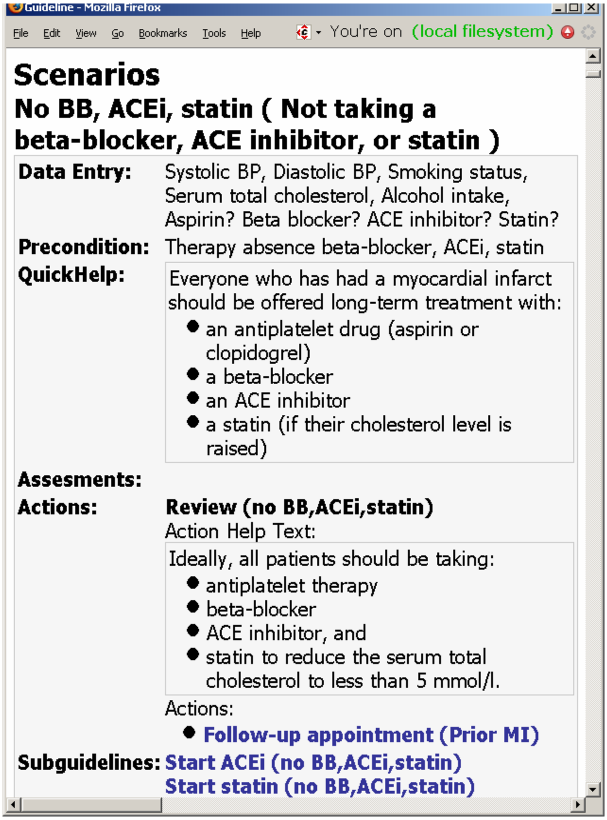
1. Context: Immunization guideline age < 7 days

- triggering_events:
  - Administrative_Event (Inpatient hospital admission event)
    - SAGE_event_type:
      Inpatient hospital admission event
- precondition:
  - Age < 7.0 day (qualifier value) [SNOMED CT])
- description: Review Hepatitis B vaccination and immunoglobulin requirements at time of newborn hospital admission or through first day of life until status verified and shots administered.

Fig. 4. Partial view of the HTML document generated from a SAGE immunization guideline. The clickable image map allows a viewer to navigate to different parts of the document. The circle represents the context of the recommendation (pediatric patient), hexagons decision nodes, and rectangles action nodes.

workbench, thus allowing document-oriented view to be generated at any stage of the knowledge development process.

For the PRODIGY project, the main use case for the document-oriented view involved creating human-readable documents for use by external groups to validate the encoded guideline. Because guideline authors were often far more comfortable authoring in the document paradigm, presenting the complex interconnected network of knowledge components in this way was seen as a significant benefit. The


Guideline - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
You're on (local filesystem)

Scenarios

No BB, ACEi, statin (Not taking a beta-blocker, ACE inhibitor, or statin)

Data Entry: Systolic BP, Diastolic BP, Smoking status, Serum total cholesterol, Alcohol intake, Aspirin? Beta blocker? ACE inhibitor? Statin?

Precondition: Therapy absence beta-blocker, ACEi, statin

QuickHelp: Everyone who has had a myocardial infarct should be offered long-term treatment with:

- an antiplatelet drug (aspirin or clopidogrel)
- a beta-blocker
- an ACE inhibitor
- a statin (if their cholesterol level is raised)

Assesments:

Actions: **Review (no BB,ACEi,statin)**

Action Help Text:

Ideally, all patients should be taking:

- antiplatelet therapy
- beta-blocker
- ACE inhibitor, and
- statin to reduce the serum total cholesterol to less than 5 mmol/l.

Actions:

- **Follow-up appointment (Prior MI)**

Subguidelines: **Start ACEi (no BB,ACEi,statin)**
Start statin (no BB,ACEi,statin)

Fig. 5. Partial view of the HTML document generated from a PRODIGY post-myocardial infarction guideline. It displays the information associated with a scenario where a patient is not recorded as taking any beta blocker, ACE inhibitor, or statin. The data entry section indicates data that should be acquired in this scenario. The actions associated with this scenario include scheduling follow-up appointment and invoking subguidelines to start ACE inhibitor and statin.

PRODIGY guideline model, as encoded in Protégé, consisted of a series of related projects that allowed the re-use of the reference drugs, clinical terms, and decision criteria. The learning curve for using Protégé exceeded what would be reasonable for external reviewers of encoded guideline content. By flattening out this structure and selectively displaying relevant slots, the document-oriented view provided a convenient method for reviewing and ensuring the quality of the encoded knowledge. One key aspect of the PRODIGY guideline was the concise narrative that was associated with each guideline step and that was displayed to the user when the guideline was executed. By emphasizing this *quick-help* slot in the HTML document (see Figure 5), the generated document allowed an effective way of quality-assuring a large quantity of text without having to access each frame individually. We also recognized the possibility of automatically generating training documentation, and, thus, using the knowledge base for multiple purposes.

The ATHENA hypertension guideline knowledge base had been developed, tested, and deployed over a number of years [7]. A simple-minded hierarchical expansion from the *Guideline* node resulted in an HTML document that contained almost 25 thousand lines. The graph of the clinical algorithm used in the knowledge base proved to be insufficient to allow easy navigation in the generated document. A large part of the hypertension guideline knowledge base dealt with the properties and usage of different classes of anti-hypertensive agents. Alternative presentations (possibly in tabular form) were needed to facilitate easier access to that information than the current hierarchical expansion. Thus, we have not yet presented the HTML document generated from ATHENA to clinicians for external review.

5 Discussion

The idea of generating human-readable documents from machine-interpretable artifacts is not new. Cawsey et al. surveyed several systems in healthcare that used text generation to provide explanations, summaries, reports, and descriptions of medical concepts [11]. The original MYCIN program, for example, included a text-generation module that produced natural-language rule translation [12]. More recently, Design-a-Trial [13] provided a clinical trial authoring environment that generated a protocol document and a Prolog-based executable knowledge base. In the wider computer-science literature, this work on creating document-oriented view of a knowledge base is closely related to Knuth's *literate programming* [14], where the construction of computer programs is seen as a task not only to instruct a computer what to do, but also to explain to human being what we want a computer to do. The experiences we gained from creating document-oriented views of guideline knowledge bases are consistent with the tenets of literate programming.

First, while it important to generate text from the knowledge base, such text does not replace the need to have well-written documentation. The PRODIGY example showed how *quick-help* texts, originally designed as an explanation aid for end users, were helpful to guideline authors for quality-assurance purpose. The SAGE clinician reviewers similarly called for overview narrative to orient readers and to record design decisions. Second, our experiments highlight the importance of mapping knowledge base structures to appropriate document structures. The SAGE guideline knowledge bases, for example, primarily consisted of recommendation sets that provide chapter-like divisions whose content were indexed by graphical algorithms. On the other hand, the ATHENA knowledge base contained heterogeneous knowledge structures that required more complex mapping to a document model.

Just as Knuth designed the WEB system so that a programmer can write documentation and code in the same *literate program* [14], our experiences suggest that computable models of clinical guidelines should be designed so that a knowledge modeler can encode an executable and a readable guideline at the same time. Results in software engineering indicate that the maintenance cost of knowledge bases is likely to exceed their initial development cost. Thus, having human-comprehensible semantics is just as important as having machine-executable formal semantics. Our work demonstrated that, with simple XML output and XSL transformations, it is possible to generate rudimentary documents from multiple guideline knowledge bases. We expect to refine our computer-interpretable guideline models and text-generation capability with the goal of producing better document-oriented views of guideline knowledge bases.

Acknowledgement. This work has been supported in part by grant 70NANB1H3049 of the NIST Advanced Technology Program. We thank Kai Rong at SCHIN for his help in generating the PRODIGY version of the HTML output. The ATHENA project was supported in part by VA HSR&D CPG-97-006, CPI-275, and RCD-96301.

References

1. Hunt, D.L., Haynes, R.B., Hanna, S.E., Smith, K.: Effects of Computer-based Clinical Decision Support Systems on Physician Performance and Patient Outcome: A Systematic Review. JAMA 270(15), 1339–1346 (1998)
2. Shiffman, R.N., Michel, G., Essaihi, A., Thornquist, E.: Bridging the Guideline Implementation Gap: A Systematic, Document-Centered Approach to Guideline Implementation. J Am Med Inform Assoc 11, 418–426 (2004)
3. Shalom, E., Shahar, Y.: A graphical framework for specification of clinical guidelines at multiple representation levels. In: AMIA Annu Symp Proc 2005, pp. 679–683 (2005)
4. Ruzicka, M., Svatek, V.: Mark-up based analysis of narrative guidelines with the Stepper tool. Stud Health Technol Inform 101, 132–136 (2004)
5. Myers, G.J.: The Art of Software Testing. John Wiley & Sons, Inc., Hoboken, NJ (2004)
6. Tu, S.W., Musen, M.A., Shankar, R., et al.: Modeling Guidelines for Integration into Clinical Workflow. Medinfo 174–178 (2004)
7. Goldstein, M.K., Hoffman, B.B., Coleman, R.W., et al.: Implementing Clinical Practice Guidelines While Taking Account of Changing Evidence: ATHENA, an Easily Modifiable Decision-Support System for Management of Hypertension in Primary Care. In: Proc AMIA Symp, pp. 280–284 (2000)
8. Johnson, P.D., Tu, S.W., Booth, N., Sugden, B., Purves, I.N.: Using Scenarios in Chronic Disease Management Guidelines for Primary Care. In: Proc AMIA Symp., pp. 389–393 (2000)
9. Tu, S.W., Musen, M.A.: From Guideline Modeling to Guideline Execution: Defining Guideline-Based Decision-Support Services. In: Proc AMIA Symp., pp. 863–867 (2000)
10. 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)
11. Cawsey, A.J., Webber, B.L., Jones, R.B.: Natural language generation in health care. J Am Med Inform Assoc 4(6), 473–482 (1997)
12. Shortliffe, E.H.: Details of the Consultation System. In: Buchanan, B.G., Shortliffe, E.H. (eds.) Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, pp. 78–132. Addison-Wesley Publishing Company, Reading (1984)
13. Modgil, S., Hammond, P.: Decision support tools for clinical trial design. Artificial Intelligence in Medicine 27(2), 181–200 (2003)
14. Knuth, D.E.: Literate Programming. The Computer Journal 27(2), 97–111 (1984)

Maintaining Formal Models of Living Guidelines Efficiently

Andreas Seyfang[1], Begoña Martínez-Salvador[2], Radu Serban[3], Jolanda Wittenberg[4], Silvia Miksch[1,5], Mar Marcos[2], Annette ten Teije[3], and Kitty Rosenbrand[4]

[1] Vienna University of Technology, Austria
[2] Universitat Jaume I, Spain
[3] Vrije Universiteit Amsterdam, The Netherlands
[4] Dutch Institute for Healthcare Improvement, The Netherlands
[5] Danube University Krems, Austria

Abstract. Translating clinical guidelines into formal models is beneficial in many ways, but expensive. The progress in medical knowledge requires clinical guidelines to be updated at relatively short intervals, leading to the term *living guideline*. This causes potentially expensive, frequent updates of the corresponding formal models.

When performing these updates, there are two goals: The modelling effort must be minimised and the links between the original document and the formal model must be maintained. In this paper, we describe our solution, using tools and techniques developed during the Protocure II project.

1 Introduction

Clinical guidelines are "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" [1]. A guideline describes the optimal care for patients and therefore, when properly applied, it is assumed that they improve the quality of care. The fast progress in medical knowledge in many fields leads to the need to update guidelines frequently, leading to the concept of *living guidelines*.

There are several formal languages to model clinical guidelines in a computer processable way, e.g., Asbru, GLIF, PROforma (see [2] for an overview and comparison). While producing a formal model of a guideline has several possible applications, it is difficult and expensive. In addition, the resulting model is often difficult to compare to the original. Furthermore, if the guideline is revised, the modeling effort is lost and it is not easy to detect which changes in the formal model are required by the changes in the original text. At the same time, changes in guidelines become more frequent as medical knowledge grows rapidly and the evidence-based guidelines replace consensus-based ones.

Our experience shows that updates of guidelines are evolutionary and complete rewrites of chapters are infrequent. This means that a method which efficiently

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 441–445, 2007.

traces small changes through several layers of models has considerable potential to save remodelling effort.

The LASSIE methodology uses information extraction to obtain as many fragments of formal representation from a guideline in free-text as possible. This approach is applied to support the idea of living guidelines [3]. It differs from our approach in the modelling methodology – information extraction versus modelling through human knowledge engineer. The combination of both methods is under exploration.

The guideline authoring tool URUZ differs from the tools used in the work described here in the modelling philosophy. The various slots (or knowledge roles) in the ontology of the target language (e.g., Asbru or GEM) are filled with informal content in a first step and optionally refined later. This requires significant training of the users to reach agreement on how to apply the modelling language and to reduce the diversity of modeller decisions [4].

In this paper, we describe our approach to modelling the changes in living guideline with minimal effort as developed during the Protocure II project. In Section 2, we describe the Protocure modelling process. We evaluate our approach in Section 3 and conclude in Section 4.

2 Method: The Protocure Modelling Process

2.1 The Models

Figure 1 gives an overview of the gradual modelling process we use. In our application example, the original guideline was written in Dutch and therefore it was translated to English and then converted to HTML. These first steps are optional, of course, and country-specific.

In Protocure I we modelled the guidelines in Asbru directly from the original text of the guideline [5]. This proved to be a complex task. In Protocure II we therefore designed the intermediate representation MHB [6] to bridge the gap between the original guideline and Asbru. Modelling a guideline in MHB means to transform it into a series of information chunks. Each chunk has aspects grouped into eight different dimensions. The most important dimensions are control flow, data flow, and evidence base.

The process of writing the MHB version of the guideline is supported by the Document Exploration and Linking Tool with Add-ons DELT/A [7]. The user marks a piece of text in the original guideline, which is displayed as HTML in the left-hand window in DELT/A, and selects a suitable macro. Each macro represents a simple pattern in MHB, e.g., a particular aspect of a chunk.

When a macro is activated, it combines the selected text in the original guideline with additional user input and predefined parts and inserts the result in the MHB file. This allows the efficient creation of MHB file with little chances to introduce errors such as typos.

Every macro includes delta-links, which connect the newly inserted elements in the MHB file with the corresponding text in the original guideline. Clicking on one of these links in the MHB file will bring up the corresponding text in the

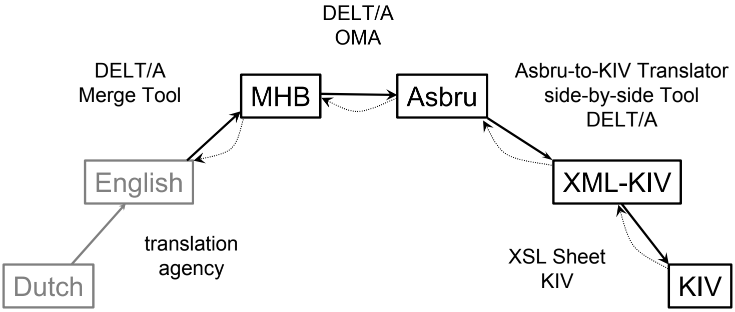


Fig. 1. Modelling steps in Protocure II

left-hand (HTML) window. Likewise, clicking on the automatically created link in the original text will bring up the corresponding MHB chunk in the right-hand window.

In a second step, we create an Asbru [8] model, based on the MHB model. This process resembles the one described above, except for the fact that the MHB model is shown at the left-hand side and the Asbru model is created in the window at the right hand, again using macros and inserting links between the two files.

The Asbru model is then automatically translated to XML-KIV using the Asbru-to-KIV translator. XML-KIV is then converted to KIV using an XSL script. The result is imported into the KIV system, where it can be interactively verified, to examine whether certain properties hold for the given guideline.

Using the delta-links it is easy to find the original guideline text for a given part of the KIV or the Asbru model (via the links in the Asbru and the MHB models) and vice versa. In addition to DELT/A, we developed two simple but powerful visualisation tools, OMA and side-by-side. The first one displays original guideline, and the MHB and Asbru models joined together using an HTML-based, tabular abstraction of the XML-based syntax of MHB and Asbru. The second tool displays two arbitrary XML files side by side as beautified HTML closely based on the XML syntax. Both map delta-links to HTML links allowing for the easy navigation through the different models.

2.2 Modifying the Models

Focusing on the highlighted changes in the text, we went through the HTML text in DELT/A (and in parallel, in the OMA output). For each part of text that was marked as either new or deleted and that had a delta-link attached to it, this link was clicked on. If there was no link, then that particular piece of text had not been modelled (e.g., scientific justification or headings) and hence did not require further actions. Clicking on the link brings up the corresponding MHB chunk, in which the necessary changes were carried out. In most cases, the changes did not affect the Asbru model (e.g., changes of background knowledge).

A (short) list of those changed chunks which could possibly lead to changes in the Asbru model was kept manually. Future versions of DELT/A should include support for versioning, which would replace the manually edited list.

Then the MHB file was displayed in the left-hand window and the Asbru file in the right-hand one. Going through the chunks on the list and clicking on the delta-links showed the relevant parts of Asbru which needed to be changed. Finally, the modified Asbru file was translated to KIV using the automatic Asbru-to-KIV translator.

3 Results

Table 1 shows the modelling effort for the original model and for the changes. It clearly shows that modelling the changes only caused a fraction of the effort of modelling the first version.

This is not only caused by the efficient methods to support living guidelines. Being acquainted to the content of the guideline and to using the tools and representation certainly contributed to reduce the effort in the second round of modelling. This was shown for Chapter 2, for which the Asbru model was redone from scratch, due to the big changes in the chapter update. This took 1.5 person months (PM) while the initial model required 4.5 PMs, i.e., 1/3 of the initial effort. The effort for remodelling other chapters ranged from 1/8 or less to 1/4 of the initial effort, i.e., it was smaller in all cases and significantly smaller than the effort observed for modelling Chapter 2 anew in many cases.

For some chapters the modeller changed between the two versions. However, the sample is far too small and the chapters too different to draw any conclusions regarding the influence of this factor from this experiment.

Overall, our experience showed that utilising the links between the different chapters greatly reduced the modelling effort and helped to ensure that all necessary changes of the model were performed.

Table 1. Effort for original modelling and for modelling the changes. (PM .. person month).

Chapter number	Original effort (PM)	Effort for changes (PM)
1	4	0.5
2	4.5	1.5
3	2	0
4	3.5	0.2
5	2	0.3
6	2	0.5

4 Conclusion

We have shown that it is a good strategy to create a formal model of a real-world clinical guideline in a multi-step process, transforming the free-text form via more and more formal representations to temporal logics with clearly defined semantics. In this translation process, it is important to maintain the connections between corresponding parts in each pair of consecutive models.

Frequent updates to the original text of the clinical guideline, which are required by the swift progress of medical knowledge, threaten to invalidate the comprehensive modelling effort within short time.

Following the links of corresponding parts in our models, we devised a procedure to introduce minimal changes to the original model. At the same time, following all relevant links guarantees the changes in the model to be exhaustive, provided that the links are complete. While there is still an important part of human modelling experience required, which precludes fully automated approaches, the effort is small compared to creating a new model, and a large part of the original modelling effort is preserved. This hypothesis was confirmed in a practical test on a real-world guideline.

Acknowledgement. This work has been supported by the European Commission's IST program, under contract number IST-FP6-508794 Protocure-II.

References

1. Field, M.J., Lohr, K.H. (eds.): Clinical Practice Guidelines: Directions for a New Program, Institute of Medicine. National Academy Press, Washington DC (1990)
2. Peleg, M., Tu, S., Bury, J., Ciccarese, P., Fox, J., Greenes, R., Hall, R., Johnson, P., Jones, N., Kumar, A., Miksch, S., Quaglini, S., Seyfang, A., Shortliffe, E., Stefanelli, M.: Comparing Computer-Interpretable Guideline Models: A Case-Study Approach. JAMIA 10 (2003)
3. Kaiser, K., Miksch, S.: Formalizing 'living guidelines' using LASSIE: A multi-step Information Extraction Method. In: Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME'07) (2007)
4. Shalom, E., Shahar, Y., Lunenfeld, E., Taieb-Maimon, M., Young, O., Bar, G., Martins, S., Vaszar, L., Liel, Y., Yarkoni, A., Goldstein, M., Leibowitz, A., Marom, T.: The Importance of Creating an Ontology-Specific Consensus Before a Markup-Based Specification of Clinical Guidelines. In: AI techniques in healthcare: evidence-based guidelines and protocols; Workshop at the 17th European Conference on Artificial Intelligence (ECAI-06) (2006)
5. ten Teije, A., Marcos, M., Balser, M., van Croonenborg, J., Duelli, C., van Harmelen, F., Lucas, P., Miksch, S., Reif, W., Rosenbrand, K., Seyfang, A.: Improving medical protocols by formal methods. Artificial Intelligence in Medicine 36, 193–209 (2006)
6. Seyfang, A., Miksch, S., Marcos, M., Wittenberg, J., Polo-Conde, C., Rosenbrand, K.: Bridging the Gap between Informal and Formal Guideline Representations. In: 17th European Conference on Artificial Intelligence (ECAI-06) (2006)
7. Votruba, P., Miksch, S., Seyfang, A., Kosara, R.: Tracing the Formalization Steps of Textual Guidelines. In: Computer-based Support for Clinical Guidelines and Protocols, pp. 172–176. IOS Press, Amsterdam (2004)
8. Seyfang, A., Kosara, R., Miksch, S.: Asbru 7.3 Reference Manual. Technical report, Vienna University of Technology (2002)

A Causal Modeling Framework for Generating Clinical Practice Guidelines from Data

Subramani Mani and Constantin Aliferis

Vanderbilt University, Nashville TN 37232, USA
{subramani.mani,constantin.aliferis}@vanderbilt.edu

Abstract. The practice of medicine is becoming increasingly evidence-based and clinical practice guidelines (CPGs) are necessary for advancing evidence-based medicine (EBM). We hypothesize that machine learning methods can play an important role in learning CPGs automatically from data . Automatically induced CPGs can then be used for further manual refinement and deployment, for automated guideline compliance checking, for better understanding of disease processes, and for improved physician education. We discuss why learning CPGs is a special form of computational causal discovery and why simply predictive (i.e., non-causal) methods may not be appropriate for this task.

1 Introduction and Background

Clinical practice guidelines (CPGs) can be broadly classified as predictive guidelines or prevention/intervention guidelines based on their goals. While predictive guidelines may be sufficient for diagnosis or assessing prognosis, we need a cause and effect interpretability for prevention and intervention. In this paper we develop a framework using the representation of causal Bayesian networks (CBNs) for automatic generation of guidelines from data with application to prevention and treatment of disease. The generated guidelines can be evaluated by experts, tested and improved before they are adopted by professional societies or hospital managements for deployment. We discuss why learning CPGs is a special form of computational causal discovery and why simply predictive (i.e., non-causal) methods may not be appropriate for this task.

A causal Bayesian network is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node (variable) and a child node, relative to the other nodes in the network [1]. We define the *causal influence* of a variable A on variable B using the *manipulation criterion* [2,3]. The manipulation criterion states that if we had a way of setting just the values of A and then measuring B, the causal influence of A on B will be reflected as a change in the conditional distribution of B. That is, there exist values a_1 and a_2 of A such that $P(B|\text{ set } A = a_1) \neq P(B|\text{ set } A = a_2)$. For a philosophical approach to causality in the health sciences see for example [4]. The two basic assumptions that are necessary for our causal discovery framework are the causal Markov condition (CMC) and the causal faithfulness condition (CFC) [2,3].

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 446–450, 2007.

The rest of the paper is organized as follows. In Section 2 we discuss the categories of causal and predictive guidelines. In Section 3 we develop the framework for learning clinical practice guidelines using CBNs and point to some preliminary results. In Section 4 we highlight the strengths of the causal modeling framework for guideline induction, point out a limitation and give direction for future research.

2 Causal Versus Predictive Guidelines

Consider the hypothetical CBN structure for the Malaria domain shown in Figure 2. Assume we have a dataset $\mathcal{D}_1$ that is faithful to the structure in Figure 1. A plausible predictive model for *Malaria* using for example, a decision tree or a rule learner is given below:

If *Intermittent Fever* and *Enlarged Spleen* Then *Malaria* Else Normal.

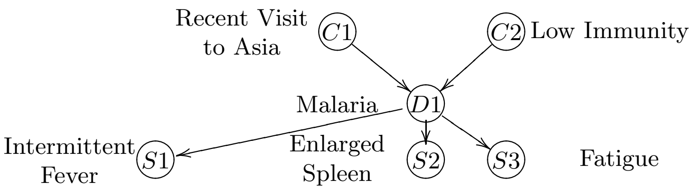


Fig. 1. A hypothetical causal Bayesian network structure for the Malaria domain

Let us call this rule the Fever Spleen (FS) rule. The FS rule would be a good predictor of Malaria. However, it is clear that guidelines to treat fever with aspirin or paracetamol will not have any effect on the distribution of Malaria.

On the other hand consider a causal relationship such as the following that may be induced from $\mathcal{D}_1$ by a causal discovery algorithm:

Recent Visit to Asia causally influences *Malaria.*

Based on this cause and effect relationship, we could propose a travel advisory warning people against traveling to Africa resulting in a reduction in Malaria in the population. While causal discovery methods focus on causal factors for proposing guidelines, predictive machine learning algorithms are known to select relevant but non-causal variables based on predictive accuracy. Moreover, the selected relevant variables may not be in the local neighborhood (direct predictors) of the class variable [5].

3 Framework for Clinical Practice Guidelines

In this section we describe a causal modeling framework using the representation of CBNs for three CPG specific tasks.

3.1 Learn the Causal Model of the Domain from Data and Propose New CPGs

Using the model shown in Figure 1 (assuming it is learned from $\mathcal{D}_1$ using a causal discovery algorithm), we could propose the following Malaria CPG:

> If *Recent Visit to Asia* and *Intermittent Fever* and *Enlarged Spleen*
> Then diagnose *Malaria*.

3.2 Learn the Causal Model from Practice Data and Recognize the Likely CPGs Being Followed

Assume we also have data (dataset $\mathcal{D}_2$) from a hospital that is following the Malaria CPG given in Section 3.1. A plausible causal model for $\mathcal{D}_2$ is shown in Figure 2. The variable $D1$ represents *Malaria Diagnosis* which is a deterministic node and functionally dependent on the three variables $C1$, $S1$ and $S2$. Note that the arc from $C2$ to $D1$ has been removed because of the functional dependency of $D1$ on $C1$, $S1$ and $S2$[1]. From this model it is possible to recognize the guideline being followed for the diagnosis of Malaria. It is known from medical domain knowledge that $S1$ (*Intermittent Fever*) and $S2$ (*Enlarged Spleen*) are symptoms (effects) of Malaria and not causes for Malaria. However, the model tells us that it is "causal" for the diagnosis of Malaria because of the Malaria CPG.

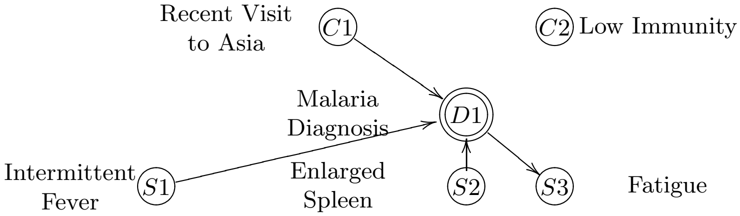


Fig. 2. A causal Bayesian network structure for Malaria diagnosis using guideline

It is not clear that a predictive rule or decision tree learner would identify the Malaria guideline because their goal is not to construct a model of the domain but maximize classification accuracy. Typical statistical and machine learning models seek to maximize predictive accuracy and will not in general generate the right causal model (see for example, [2, chapter 8] for related limitations of regression and [5] for causal limitations of SVMs). We note that causal validity is of the essence for the task of guideline compliance checking as discussed in Section 3.3.

Figure 3 shows the Malaria domain shown in Figure 1 augmented with the Malaria CPG. In Figure 3 node $D1$ denotes the disease Malaria and node $D2$ denotes Malaria diagnosis.

[1] On the other hand, if a variable X is manipulated (set randomly to one of its states), all the incoming arcs of X will be removed. The manipulation model is applicable to a randomized controlled trial (RCT).

3.3 Perform Compliance Checks for CPGs and Quantify the Degree of Compliance

Compliance check of the Malaria CPG can be done using the CBN in Figure 2. By instantiating the nodes $C1$, $S1$ and $S2$ and propagating the evidence through the network we can ascertain the probability of the disease given the evidence, that is, $\mathrm{P}(D1|C1, S1, S2)$. See [1,6] for a discussion of the algorithms for evidence propagation in a CBN. The quantification of compliance can be read from the conditional probability table for $D1$ in Figure 2 as the variables $C1, S1$ and $S2$ are the parents of $D1$. The estimated $\mathrm{P}(D1|C1, S1, S2)$ from $\mathcal{D}_2$ will give the degree of compliance for the Malaria CPG for the hospital from which $\mathcal{D}_2$ was obtained.

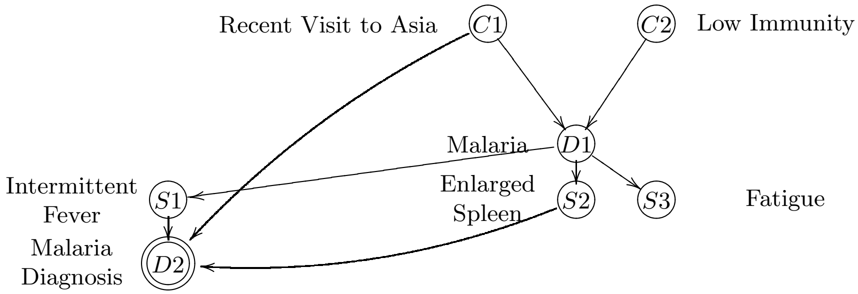


Fig. 3. An augmented causal Bayesian network structure for the Malaria diagnosis using guideline

3.4 Preliminary Results

Preliminary results using the causal discovery algorithm FCI [2, Chapter 6] on a population-based dataset for a high blood pressure (HBP) study output a causal model that incorporated the guideline used in the study:

> If *Outpatient Blood pressure* = 1 or *Blood Pressure medication* = 1, HBP = H; else HBP = N.

The causal model output by FCI for the HBP domain had directed arcs from *Outpatient Blood pressure* and *Blood Pressure medication* to HBP. See [7] for additional details.

4 Discussion and Conclusion

In this paper we presented a framework based on causal Bayesian networks for (1) *de novo* generation of cause and effect clinical practice guidelines from data, (2) recognition of a CPG from clinical data and (3) compliance checking and quantification of the degree of compliance of a known guideline. Note that traditional machine learning algorithms such as decision trees and rules will generate useful predictive models based on accuracy for the class variable (for example, diagnosis). However, they may not be robust under violations of iid (independent and identically distributed) and this may affect generalizability to other populations. Moreover, they will not be appropriate for prevention and intervention as they can neither model nor infer the causal interactions in the domain correctly. Predictive modeling techniques also cannot identify unobserved (hidden) variables or confounding.

Many practical problems exist before causal discovery methods such as the one discussed here will be able to routinely handle guideline discovery and compliance checking. An open problem is to understand how deterministic variables that result from application of clinical guidelines will impact guideline recognition from practice data using a CBN framework. Deterministic variables can result in violations of the faithfulness assumption that is typically required for causal discovery [8,9]. There are also situations where the guidelines being followed may be implicit and the focus in learning such implicit guidelines is to understand and record what causes decision makers to arrive at certain decisions. In [10] it is shown that physicians are often non-compliant to the gold-standard guideline when they believe they are actually implementing them correctly.

References

1. Pearl, J.: Probabilistic Reasoning in Intelligent Systems, 2nd edn. Morgan Kaufmann, San Francisco, California (1991)
2. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, 2nd edn. MIT Press, Cambridge, MA (2000)
3. Glymour, C., Cooper, G.F. (eds.): Computation, Causation, and Discovery. MIT Press, Cambridge, MA (1999)
4. Russo, F., Williamson, J.: Interpreting causality in the health sciences. International Studies in the Philosophy of Science (in press, 2007)
5. Statnikov, A., Hardin, D., Aliferis, C.: Using SVM weight-based methods to identify causally relevant and non-causally relevant variables. In: NIPS workshop poster (2006)
6. Lauritzen, S.L., Spiegelhalter, D.: Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society Series B 50, 157–224 (1988)
7. Mani, S., Aliferis, C., Krishnaswami, S., Kotchen, T.: Learning causal and predictive clinical practice guidelines from data. In: Proceedings of MedInfo, IOS Press, Amsterdam (in press, 2007)
8. Pearl, J., Geiger, D., Verma, T.: The logic of influence diagrams. In: Oliver, R.M., Smith, J.Q. (eds.) Influence diagrams, belief nets and decision analysis, pp. 67–87. Wiley, New York (1990)
9. Cooper, G.F.: An Overview of the Representation and Discovery of Causal Relationships Using Bayesian Networks. In: Glymour, C., Cooper, G.F. (eds.) Computation, Causation, and Discovery, pp. 3–62. MIT Press, Cambridge, MA (1999)
10. Sboner, A., Aliferis, C.: Modeling clinical judgment and implicit guideline compliance in the diagnosis of melanomas using machine learning. In: Proceedings of the AMIA fall symposium (2005)

Semantic Web Framework for Knowledge-Centric Clinical Decision Support Systems

Sajjad Hussain, Samina Raza Abidi, and Syed Sibte Raza Abidi

NICHE Research Group, Faculty of Computer Science, Dalhousie University,
Halifax, CANADA B3H 1W5
{hussain,abidi,sraza}@cs.dal.ca

Abstract. Lately, there have been considerable efforts to computerize Clinical Practice Guidelines (CPG) so that they can be executed via Clinical Decision Support Systems (CDSS) at the point of care. We present a Semantic Web framework to both model and execute the knowledge within a CPG to develop knowledge-centric CDSS. Our approach entails knowledge modeling through a synergy between multiple ontologies–i.e. a domain ontology, CPG ontology and patient ontology. We develop decision-rules based on the ontologies, and execute them with a proof engine to derive CPG-based patient specific recommendations. We present a prototype of our CPG-based CDSS to execute the CPG for Follow-up after Treatment for Breast Cancer.

Keywords: Clinical Decision Support Systems, Semantic Web, Ontologies, Clinical Practice Guidelines, Breast Cancer.

1 Introduction

Clinical Practice Guidelines (CPG) are evidence-based recommendations to assist clinical decision-making [1]. Clinical Decision Support Systems (CDSS), developed using CPG-mediated knowledge, can offer the functionality to (a) Execute the CPG at the point of care; (b) Guide healthcare practitioners to make evidence based decisions, actions and recommendations; (c) standardize the delivery of care at a particular healthcare setting; and (d) Collect all necessary and relevant patient data.

From a technical perspective, the key challenges in developing CPG-guided CDSS are: (i) The capture of the disease-specific knowledge inherent within a CPG, whilst maintaining the underlying semantics, avoiding ambiguities and identifying the key decisional elements; (ii) The specification of the interactions between the key CPG elements to realize an executable CPG plan; (iii) The transformation of the CPG decision logic into medically salient and executable logic-based decision rules; (iv) The execution of a computerized CPG based on both acquired and inferred patient information; and (v) The explanation of the CDSS recommendations in order to establish 'trust' with the user.

The emerging Semantic Web [2] approach offers interesting and pragmatic solutions to model CPG for developing CPG-guided CDSS. The Semantic Web purports the semantic modeling and markup of knowledge, its properties and its relations using well-defined semantics such as formal definitions of terms, ontological

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 451–455, 2007.

definitions of domain concepts and resources. For developing CDSS, the Semantic Web offers a logic-based framework to (a) semantically model the structure of a paper-based CPG to a semantically enriched formalism, such as a CPG ontology; (b) annotate the CPG, based on the CPG ontology elements, in terms of the Resource Description Framework (RDF); (c) represent the underlying clinical concepts and relationships inherent within a CPG in terms of a domain ontology, using the Web Ontology Language (OWL); (d) represent the different patient data sources in a semantically enriched formalism; (e) specify the CPG based decision-making logic in terms of symbolic rules, represented as N3 triples, that can be executed using proof engines; (f) ensure interoperability between multiple ontologically defined knowledge resources; and (g) provide a justification trace of the inferred recommendations.

In this paper, we present a Semantic Web based framework for computerizing CPG to develop a CDSS. We demonstrate the modeling of a Breast Cancer Follow-up (BCF) CPG, leading to the development of a BCF-CDSS that is deployed at the point of care in primary care settings. Figure 1 shows the architecture of our Semantic Web based CDSS. Our CDSS development framework constitutes three key elements:

a) *Modelling* the overall declarative and procedural knowledge required for decision support. We developed three independent, yet interacting ontologies as follows: (i) A CPG Ontology that models the computerized structure of the CPG in terms of the Guideline Element Model (GEM) [3]; (ii) A Domain Ontology that models the medical knowledge pertaining to the domain of the CPG. The Domain Ontology represents both the CPG's concepts and relationships between these concepts as OWL classes and properties, respectively; and (iii) A Patient Ontology that models the patient in terms of the longitudinal medical record of patient. We used Protégé [4] to develop all the ontologies.
b) *Authoring* of CPG-medicated decision rules using the CPG Rule Authoring Module. We developed a simple rule syntax to author rules.
c) *Execution* of the CPG to provide case-specific decision support. The CPG Execution Module, built using the JENA–a logic-based proof engine [5], allows the execution of decision logic rules based on patient data. A novel feature of our CDSS is that it provides a justification trace of all inferred recommendations.

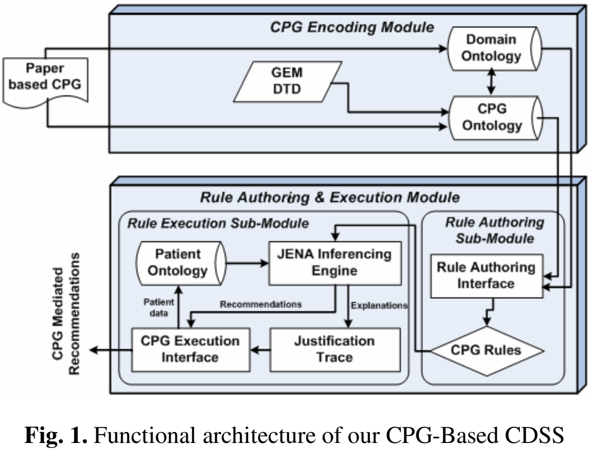


Fig. 1. Functional architecture of our CPG-Based CDSS

2 Clinical Practice Guideline Modeling Via Ontologies

The Canadian Steering Committee on CPG for the Care and Treatment of Breast Cancer (BC) has developed and updated the guideline on follow-up care after treatment for BC [6], with a special emphasis on the needs of primary care physicians. The challenge was to operationalize the BCF-CPG within the family physician's clinical workflow so that BC follow-up can be offered in a primary care setting.

Computerization of the BCF-CPG was achieved using GEM, whereby the main task was to determine the function of a specific CPG text and annotate it using the relevant GEM tag. Next, we developed a *Domain Ontology* that models the knowledge encapsulated within the BCF-CPG. We used Protégé ontology editing environment to build our BC ontology using Protégé OWL. We defined twelve main classes, namely; *Patient_Type*, *Physician_Type*, *Illness*, *Menstrual_Status*, *Recommendation*, *Symptom*, *Diagnostic_Tests*, *Treatment*, *Age*, *Risk_Factor*, *Weight_Status* and *Patient_Wish*. Full details of the BC ontology are reported in [7].

3 CPG Modeling Module

We developed a CPG encoding tool that computerizes a paper-based CPG so that it can be executed by our CPG execution engine. The CPG encoding module comprises two components: (a) A GEM-based CPG representation formalism to convert the paper-based CPG into an electronic format (described earlier); and (b) A CPG ontology to model the structure of the CPG.

Our CPG Ontology is based on the GEM DTD [3]. The main CPG knowledge is represented in the Knowledge Component (KC) class in the CPG Ontology as they describe the procedural, conditional or imperative knowledge (as shown in Figure 2). We defined a *Recommendation* class to describe the recommended actions that are classified as being either imperative or conditional. Imperative recommendations are applicable to the entire eligible population, whereas conditional recommendations describe the clinical conditions/scenarios that demand specific actions. We represent these clinical conditions as *decision.variable* class in the CPG Ontology. We also made use of the property logic to define the decision logic for all recommendations based on the conditions for the various actions. Each *decision.variable* instance with a property *variable.name* is annotated with a property from the Domain Ontology.

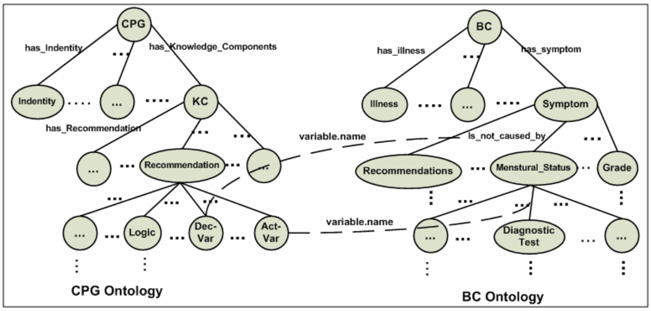


Fig. 2. Relation between CPG and Domain Ontology

4 CPG Authoring and Execution Module

The CPG Rule Authoring and Execution module is designed to encapsulate the clinical decision logic inherent within a CPG in terms of logical rules–such logical rules are executed by a reasoning engine to derive CPG-based recommendations. To achieve the above functionality, we built three sub-modules as follows:

4.1 Rule Authoring Sub-module

Rule Authoring is performed by defining decision rules in the logic tag of CPG ontology as follows: Step 1: Select decision variables, which represents the body (premises) of the rule; Step 2: Select the action variable, which represents the head (conclusion) of the rule; Step 3: For each decision variable and action variable in the rule, an equality/inequality relation can be defined with either a variable, a value, a binary algebraic formula, another decision variable or list of decision variables. An example conditional recommendation from the BCF-CPG "*When such bleeding (Vaginal Bleeding) is present in the absence of obvious cause, endometrial biopsy should be carried out",* can be defined in terms of a CPG decision rule via following rule authoring steps:

Step # 1: First, we identified conditions in the above recommendation (marked as underline), and defined them as decision variables *dv1=has_symptom*, *dv8=is_not_cause_by*, *dv9=ms_apply_to_diagnostic_test* and *a1=is_Recommended*
Step # 2: Then, we defined a rule in the logic tag of CPG Ontology by selecting decision and action variables that serve as conditions/premises and conclusion, respectively.

IF dv1, dv8, dv9 THEN a1

Step # 3: Finally, we defined this rule for a general case by quantifying decision variables for all such scenarios (represented as ?), as follows:

IF dv1=?,dv8=?,dv9=? THEN a1=[dv9]

Upon completion of the rule authoring process, we apply a rule transformation algorithm to transform the CPG rules into the JENA syntax so that they can be executed in the Execution Sub-Module that uses the JENA inference engine [5]. The above CPG rule R is transformed into JENA syntax via our rule transformation algorithm as follows:

Transform (R) = [conditional1: (?X2 bc:has_symptom ?X5), (?X5 bc:is_not_cause_by ?X1), (?X1 bc:ms_apply_to_diagnostic_test ?X6) -> (?X2 bc:is_Recommended List(?X6))]

4.2 Execution Sub-module

The Execution Sub-Module invokes the JENA inference engine to execute a CPG—i.e. infer recommendations based on patient data. We model instances from the Domain Ontology, CPG Ontology and Patient Ontology as RDF graphs, which serve as the knowledge base for JENA. The JENA inference engine employs backward reasoning to infer CPG-mediated recommendations based on the given patient scenario, encoded clinical knowledge in the Domain Ontology and CPG Ontology.

4.3 Justification Trace Sub-module

Justification Trace Sub-Module generates a justification trace of the rule execution to assist medical practitioners in understanding the rationale behind the inferred recommendations. The justification derivation includes the linear representation of CPG rules that were satisfied to derive the stated recommendation. The justification trace initiates with an inferred patient recommendation (derived facts) and generates facts which served as premises for deriving the patient recommendation, recursively. The process terminates, if all the premises are ground instances (known facts). Below we show the justification trace for recommending *Endometrial Biopsy* for a BC patient *Jane*, who has symptom of *Vaginal Bleeding.*

```
Jane --> has_symptom = Vaginal_Bleeding
Vaginal_Bleeding --> is_not_cause_by = Mensturation_or_Obvious_Cause
Mensturation_or_Obvious_Cause --> ms_apply_to_diagnostic_test = Endometrial_Biopsy
Jane --> is_Recommended = Endometrial_Biopsy
```

5 Concluding Remarks

In the realm of healthcare knowledge management the modeling of CPG provides interesting opportunities to develop CDSS that support evidence-guided recommendations. In this paper, we demonstrated the application of the Semantic Web to integrate multiple ontologies to develop a CDSS. Our CDSS approach is quite generic and can be extended to other medical problems. We tested our CDSS with a number of real-life clinical cases and both the recommendations and their justifications were validated by medical practitioners.

Acknowledgement. This project is supported by a research grant from Nova Scotia Health Research Foundation. Thanks to Dr. Grunfeld for the Breast Cancer CPG [6].

References

1. Field, M.J., Lohr, K.N. (eds.): Clinical Practice Guidelines: Directions for a New Program, Institute of Medicine, National Academy Press, Washington, DC (1990)
2. Berners-Lee, T.: The semantic web. Scientific American 284(5), 34–43 (2001)
3. 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. J Am Med Informatics Assoc 7, 488–498 (2000)
4. Knublauch, H., Fergerson, R., Noy, N.F., Musen, M.: The Protégé OWL Plugin: An open development environment for semantic web applications. In: McIlraith et al., pp. 229–243
5. JENA: Semantic Web Framework http://jena.sourceforge.net/documentation.html
6. Grunfeld, E., Dhesy-Thind, S., Levine, M.: Clinical practice Guidelines for the Care and Treatment of Breast Cancer: 9. Follow-up After Treatment for Breast Cancer (2005 update) Can. Med. Assoc. J 172, 1319–1320 (2005)
7. Abidi, S.: Ontology-based Modeling of Breast Cancer Follow-up Clinical Practice Guideline for Providing Clinical Decision Support. In: 20th IEEE Symposium on Computer-Based Medical Systems, Maribor, Slovenia, June 20-22, 2007, IEEE Press, Los Alamitos (2007)

Inference in the Promedas Medical Expert System

Bastian Wemmenhove[1], Joris M. Mooij[1], Wim Wiegerinck[1], Martijn Leisink[1], Hilbert J. Kappen[1], and Jan P. Neijt[2]

[1] Department of Biophysics, Radboud University Nijmegen, 6525 EZ Nijmegen, The Netherlands
[2] Internal Medicine, University Hospital Utrecht Utrecht, The Netherlands

Abstract. In the current paper, the Promedas model for internal medicine, developed by our team, is introduced. The model is based on up-to-date medical knowledge and consists of approximately 2000 diagnoses, 1000 findings and 8600 connections between diagnoses and findings, covering a large part of internal medicine. We show that Belief Propagation (BP) can be successfully applied as approximate inference algorithm in the Promedas network. In some cases, however, we find errors that are too large for this application. We apply a recently developed method that improves the BP results by means of a loop expansion scheme. This method, termed Loop Corrected (LC) BP, is able to improve the marginal probabilities significantly, leaving a remaining error which is acceptable for the purpose of medical diagnosis.

1 Introduction

In this paper we present the Promedas medical diagnosis model. It is an expert system for doctors based on a Bayesian network structure for which the calculation of marginal probabilities is tractable for many cases encountered in practice. For those cases that are intractable (i.e. a junction tree algorithm is not applicable), alternative algorithms are required. A suitable candidate for this task is Belief Propagation (BP), which is a state-of-the art approximation method to efficiently compute marginal probabilities in large probability models [1,2]. Over the last years, BP has been shown to outperform other methods in rather diverse and competitive application areas, such as error correcting codes [3,4], low level vision [5], combinatoric optimization [6] and stereo vision [7].

In medical expert systems, so far the success of BP has been limited. Jaakkola and Jordan [8] successfully applied variational methods to the QMR-DT network [9] but BP was shown not to converge on these same problems [2]. We find that BP does converge on all Promedas cases studied in the current paper. Although this does not guarantee convergence in all possible cases, we note that double loop type extensions to BP [10] may be applied when convergence ceases. Here we compute the marginal errors of BP and apply a novel algorithm, termed Loop Corrected Belief Propagation (LCBP) [11] to cases in which the error becomes unacceptable. We argue that this method potentially reduces the error to values acceptable for medical purposes.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 456–460, 2007.

Recently a company was founded that uses the Promedas network to develop a commercially available software package for medical diagnostic advise. A demonstration version can be downloaded from the website www.promedas.nl. The software will become available as a module in third party software such as laboratory or hospital information systems or stand alone designed to work in a hospital network to assist medical specialists. In all cases the software will be connected to some internally used patient information system. This year the Promedas software will be available via a web portal as well. This might be operational at the time of the AIME congress. Physicians can visit the website, enter medical characteristics of a specific case and immediately obtain a list of most probable diagnoses. The Promedas web portal uses the full available database of diagnoses and findings.

2 Inference in the Promedas Graphical Model

The global architecture of the diagnostic model in Promedas is similar to QMR-DT [9]. It consists of a diagnosis-layer that is connected to a layer with findings. Diagnoses (diseases) are modeled as a priori independent binary variables $d_j \in \{0, 1\}$, $j \in \{1, \ldots, N_{\mathrm{D}}\}$, causing a set of symptoms or findings $f_i \in \{0, 1\}$. In the user interface, a significant part of the findings are presented as continuous variables. These are discretized in a medically sensible way. The interaction between diagnoses and findings is modeled with a noisy-OR structure, indicating that each parent j has an individual probability of causing a certain finding i to be true if it is in the parent set $V(i)$ of i, and there is an independent probability λ_i that the finding is true without being caused by a parent (disease). Thus

$$\begin{aligned} p(f_i = 0|\mathbf{d}) &= [1 - \lambda_i] \prod_{j \in V(i)} [1 - w_{ij} d_j] \\ p(f_i = 1|\mathbf{d}) &= 1 - p(f_i = 0|\mathbf{d}) \end{aligned} \tag{1}$$

The parameters $\{\lambda_i\}, \{w_{ij}\}$, together with the disease prevalences (ranging from 0.001 to 0.1) are the model parameters determined by the medical experts. The disease nodes are coupled to risk factors, such as, e.g., concurrent diagnoses and nutrition. Risk factors are assumed to be observed and to modify the prevalences of the diagnoses. From a database of model parameters, the graphical model and a user-interface for Promedas are automatically compiled. This automatic procedure greatly facilitates changes in the model, such as adding or removing diseases, as required in the design phase of the model. Once the graphical model has thus been generated, we use Bayesian inference to compute the probability of all diagnoses in the model given the patient data. Before computation, we remove all unclamped (i.e. unobserved) findings from the graph, and we absorb negative findings in the prevalences [8]. Only a network of positively clamped findings and their parents remain. Using standard techniques for the calculation of posterior distributions directly on the factor graph in the above representation, either with a junction tree algorithm ([12]) or approximation techniques, is limited to cases in which the size $|V(i)|$ of the interaction factors is not too large. In Promedas, however, sets containing 30 nodes (i.e. findings that may have 30 different causes) are not uncommon. Thus it is helpful to reduce the maximum number of members of factor potentials, which may be achieved by adding extra (dummy) nodes to the graph [13,14]. The version of Promedas that is studied in this paper contains over 10000 variables, including about 2000 diagnoses, and 8600 connections between diagnoses and findings.

Despite these measures computation can still be intractable when the number of positive patient findings becomes large [8]. In that case, we must resort to approximations. The feasibility of this approach is studied in the remainder of the paper. In the next section we first report results of applying Belief Propagation to a number of "virtual patient" cases, followed by tests of a version of the recently developed Loop Corrected Belief Propagation [15,11] algorithm on these cases. The idea of LCBP can be understood as follows. BP is a method that is exact on graphs that are tree-like. This means that if one removes a node from the graph, the probability distribution on its neighbors (the so-called cavity distribution) factorizes. When BP is applied to graphs with loops, this is no longer true and the cavity distribution contains correlations. The LCBP method incorporates estimates of these correlations in a message passing scheme. For more details see [11].

3 Simulations with Virtual Patient Data

Using the model first as a generator of virtual patients, we generated, 1000 patient cases with $N_\mathrm{d} = 1$ and another 1000 with $N_\mathrm{d} = 4$ where N_d represents the number of randomly selected true diseases for the generation of patient data. The first result we report is the fact that on all cases that we generated BP converged. This contrasts with previous results by Murphy et. al. [2], found for the QMR-DT network, where the small prior probabilities seemed to prevent convergence in a couple of complex cases. The maximal marginal errors in the BP results are typically small but may occasionally be rather large. In fig. 1 left we plot the error versus the tree width of the JT method, which is an indication of the complexity of the inference task. From fig. 1 left we conclude that the quality of the BP approximation is only mildly dependent on this complexity. It follows that for patient cases where exact computation is infeasible, BP gives a reliable alternative for most cases. To avoid cases where the error is unacceptably large we propose to use the so-called Loop Corrected BP method.

The right picture of fig. 1, displays results of applying LCBP to a set of 150 $N_d = 1$ virtual patient cases. Horizontally, the maximal error in BP single node marginals is plotted, and vertically the maximal error after applying the loop correction scheme. Only cases with nonzero error (i.e. loopy graphs) are plotted, 86 in total. The maximal error in the marginals produced by BP typically reduces one order of magnitude after applying LCBP. The largest maximum error over all cases in this sample reduced from 0.275 to 0.023.

As a second test, we applied the method to a few cases in the left picture of figure 1, where we attempted to reduce large BP errors of these complex multiple

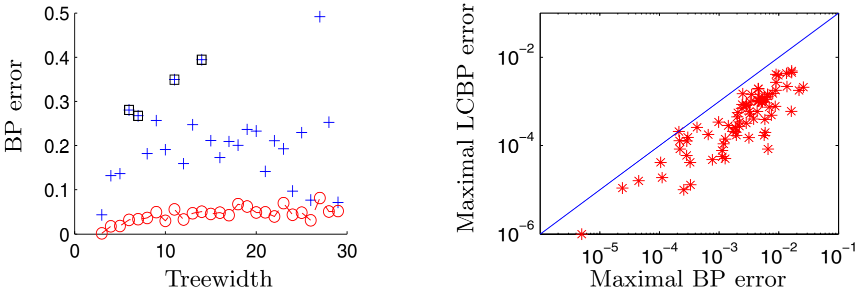


Fig. 1. *Left*: BP maximal error(○) averaged over instances, largest maximal error (+) as a function of treewidth for $N_d = 4$. The squares mark instances with large error which we have later subjected to LCBP (see table 1). *Right*: $N_d = 1$ Maximal single node marginal error of LCBP (vertical) versus BP (horizontal). All data lie on the side of the line where the LCBP error is smaller than the BP error.

disease errors. A drawback of the current implementation of LCBP is its rather large computation time when Markov blanket sizes grow large. For the implementation of LCBP that we used, computation time grows as N^2 (assuming constant maximal degree per node), but also grows exponentially in the number of nodes in the largest Markov blanket. The exponential scaling of the algorithm in Markov blanket size forced us to look at a few relatively easy cases only. Results for the BP errors marked by a black square in figure 1 are reported in table 1:

Table 1. LCBP results on complex instances with large errors

Treewidth	rms error BP	max error BP	rms error LCBP	max error LCBP
6	0.0336	0.2806	0.0021	0.0197
7	0.0429	0.2677	0.0017	0.0102
11	0.0297	0.3494	intractable	intractable
14	0.0304	0.3944	0.0011	0.0139

The maximal error of LCBP clearly reduces to acceptable levels, but the computation time is prohibitive for complex cases. The solution to this problem may be an alternative implementation, taking into account only nontrivial correlations between pairs of variables in the Markov blanket (see [15]), and consequently scales polynomially in the Markov blanket size. We did not consider this algorithm in the current investigation, since its implementation is much more involved, but the promising results obtained here motivate us to do so in the future.

4 Conclusions

In this paper we have shown that BP is an attractive alternative for exact inference for complex medical diagnosis inference tasks. In some isolated instances, BP produces large errors and we have shown that loop corrected BP can significantly reduce these errors. Therefore, for practical purposes it seems worthwile to further develop an efficient version of LCBP that scales polynomially in the Markov-blanket size, such as the one proposed in [15].

References

1. Pearl, J.: Probabilistic reasoning in intelligent systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco, California (1988)
2. Murphy, K.P., Weiss, Y., Jordan, M.I.: Loopy belief propagation for approximate inference: An empirical study. In: Proceedings of Uncertainty in AI, pp. 467–475 (1999)
3. Gallager, R.G.: Low-density parity check codes. MIT Press, Cambridge (1963)
4. McElice, R., MacKay, D., Cheng, J.: Turbo decoding as an instance of pearl's belief propagation algorithm. Journal of Selected Areas of Communication 16, 140–152 (1998)
5. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. Int. J. Comp. Vision 40, 25–47 (2000)
6. Mezard, M., Parisi, G., Zecchina, R.: Analytic and algorithmic solution of random satisfiability problems. Science 297 (2002)
7. Sun, J., Li, Y., Kang, S.B., Shum, H.-Y.: Symmetric stereo matching for occlusion handling. Proceedings CVPR 2, 399–406 (2005)
8. Jaakkola, T., Jordan, M.I.: Variational probabilistic inference and the QMR-DT network. Journal of artificial intelligence research 10, 291–322 (1999)
9. Shwe, M.A, Middleton, B., Heckerman, D.E., Henrion, M., Horvitz, E.J., Lehman, H.P., Cooper, G.F.: Probabilistic Diagnosis Using a Reformulation of the Internist-1/ QMR Knowledge Base. Methods of Information in Medicine 30, 241–255 (1991)
10. Heskes, T., Albers, K., Kappen, H.J.: Approximate inference and constraint optimisation. In: Proceedings UAI, pp. 313–320 (2003)
11. Mooij, J.M., Wemmenhove, B., Kappen, H.J., Rizzo, T.: Loop corrected belief propagation. In: Proceedings of AISTATS (2007)
12. Lauritzen, S.L., Spiegelhalter, D.J.: Local computations with probabilties on graphical structures and their application to expert systems. J. Royal Statistical society B 50, 154–227 (1988)
13. Heckerman, D.: A tractable inference algorithm for diagnosing multiple diseases. In: Proceedings UAI, pp. 163–171. Elsevier, Amsterdam (1989)
14. Takinawa, M., D'Ambrosio, B.: Multiplicative factorization of noisy-MAX. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence UAI99, pp. 622–30 (1999)
15. Montanari, A., Rizzo, T.: How to compute loop corrections to the bethe approximation. Journal of Statistical Mechanics P10011 (2005)

Computerised Guidelines Implementation: Obtaining Feedback for Revision of Guidelines, Clinical Data Model and Data Flow

S. Panzarasa[1], S. Quaglini[2], A. Cavallini[3], S. Marcheselli[3],
M. Stefanelli[2], and G. Micieli[4]

[1] CBIM, Pavia
[2] Dept of Computer Science and Systems, University of Pavia
[3] IRCCS Foundation "C. Mondino", Pavia
[4] IRCCS Istituto Clinico "Humanitas", Rozzano (MI), Italy

Abstract. In this paper we describe a module that allows to collect (a) motivations for non-compliance to guidelines, (b) motivations for poor data entry into the electronic patient record, and (c) comments on medical aspects of guideline recommendations, on their formalisation into computerised rules, and on the guideline integration into the computerised clinical chart. We organised a well-structured taxonomy of non-compliance motivations in such a way that the main hierarchical levels correspond to different medical or technical roles suitable for feedback managing. We analysed about 400 consecutive cases of patients with ischemic stroke. About 40 non-compliances, as well as several incomplete data forms have been identified and motivated.

1 Introduction

Despite wide diffusion of clinical practice guidelines (GLs), there is often a feeling of a poor implementation degree. In our opinion, one cause could be the few attention on the final users' feedback. Comments on the practical use of GLs are limited to discussions among end-users, rarely collected systematically to be reported at higher levels (e.g. consensus conferences, expert panels, etc.) and to be used for GLs revisions. Computerised GLs suffer from this problem too, because they mainly focus on representing medical knowledge and running it on patient's data, while tools for systematic collection of users' comments are not widespread. Problems related to the functionality of the computerized system are solved through isolated, ad-hoc interventions without keeping track of them. Levels of use of the software are often low [1], and becomes lower and lower if technical and organizational problems persist and computer-based systems are perceived by the users as an extra working-load. We think that any GL implementation, either computer-supported or not, needs a tool for users' feedback collection. Of course, if the GL is computerized, the task is facilitated and the tool may be integrated in the decision support system itself. In such a way, systematic data collection on systems usability is more guaranteed. Feedbacks may be used at different levels, e.g. towards medical teams, developers of computerized clinical chart, responsibles of the workflow management systems, and the experts team caring for GL development, revision and

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 461–466, 2007.

dissemination. We show some results from the Italian SPREAD GLs [2] implementation in a Stroke Unit (SU). The paper illustrates some paradigmatic exemplars to show how they can be used for improving the whole careflow (Cf) and the GLs themselves.

2 The Careflow Management System

Theoretical basis of careflow management systems (CfMSs) and the application implemented in the SU of the Mondino Hospital have been extensively described previously [3,4]. Briefly, the GL recommendations have been implemented through a Workflow model with Oracle Workflow^{TM}. A Computerised Clinical Chart (CCC), developed with Wincare®[1], was used in the SU since many years. A middleware layer developed in PL/SQL allows the communication between the CfMS and the CCC, and attention has been put in maintaining the end-user interface as much as possible unchanged, so that users perceive the new system just as an update of CCC, with some new functionalities, that are: a) the sections of the clinical chart waiting for new data are listed runtime in an "intelligent", patient-specific, dynamic way; b) data relevant for interpretation of the GL rules appear as yellow field ("*yellow-data*"), in order to facilitate a complete record filling.; c) according to the urgency of the recommendation, messages and reminders are shown runtime directly on the screen or through a communication box. The decision support system has been installed in April 2006, but the data model of the CCC is the same from January 1st 2005. Since that time to mid January 2007, about 400 ischemic stroke patients have been admitted to the SU. In a recent work [5] we described the preliminary results. This real-time decision support system has been designed to be as less intrusive as possible. Thus, in no case the system asks in real-time for non-compliance (NC) motivation or other comments. On the other hand, collecting these motivations is extremely important. Thus we decided to manage NC in a less invasive way, as illustrated in the next sections.

3 Problems Encountered During Implementation

The system malfunctioning or "user-perceived" malfunctioning, reported during the very first period of the system enactment, must be promptly solved, to remove confounding factors that could bias further analysis of NC. Different types of problems have been encountered during the daily use of the system indeed:

- **Timely data sharing** - Communication between physicians, that prescribe the therapy, and nurses, that administer it, is normally paper-based, and this implies late data entry in the CCC.
- **Wrong rule formalization** - Let us take the SPREAD recommendation *"In patients at high risk of deep venous thrombosis (DVT) (i.e. presenting with plegic limbs, reduced consciousness, obesity, previous lower-limb venous diseases) prophylaxis with … heparin … is recommended starting since hospital admission."* "High risk of DVT" was at first represented as the OR of the listed risk factors, but this produced several false positive results: a paradigmatic case was a patient with a very high

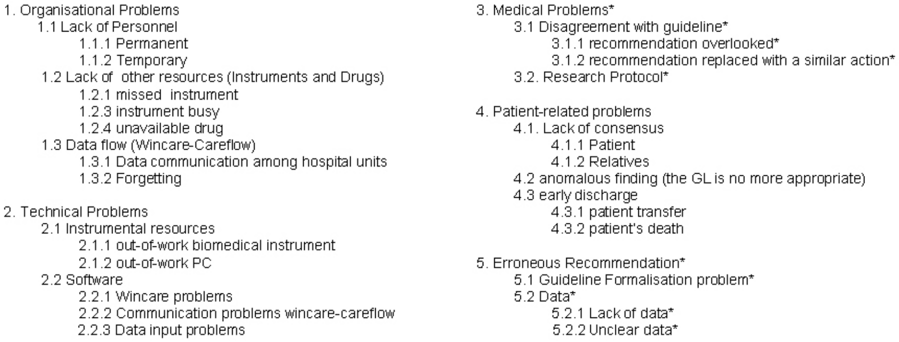
[1] Wincare® is a product by TSD (Pedrino di Vignate-Milan, Italy).

Body Mass Index but a normal mobility. This observation led to re-think about formalization of the rule: “obesity” must be read as “obesity impairing mobility”.

- **Incomplete data** - Erroneous system suggestions may be due to incomplete data, i.e. data that are known to the physicians but not stored in the CCC yet.
- **Graphical Interface issues -** The system and the user interface were designed to be minimally intrusive, but sometimes this causes users ignoring the system hints.

4 The Problems Capturing Module

Implementing ad-hoc solutions for any problem that we were encountering was not efficient, therefore we decided to implement a unique module able to: a) capture and report problems; b) ask for motivations or comments; c) store results; d) analyse data to elaborate a solution strategy, provide organisational feedback for healthcare administrators, provide medical feedback for the SU team and GL reviewer and provide technical feedback for CCC and Cf developers. Physicians decided to accomplish this documentation task at the patient’s discharge, when they summarize the patient’s hospital stay, and reason about: that's why we labelled such module RoMA (Reasoning on Medical Actions). For each patient, the module: a) checks the completeness of yellow data; b) lists the recommendations the patient was eligible for; c) lists the NCs.



1. Organisational Problems
1.1 Lack of Personnel
1.1.1 Permanent
1.1.2 Temporary
1.2 Lack of other resources (Instruments and Drugs)
1.2.1 missed instrument
1.2.3 instrument busy
1.2.4 unavailable drug
1.3 Data flow (Wincare-Careflow)
1.3.1 Data communication among hospital units
1.3.2 Forgetting

2. Technical Problems
2.1 Instrumental resources
2.1.1 out-of-work biomedical instrument
2.1.2 out-of-work PC
2.2 Software
2.2.1 Wincare problems
2.2.2 Communication problems wincare-careflow
2.2.3 Data input problems

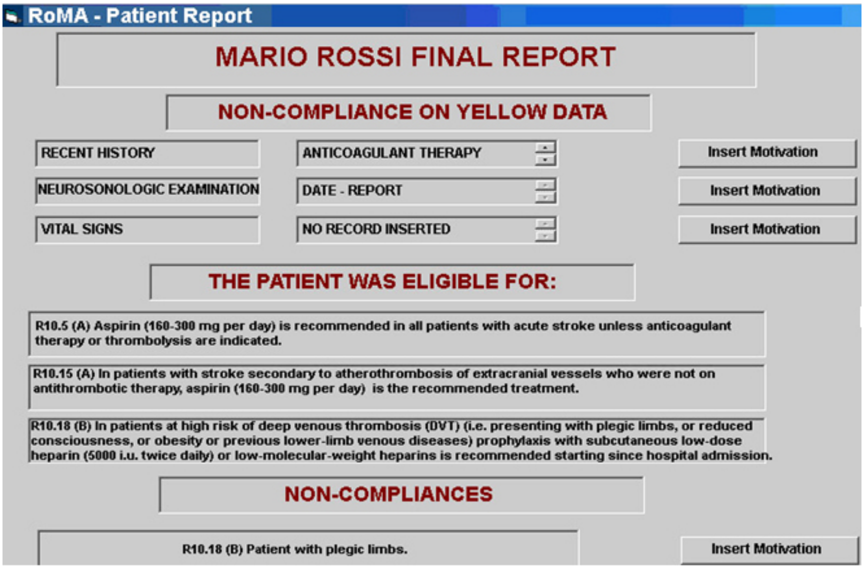
3. Medical Problems*
3.1 Disagreement with guideline*
3.1.1 recommendation overlooked*
3.1.2 recommendation replaced with a similar action*
3.2. Research Protocol*

4. Patient-related problems
4.1. Lack of consensus
4.1.1 Patient
4.1.2 Relatives
4.2 anomalous finding (the GL is no more appropriate)
4.3 early discharge
4.3.1 patient transfer
4.3.2 patient's death

5. Erroneous Recommendation*
5.1 Guideline Formalisation problem*
5.2 Data*
5.2.1 Lack of data*
5.2.2 Unclear data*

Fig. 1. The taxonomy of the motivations for a non-compliance to the GL or for missing *yellow data*. Starred items are only related to non-compliance motivations.

In case of uncompleted data or NC, the user may provide a motivation. We developed a taxonomy of these motivations, illustrated in Figure 1. The user may select any level of the taxonomy and also provide comments in free text. The more general levels (i.e. Organisational, Technical, Medical, Patient-related Problems and Erroneous Recommendation) have been devised in such a way to facilitate the feedback communication to the right people. Of course, more specific the level, more appropriate can be the feedback (i.e. a recurrent “out of work” of an instrument will be notified to the clinical engineer department). Figure 2 shows the result of a patient’s record check. Concerning “yellow data”, different messages are shown if a data form has never been filled or if a data form is filled but at least one of its yellow fields is lacking. Recommendations the



MARIO ROSSI FINAL REPORT

NON-COMPLIANCE ON YELLOW DATA

RECENT HISTORY	ANTICOAGULANT THERAPY
NEUROSONOLOGIC EXAMINATION	DATE - REPORT
VITAL SIGNS	NO RECORD INSERTED

THE PATIENT WAS ELIGIBLE FOR:

R10.5 (A) Aspirin (160-300 mg per day) is recommended in all patients with acute stroke unless anticoagulant therapy or thrombolysis are indicated.

R10.15 (A) In patients with stroke secondary to atherothrombosis of extracranial vessels who were not on antithrombotic therapy, aspirin (160-300 mg per day) is the recommended treatment.

R10.18 (B) In patients at high risk of deep venous thrombosis (DVT) (i.e. presenting with plegic limbs, or reduced consciousness, or obesity or previous lower-limb venous diseases) prophylaxis with subcutaneous low-dose heparin (5000 i.u. twice daily) or low-molecular-weight heparins is recommended starting since hospital admission.

NON-COMPLIANCES

R10.18 (B) Patient with plegic limbs.

Fig. 2. The output of the RoMA module for a specific patient

patient was eligible for are listed with their scientific evidence degree (decreasing from A to D). Clicking on the rule gives access to the right SPREAD paragraph on the SPREAD web site, thus providing an educational aspect. NCs are shown and physicians may enter motivations: clicking the button, the taxonomy in Figure 1 is open in selection mode. Users may enter a comment also when fully compliant, because compliance could be due to imposition, or protection against possible malpractice-related complaints [6].

5 Results

We collected the motivations for the 40 detected NCs, and comments that physicians wrote in 26% of the cases. The following distribution of NC types have been derived: Patient-related Problems (17%); Technical Problems (18%); Medical Problems (24%); Erroneous Recommendation (41%). It must be said that only two SPREAD chapters have been implemented, namely the diagnosis and treatment in the acute phase, thus NC number do not provide a complete view of the physicians' adherence to GLs. Most of NC highlighted a wrong rule formalisation, which may derive from either mistakes in the text interpretation, or by ambiguities or omissions in the text itself. One example has already been mentioned in section 3. Another one refers to recommendation related to aspirin, that does not mention "intolerance" as a contraindication. This rises a more general issue: we cannot pretend that *everything* is written in a GL. Some rules are so obvious to be taken for granted. This is the case of drug intolerance and allergies: a module checking for such events must be arranged for every rule dealing with drugs. Other NCs were in fact false-positive results, mainly caused by lack of a well-structured therapy record. In fact, 50% of the physicians' comments refer to therapy data that have not been recognised by the system. This result has been an important message for Wincare® company and it is now used to redesign part of the data model and user's interface. Additional false positive results were related to patients died during the hospital stay: from the physicians' comments we realised that, in case of death, only the discharge form (useful to administrative purposes), but not the discharge letter, is filled. Since such letter is an important source of information for our system, we missed data for these patients. The solution is to move all the yellow data from the discharge letter to the discharge form. This could also be useful for easier data retrieval in case of complaints from the patient's relatives. Among medical problems, the most frequent one (80%, 32 NCs) has been the disagreement with recommendation for DVT risk. We already remarked that this rule is not very clear but, even when there is no doubt on risk factors for DVT, physicians disagree with recommended administration of heparin. Most of NCs were for patients with blood in their brain. In these cases, several physicians prefer to not administer anticoagulants, even if scientific evidence shows that low doses/low molecular weight heparin does not have a significant hemorrhagic effect. All these comments are extremely important for the GL future versions and the next SPREAD consensus conferences will take into account these notes to produce more clear and convincing GLs. Concerning incomplete data input, the most common cause was impossibility of directly coding data from laboratory and radiology departments into the CCC.

6 Conclusions and Future Work

As soon as the decision support system was installed in the clinical routine, we realised that users were highlighting not only technical problems but also (and mainly) problems related to the GL implementation. Interpretation of the GL text that was carried out together by physicians and knowledge engineers required revisions, showing that GL text is often ambiguous or incomplete. Moreover, some recommendations have been shown to be hardly accepted by physicians, even if they are evidence-based. The lesson learned is that problems of GL implementation may be detected only when *real cases* are considered, and solutions are greatly facilitated by a systematic documentation of problems and collection of users' comments. Results of this study will be of great advantage for the next release of the SPREAD GLs.

References

1. Eccles, M., McColl, E., Steen, I., Rousseau, N., Grimshaw, J., Parkin, D., Purves, I.: Effect of computerized evidence based guidelines on management of asthma and angina in adults in rimary care: cluster randomized controlled trial. British Medical Journal 525, 941–944 (2002)
2. The Stroke Prevention and Educational Awareness Diffusion (SPREAD) Collaboration. The Italian Guidelines for stroke prevention. Neurol Sci 2000, 21, 5-12, last version (2005) at http://www.spread.it
3. Panzarasa, S., Madde, S., Quaglini, S., Pistarini, C., Stefanelli, M.: Evidence-based careflow management systems. Journal of Biomedical Informatics 35, 123–139 (2002)

4. Panzarasa, S., Quaglini, S., Cavallini, A., Micieli, G., Pernice, C., Pessina, M., Stefanelli, M.: Workflow Technology to Enrich a Computerized Clinical Chart with Decision Support Facilities. Proceedings AMIA 2006, pp. 619–623 (2006)
5. Panzarasa, S., Quaglini, S., Micieli, G., Marcheselli, S., Pessina, M., Pernice, C., Cavallini, A., Stefanelli, M.: Improving compliance to guideline through workflow technology: implementation and results in a Stroke Unit. Presentation to MEDINFO 2007 (submitted, 2007)
6. Giardini, G., Bottacchi, E., Corso, G., Carenini, L., Di Giovanni, M., Veronese, M., Cordera, S.: Eleggibilità alla trombolisi in pazienti con ictus ischemico acuto: i dati dello Stroke Registry della Valle d'Aosta. Il Giornale dello Stroke 1, 8–11 (2006)

Part IX

Workflow Systems

Querying Clinical Workflows by Temporal Similarity*

Carlo Combi[1], Matteo Gozzi[1], Jose M. Juarez[2], Roque Marin[2], and Barbara Oliboni[1]

[1] Department of Computer Science – University of Verona – Italy
{combi|gozzi|oliboni}@sci.univr.it
[2] Dept. of Information and Communication Engineering – Universidad de Murcia – Spain
{jmjuarez|roque}@dif.um.es

Abstract. The degree of fulfillment of clinical guidelines is considered a key factor when evaluating the quality of a clinical service. Guidelines can be seen as processes describing the sequence of activities to be done. Consequently, workflow formalisms seem to be a valid approach to model the flow of actions in the guideline and their temporal aspects. The application of a guideline to a specific patient (guideline instance) can be modeled by means of a workflow case. The best (worst) application of a guideline, represented as a reference workflow case, can be used to evaluate the quality of the service, by comparing the optimal case with specific patient instances. On the other hand, the correct application of a guideline to a patient involves the fulfillment of the guideline temporal constraints. Thus, the evaluation of the temporal similarity degree between different workflow cases is a key aspect in evaluating health care quality. In this work, we represent a portion of the stroke guideline using a temporal workflow schema and we propose a method to evaluate the temporal similarity between workflow cases. Our proposal, based on temporal constraint networks, consists of a linear combination of functions to differentiate intra-task and inter-task temporal distances.

1 Introduction

In the past years, clinical guidelines have received an increasing attention in the medical community, but also in the academic context for research issues related to clinical guidelines modeling [9]. Clinical guidelines describe, in natural language, the recommended behaviour of a medical team, the activities to apply to the patient, and their fulfillment with respect to the time and to the the state of patient health, for defining the best way to manage patients. The number of clinical guidelines, covering almost all major branches of medicine, is growing up, together with their updates delivered after regular reviews and new scientific discoveries.

In this work we consider the Italian guideline for stroke prevention and management (SPREAD) [8]. This guideline aims to provide knowledge and recommendations about primary and secondary prevention of stroke in clinical practice.

The diffusion of guidelines in an electronic form is spreading out, and allows the physicians to compare and evaluate clinical guidelines coming from different countries but focusing on the same clinical activities. Despite the diffusion of electronic versions

* This work was partially supported by the Spanish MEC under the FPU national plan (grant ref. AP2003-4476) and the national projects TIC2003-09400-C04 / TIN2006-15460-C04-01.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 469–478, 2007.

of guidelines, their consultation and interpretation may be very difficult. Due to the fact that the textual version can be very long, some information can be distributed along the document, and the sequence of medical activities can be difficult to understand, and may be ambiguous.

For these reasons, issues related to the formal representation of guidelines have been considered by several research teams [7,9]. On one hand, in the clinical context, guidelines describe a sequence of activities to be done. On the other hand, in the business context, a business process can be defined as a description of tasks and consists of subprocesses, decisions and activities. In both cases, a sequence of activities must be done to reach a (given) goal, in the former case to manage in a correct way the patient situation, while in the latter case to satisfy the business needs. This means that guidelines can be seen as processes, and can be managed by means of business modeling tools such as Workflow Management Systems (WfMS) [7].

In general, Workflow Management Systems (WfMSs) allow one to specify, control, and coordinate the flow of work cases (sequences of activities which form a business process). In the clinical workflow context, an important aspect to consider is time: activities described in a guideline must be done satisfying coordination rules expressing constraints with respect to time.

The rules a process has to follow are described in the workflow schema and must be satisfied by the instances (cases) of the process itself. A workflow schema describes the structure of the cases with respect to the coordination of the activities (parallel activities, activity sequences, total/partial fork, and total/partial join). Moreover, the schema may contain qualitative and quantitative temporal constraints.

Workflow cases, instances of the same workflow schema, can be different with respect to the structure, i.e. to the activities composing the cases, and to their (temporal) order and length. When a workflow schema represents a guideline, its cases represent different instances deriving from the application of the guideline to different patients in different situations. The best (worst) application of the guideline can be represented by means of a workflow case and can be used to evaluate the quality of the service comparing the similarity between the optimal case and the other clinical cases. Moreover, a given case, representing something interesting, can be used to retrieve a particular class of cases similar to the given one. Thus, information retrieval can be done evaluating the similarity between workflow cases.

The evaluation of temporal similarity seems to be an important issue in the clinical context, where instances (cases) are slightly different accordingly to the patient situation. In this work, we propose an approach to evaluate temporal similarity between instances (cases) of the workflow schema by the use of temporal constraint networks, and, as a motivating scenario, we represent a portion of the SPREAD guideline by means of a workflow schema.

The structure of the paper is as follows: Section 2 reviews some of the most sound approaches to temporal similarity for clinical scenarios. Section 3 describes a portion of the considered guideline represented by a workflow model and the execution of its tasks. Section 4 describes the main aspects of our approach to evaluate similarity between workflow cases. Finally, this paper offers the conclusions and future work.

2 Related Work

In clinical workflows, and in most models for clinical scenarios, it is fundamental to find the best way for representing time, processing temporal data, and comparing temporal information by similarity techniques. In general, there are two kinds of medical temporal data: time series (biosignals), and temporal sequences (time-stamped clinical data). Time series similarity proposals usually work with raw time series data (e.g. ECG or EEG directly obtained from monitoring) and aim to derive the most representative features from a large amount of data [5]. Some of the most successful strategies are based on the dimensionality reduction (Discrete Fourier Transform, Discrete Wavelet, Time Warping Transformations), in order to obtain a feature vector or model parameters [5].

Temporal sequences are collections of occurrences of different *event* types, as, for example, the set of test results of a patient during a week in the Intensive Care Unit. Occurrences are usually associated to single time points. In [6], the similarity evaluates the relative position of an event occurrence within a window context. That is, event occurrences are similar if they occur in a similar context, and contexts are defined as the set of events happening within a predefined time window.

Furthermore, it is common to find sequences composed by *facts* holding on intervals, such as the description of protocols, treatments, patient symptoms, or parameters abstracted from biosignals (e.g., ST-segment elevation of an ECG). This kind of temporal data is called *interval sequence* and it is of increasing interest in many research fields, as in temporal clinical abstraction [2,10]. Focusing on the few proposals dealing with similarity for interval sequences, in [12], the authors discuss different solutions for defining similarity between two temporal sequences, according to the distance between their composing intervals. In [3], the authors consider the issue of recognizing similar clinical scenarios, composed by both events (point-based) and facts (interval-based); as the scenarios are represented through temporal constraint networks, similarity is led back to the fusion of both networks. Temporal Constraint Networks are a powerful approach for representing and querying temporal information. They are represented as a Constraint Satisfaction Problem (CSP) [11], where variables denote event types and constraints represent the temporal relations amongst them. The interval algebra (IA), introduced by James Allen to represent and manage interval relations [1], is one of the most considered models and has obtained a large number of theoretical results.

3 A Motivating Problem

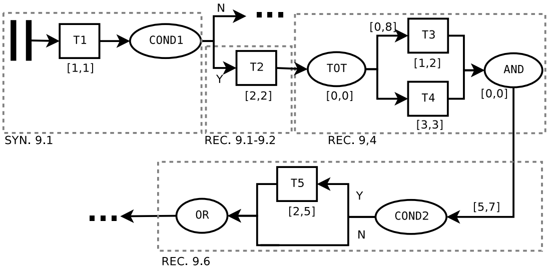
In this paper, we consider the problem of properly managing a patient possibly having a stroke. In particular we focus on the Italian Guideline for Stroke Prevention and Management [8]. We will represent the suitable portion of the considered guideline by using a temporally extended workflow model, which allows one to simply show the required clinical tasks (i.e., activities), the flow of the execution of tasks, and the temporal constraints on them.

Let us consider the following fragments of the guideline:

*"**Synthesis 9.1:** A stroke victim should rapidly be assessed after hospitalization (`T1`), by means of a general examination and [...]*

***Recommendation 9.1 and 9.2:** an early and standardized neurological evaluation (`T2`) is recommended in the setting of a qualitatively adequate management of acute stroke (`Cond1`).*
***Recommendation 9.4:** [...] the following blood exams are recommended: complete blood count including platelets (`T3`), [...], and coagulation tests (`T4`) [...]*
***Recommendation 9.6:** The electrocardiogram (`T5`) is recommended in all suspected stroke victims who are admitted to an Emergency Room (`Cond2`).*"

In Figure 1 we show the considered portion of the guideline by means of a workflow schema, where boxes represent tasks (`T1`, `T2`, ...), ovals represent connectors specifying different possible flows (`Cond1`, `AND`, ...), arrows associate successive tasks/connectors, and a double line is for the start point of the workflow. This representation is enriched by additional temporal information, such as minimum and maximum duration of a task (e.g., `[1,2]` is the allowed interval for duration of task `T3`) or the minimum and maximum delay between two consecutive tasks. According to the specified schema, there are several possible workflow instances (hereinafter *cases*), which correspond to the clinical treatment of different patients. Figure 2 depicts three different cases for the discussed workflow schema: each case is represented as a sequence of intervals labeled by the corresponding task name on the timeline having the start of the case as origin.



Workflow	Clinical Guideline
T1	Hospital Admission
COND1	Suspect Stroke ? (Y/N)
T2	Neurological Assessment
TOT-AND	(do in parallel)
T3	Blood count+platelets
T4	Coagulation tests
COND2	Admitted to E.R.? (Y/N)
T5	Electrocardiogram

Fig. 1. The workflow schema of the considered guideline portion enriched by additional temporal information

In general, cases are only temporally constrained by the specification of the workflow schema. Thus, cases of the same schema could differ with respect to the order and duration of tasks, and with respect to the presence of different tasks due to alternative paths. For instance, the first and the second cases in Figure 2 differ on the order between tasks `T3` and `T4`, while the first (second) case and the third one differ on tasks, due to the alternative paths induced by the connector `Cond1`.

Note that the correct application of clinical guidelines and protocols is considered a quality of service indicator. In order to measure this indicator, one essential factor is the temporal dimension. Thus, according to the given temporal scenario, a huge amount of cases for the same guideline will be stored by a hospital stroke unit. Querying and analyzing this database is, thus, extremely important for several clinical applications: for example, to evaluate the quality of the provided care, we could compare the similarity between the best case and the real clinical cases in the database. Moreover, a

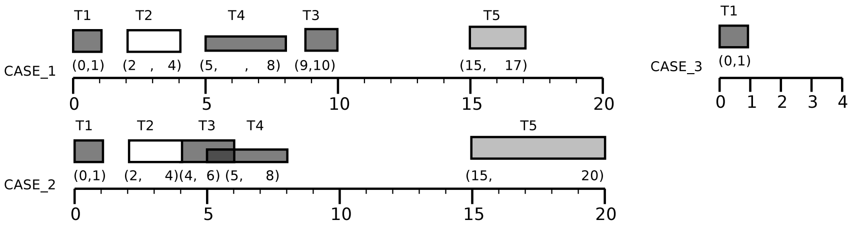


Fig. 2. Three examples of workflow cases

given case, representing something clinically interesting, may be used to retrieve a set of cases similar to the given one. A proper definition of (temporal) similarity for cases needs to be deeply studied.

4 A Similarity Proposal for Clinical Cases

In this section, we propose an approach to evaluate the similarity between two clinical (workflow) cases considering: (i) the comparison between the corresponding performed tasks; (ii) the comparison between the qualitative/quantitative temporal relations between corresponding tasks; and (iii) the presence/absence of some task.

Our proposal compares workflow cases by means of an interval similarity function and is based on the following steps:

1. express both clinical cases through interval constraint networks;
2. evaluate the intra-task distance, i.e. the distance between intervals representing corresponding tasks;
3. evaluate the inter-task distance, i.e. the distance between the relations between corresponding tasks;
4. compute the overall similarity, by considering possible dissimilarities of cases with regard to the occurring tasks.

In the following, we will present and discuss the details of each step.

A workflow case (C) is a set of labeled task intervals. A labeled task interval (t) is a triple $(taskName, t^-, t^+)$: $taskName$ is a task label (e.g. `T1`, `T2`, `T3`), and t^-, t^+ are timestamps describing the beginning and ending time of the task interval. Moreover, C is an ordered set of task intervals:

$$C = \{(taskName_1, t_1^-, t_1^+), (taskName_2, t_2^-, t_2^+), \dots, (taskName_n, t_n^-, t_n^+)\}$$

where $\forall t_i \in C,\ t_i^- \leq t_{i+1}^-, i = 1, \dots, n-1$

Temporal constraint networks are a temporal modeling approach that provides an explicit representation of temporal relations for a given scenario. This is an advantage when two temporal scenarios must be compared. Moreover, these constraints can also contain different aspects of the temporal information (e.g. quantitative/qualitative, crisp/fuzzy) enriching its description and providing a flexible representation for different purposes.

The first step, therefore, of our approach is to obtain for each clinical workflow case a temporal constraint network. The considered network is composed by nodes and edges: nodes represent task intervals, while edges stand for qualitative relations, enriched by some quantitative information, between two task intervals. Nodes are labeled by the corresponding task name and by the related (upper and lower bounds of) interval durations. Multiple instances of a single task (i.e., from a loop in the workflow schema) are labeled by different identifiers (e.g. instances of task T1 would be named T1.1, T1.2, and so on). Edges are labeled by a single Allen's interval relation enriched with some quantitative data. Each task interval has a corresponding node in the network, while edges are introduced only for relations between each task t_i and its successive one t_{i+1}. For example, Figure 3 depicts the three networks corresponding to the cases reported in Figure 2: as for CASE_1, tasks T1 and T2 are represented by labeled nodes and their relation is represented as a directed edge, labeled by b, standing for the relation *before*, and by the interval [1,1], describing the (minimum and maximum) delay between the end of T1 and the beginning of T2[1].

Once the workflow cases are translated into temporal networks, in order to perform a comparison between cases, we need to establish a correspondence between nodes and edges of two different networks. As the task names univocally identify tasks within a case, the correspondence between nodes is built through task names; the correspondence between edges is built up on the correspondence of the connected nodes; when two connected nodes are consecutive in a case and are not consecutive in the other one, we need to derive the missing edge. For example, in Figure 3 derived edges are represented through dashed edges.

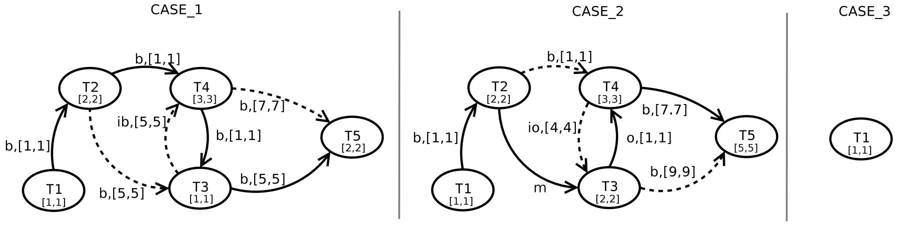


Fig. 3. The temporal networks obtained from the workflow cases of Figure 2

The *intra-task distance* provides a direct method to evaluate the similarity of corresponding tasks, with respect to their durations and to other atemporal features within the workflow process (such as the agent that performed the task...).

[1] Note that the network corresponding to a case has no uncertainty for task durations and delays, and therefore the given ranges are redundant. We maintain this redundancy in the graphical notation to adhere to the usual notation for temporal networks and to adopt the related algorithms.

The intra-task distance (d_{in}) is based on the duration of the task interval ($D_t = (t^+ - t^-)$). Given two corresponding tasks t' and t'', having the same task name, the intra-task distance function is defined as follows:

$$d_{in}(t', t'') = \alpha \frac{|(D_{t'} - D_{t''})|}{|D_{t'}| + |D_{t''}|} + (1 - \alpha) d_{wf}(t', t'')$$

where $\alpha \in [0, 1]$ is the weight of the duration in the function.

The distance function d_{wf} measures other workflow parameters not related to the temporal dimension.

Table 1 shows the intra-task distance values of the problem example between the cases CASE_1 and CASE_2, with $\alpha = 1$. Between the cases CASE_1 and CASE_3 ($\alpha = 1$), $d_{in}(\mathtt{T1}_{\mathtt{CASE\_1}}, \mathtt{T1}_{\mathtt{CASE\_3}}) = 0$.

Table 1. Example of the intra-task distance between CASE_1 and CASE_2

	T1	T2	T3	T4	T5	$\sum d_{in}$
d_{in}	0	0	1/3	0	3/7	0.7619

After considering intra-task distances, we have to take into consideration the inter-task similarity. It deals with similarities for temporal relations between corresponding tasks. To this end, we define an inter-task distance function (d_{IN}), which takes into account both quantitative and qualitative components of the edge labels in the temporal network, by using functions q and Q, respectively.

Given two corresponding relations r', r'' in two different workflow clinical cases, the inter-task distance function is defined as follows:

$$d_{IN}(r', r'') = \begin{cases} Q(v', v'') & \text{if } Q(v', v'') > 0 \\ \beta q(m', m'') & \text{if } Q(v', v'') = 0 \end{cases}$$

where $r' = (v', m')$ ($r'' = (v'', m'')$) represents the relation between the tasks of the first (second) case, through a qualitative value v' (v''), i.e., one of the Allen's relations, and a quantitative value m' (m''), i.e., the distance between the end time of the first task and the beginning of the second one. β is a weight for the quantitative component within the function.

The distance between two Allen's relations is evaluated according to the distance of the considered relations on one of the neighbour graphs proposed by Freksa in [4]: two interval relations between the same intervals are neighbours if it is possible to directly move from one relation to the other one, by continuously deforming the intervals (i.e. shortening, moving...). For example, if we have a *before* relations between two intervals, we can move from *before* to *meets*, by simply moving the first interval to be contiguous to the second one: in this case, the distance between *before* and *meets* is 1. In this work, we adopted, without loss of generality, the A-neighbours graph (a particular neighbour graph obtained by fixing 3 of the 4 bounds of the two intervals), as depicted in Figure 4.

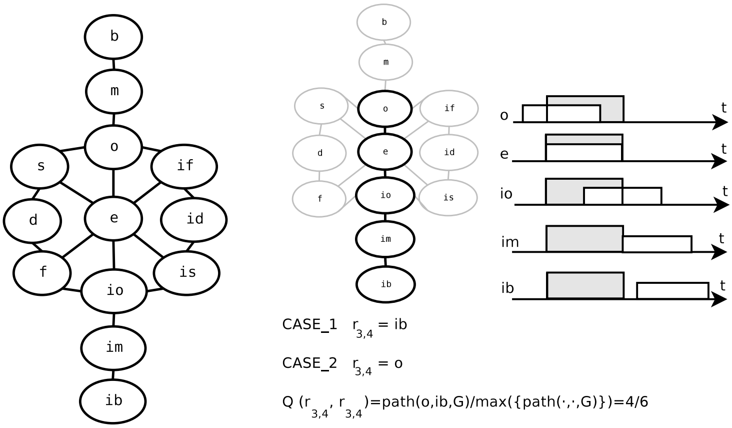


Fig. 4. The A-neighbours graph proposed by Freksa and an example of the Q function

The function Q evaluating the distance between two qualitative temporal relations and the function q, considering the quantitative part, are defined as:

$$Q(v', v'') = \frac{path(v', v'', G)}{max(\{path(\cdot, \cdot, G)\})} \ , \ q(m', m'') = \frac{|m' - m''|}{|m' + m''|}$$

where G is the chosen neighbours graph, $path(v', v'', G)$ stands for the length of the shortest path in G between relations v' and v'', and normalization is performed with respect to the longest path among the shortest ones in G (i.e., the maximum value of the function $path(x, y, G)$ for any couple of nodes (x, y) in G).

Table 2 shows the inter-interval distance values of the problem example between the cases `CASE_1` and `CASE_2` where $\beta = 0.1$.

Table 2. Example of the inter-interval distance. $r_{i,k}$ stands for corresponding relations (for example, $r_{1,2}$ stands for the corresponding relations between tasks `T1` and `T2`).

	$r_{1,2}$	$r_{2,3}$	$r_{2,4}$	$r_{3,4}$	$r_{3,5}$	$r_{4,3}$	$r_{4,5}$
Q	0	1/6	0	2/3	0	2/3	0
q	0	-	0	-	2/7	-	0
d_{IN}	0	1/6	0	**2/3**	2/70	**2/3**	0

In general, the similarity is inversely proportional to the distance between the elements (i.e., tasks and temporal relations), to be compared. Until now, we have defined the concept of distance between corresponding elements. In our case, when defining the overall similarity between two clinical workflow cases, we have to consider also the presence of tasks in one workflow case without corresponding tasks in the other workflow case. Such a presence of non corresponding tasks is suitably represented in the overall similarity function we will define for clinical workflow cases, by the function p.

Given two workflow cases C' and C'', represented by sets T', T'' of tasks, and by sets R', R'' of temporal relations in the corresponding temporal networks, the overall similarity is defined as:

$$similarity(C', C'') = 1 - (\gamma d(C', C'') + (1 - \gamma)p(C', C''))$$

where $\gamma \in [0, 1]$ is the weight for distance similarities and

$$d(C', C'') = \delta \frac{\sum_{t' \in T', t'' \in T''} d_{in}(t', t'')}{|T' \cap T''|} + (1 - \delta) \frac{\sum_{r' \in R', r'' \in R''} d_{IN}(r', r'')}{|R' \cap R''|}$$

while

$$p(C', C'') = \frac{|(T' \cup T'') \backslash (T' \cap T'')|}{|T'| + |T''|}$$

Note that the $\sum d_{in}$ must calculate the inter-interval distance avoiding to measure redundant temporal information. For instance, in Figure 2 nodes i_3 and i_4 (CASE_1 and CASE_2) have both the temporal redundancy of r_{34} and r_{43}, obviously having the same d_{IN} value (see Table 2). In that particular cases, the d function only considers one of the two constraints (ignoring the inverse). In the motivating example, assuming that $\gamma = 0.5$ and $\delta = 0.5$, the similarity between cases CASE_1 and CASE_2 can be calculated given d_{in} and d_{IN} (results calculated previously in this paper) as follows:

$$d(\mathtt{CASE\_1}, \mathtt{CASE\_2}) = 0.5 \frac{0.761904}{5} + 0.5 \frac{0.8619}{5} = 0.16238$$

$$p(\mathtt{CASE\_1}, \mathtt{CASE\_2}) = \frac{0}{5 + 5} = 0$$

$$similarity(\mathtt{CASE\_1}, \mathtt{CASE\_2}) = 1 - (0.5 d(\mathtt{CASE\_1}, \mathtt{CASE\_2})) = 0.9188$$

And the similarity measure between CASE_1 and CASE_3 is:

$$d(\mathtt{CASE\_1}, \mathtt{CASE\_3}) = 0 \ , \ p(\mathtt{CASE\_1}, \mathtt{CASE\_3}) = \frac{\{\mathtt{T1}, \mathtt{T2}, \mathtt{T3}, \mathtt{T4}, \mathtt{T5}\} \backslash \{\mathtt{T1}\}}{5 + 1} = 2/3$$

$$similarity(\mathtt{CASE\_1}, \mathtt{CASE\_3}) = 1 - (0.5p) = 0.66$$

5 Discussion and Conclusions

This work deals with the representation of clinical guidelines by using temporally extended workflow modeling techniques. In particular, we propose an approach to evaluate the temporal similarity between workflow cases representing different applications of the same guideline. The similarity measure proposed in this paper provides a simple but powerful way to compare workflow cases by using temporal constraint networks, providing explicit temporal information about interval distances. We propose a general method that can be also applied for non medical applications; however, its use in clinical domains is essential due to the importance of the temporal dimension in many clinical procedures.

Related proposals in the literature concern about event sequences [6], or interval sequences, like in [12]. Unlike our proposal, these sequence similarity approaches do not consider (or consider partially) the relative order position of intervals within the overall sequence. Another relevant aspect of our approach is the potential capability of managing and inferring temporal knowledge by inferring temporal information from the temporal network. Moreover, the use of temporal constraint networks also provides a flexible representation for evaluating the similarity of uncompleted and imprecise descriptions of a temporal scenario (e.g. when it is not possible to obtain a crisp duration of the tasks in a workflow case). In [3], temporal constraint networks represent clinical scenarios and the consistency of the fusion of networks (incompatible, compatible, or satisfactory) is used as a qualitative similarity evaluation. In this sense, our proposal also covers this aspect but considering also the absence of some tasks in the scenario.

Our future work will focus on the description of specific temporal constraint network models to obtain an efficient similarity function and its evaluation in a concrete medical domain.

References

1. Allen, J.F.: Maintaining knowledge about temporal intervals. Communications of the ACM 26, 832–843 (1983)
2. Chittaro, L., Combi, C.: Visualizing queries on databases of temporal histories: new metaphors and their evaluation. Data Knowl. Eng. 44(2), 239–264 (2003)
3. Dojat, M., Ramaux, N., Fontaine, D.: Scenario recognition for temporal reasoning in medical domains. Artificial Intelligence in Medicine 14(1-2), 139–155 (1998)
4. Freksa, C.: Temporal reasoning based on semi-intervals. Artificial Intelligence 54(1), 199–227 (1992)
5. 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 (2001)
6. Mannila, H., Moen, P.: Similarity between event types in sequences. In: DaWaK '99: Proceedings of the First International Conference on Data Warehousing and Knowledge Discovery, London, UK, pp. 271–280. Springer, Heidelberg (1999)
7. Panzarasa, S., Stefanelli, M.: Workflow management systems for guideline implementation. Neurological Sciences 27 (2006)
8. The Stroke Prevention and Educational Awareness Diffusion (SPREAD) Collaboration. The italian guidelines for stroke prevention. Neurological Sciences, 21 (2000)
9. Quaglini, S., Ciccarese, P.: Models for guideline representation. Neurological Sciences 27 (2006)
10. Shahar, Y., Musen, M.A.: Knowledge-based temporal abstraction in clinical domains. Artificial Intelligence in Medicine 8(3), 267–298 (1996)
11. Vilain, M., Kautz, H.: Constraint propagation algorithms for temporal reasoning. In: Proceedings of the National Conference on Artificial Intelligence (AAAI-86), USA, vol. 6, pp. 132–144 (1986)
12. Yi, B.-K., Roh, J.-W.: Similarity search for interval time sequences. In: DASFAA, vol. 2973, pp. 232–243 (2004)

Testing Careflow Process Execution Conformance by Translating a Graphical Language to Computational Logic

Federico Chesani[1], Paola Mello[1], Marco Montali[1], and Sergio Storari[2]

[1] DEIS – Università di Bologna
viale Risorgimento, 2 – 40136 – Bologna, Italy
{fchesani|pmello|mmontali}@deis.unibo.it
[2] ENDIF – Università di Ferrara
Via Saragat, 1 – 44100 – Ferrara, Italy
strsrg@unife.it

Abstract. Careflow systems implement workflow concepts in the clinical domain in order to administer, support and monitor the execution of health care services performed by different health care professionals and structures. In this work we focus on the monitoring aspects and propose a solution for the conformance verification of careflow process executions.

Given a careflow model, we have defined an algorithm capable of translating it to a formal language based on computational logic and abductive logic programming in particular. The main advantage of this formalism lies in its operational proof-theoretic counterpart, which is able to verify the conformance of a given careflow process execution (in the form of an event log) w.r.t. the model.

The feasibility of the approach has been tested on a case study related to the careflow process described in the cervical cancer screening protocol.

Keywords: Careflow management, Clinical practice guidelines, Conformance verification, Computational logic.

1 Introduction

In modern health care organizations, clinical decisions are progressively based on evidence-based care [1]. In order to achieve the goals of this approach, the adoption of clinical practice guidelines and computer-based guideline management systems is considered very important. These guidelines are effectively used in practice if they are managed by systems that are deeply integrated with the health care information systems and take care of the aspects related to the clinician's workflow (namely *careflow*).

As described in [2], careflows focus on the behavioural aspects of medical work described in clinical practice guidelines. Careflow systems implement workflow concepts in the clinical domain, coordinating the execution of health care services performed by different health care professionals and structures. The literature proposes several formalisms to represent workflows/careflows. Usually, they are graphical flow-charts that clearly express the sequence of activities to be performed. This formalization is used by the workflow management systems to administer, support and monitor the process execution.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 479–488, 2007.

We focus on the monitoring aspects and present a preliminary result of our research activity aimed to perform the conformance verification of careflow process executions. We have developed a tool for describing careflow models in a graphical language (named GOSpeL [3]). In this work we present an algorithm for translating a GOSpeL model into a set of rules written in a declarative language based on computationl logic, named $\mathcal{S}$CIFF [4]; careflow participants are mapped to agents, activities to events and medical knowledge to a knowledge base. Such a formalization is used by the operational proof-theoretic counterpart of the $\mathcal{S}$CIFF language, which is able to verify the conformance of a given careflow process execution (in the form of a log of the events) w.r.t. the model. We discuss how the result of the verification step could be used for better understanding the careflow model as well as for pointing out possibly non-forecasted behaviours.

2 An Overview of GOSpeL

GOSpeL [3] is a software tool for specifying careflows by means of a graphical language. In GOSpeL, a careflow is represented in terms of a (1) *flow chart*, which models the process evolution; and (2) an *ontology*, which describes at a fixed level of abstraction the application domain and provides a semantics to the flow-chart.

Careflow process evolutions are described by means of blocks (shown in Table 1) and relations between blocks. GOSpeL blocks accounts for *activities* (work units at the desired abstraction level), *gateways* (to manage the convergence and the divergence of control flows), and *start/completion point* of complex activities. A complex activity models a composite unit of work, defined at a lower abstraction level by a new (sub)process, thus allowing a top-down approach for the careflow definition. GOSpeL blocks can be connected by using *order relations* (that represent the flow path), or by using *temporal relations* (to model temporal constraints among activities like e.g. deadlines and delays). GOSpeL ontologies follow a simple template, supporting mainly two taxonomies: *Activities* to model the atomic ontological activities of the target domain; *Entities* to model actors and objects involved in an activity execution.

3 A Brief Description of the $\mathcal{S}$CIFF Framework

The $\mathcal{S}$CIFF framework was originally developed in the context of the SOCS European project [5] for the specification and verification of agents interaction protocols within open and heterogeneous societies. The framework is based on abduction, a reasoning paradigm (well suitable in the medical field) which allows to formulate hypotheses (called *abducibles*) accounting for observations. In most abductive frameworks, *integrity constraints* are imposed over possible hypotheses in order to prevent inconsistent explanations. The idea behind the SOCS framework is to formalize domain knowledge in terms of abductive logic programming and expectations about participants behaviour as abducibles, and to use Social Integrity Constraints (IC_S) to detect such behaviour that is not compliant with interaction protocols. It is assumed that the ongoing social participants behaviour is accessible and represented by a set of (ground) facts called *events*. Happened events are denotated by the functor $\mathbf{H}$: e.g.,

Table 1. GOSpeL elements

family	**type**	**notation**	**description**
Activities	atomic activity	🟦	single atomic unit of work within the guideline
	complex activity	⤢	Non-atomic unit of work. It encapsulates a new (sub)process definition.
	iteration	◎	For-like cyclical complex activity
	while	◈	While-like cyclical complex activity
Gateways	exclusive choice	❖	Data-based choice; each outgoing relation is associated to a logical guard, and at evaluation-time, one of the path which has a a true condition is chosen.
	deferred choice	◆	Non-deterministic choice, without explict logical conditions; the choice is delayed until one of the possible paths is actually performed by participants.
	parallel fork	🟩	Point at which multiple threads of execution are spanned
	parallel join	🟩	Synchronization of multiple threads of control
Start/ Completion Blocks	start	▶	Start point of a complex activity
	cyclic start	▶	Start point of a cyclical complex activity
	completion	◉	Completion point of a complex activity
	abort	●	Abort the entire guideline

$\mathbf{H}(enter(p, emergency\_ward), 7)$ represents the fact that p has entered into the hospital's $emergency\_ward$ at time 7.

Generally speaking, the participants behaviour is unpredictable. However, interaction protocols provides hints about which are the possible expectations about future events. This represents in some sense the "ideal" behaviour. Events expected to happen are indicated by the functor $\mathbf{E}$ and have the same format as happened events but they will typically contain variables, to indicate that expected events are not completely specified. These variables may be constrained by using CLP constraints [6] or bound by evaluating predicates defined in the $\mathcal{S}$CIFF abductive knowledge base (named SOKB).

Given the happened events, IC_s specifies how to generate expectations. An IC is a rule of the form $body \rightarrow head$, expressing that when $body$ becomes true then $head$ is expected. Protocols are defined as sets of rules, relating happened events to expectations about future events. E.g.:

$$\begin{aligned} &\mathbf{H}(enter(Pat, emergency\_ward), T_1) \wedge high\_priority(Pat) \\ \rightarrow &\mathbf{E}(examine(Phy, Pat), T_2) \wedge T_2 > T_1 \wedge T_2 < T_1 + 15 \end{aligned} \tag{1}$$

states that, if a patient Pat enters into the emergency ward and it is evaluated as "high priority", then a physician Phy is expected to examine him/her within a quarter of an hour (supposing that times are expressed in minutes). The $high\_priority/1$ predicate is defined in the SOKB. Notice that temporal constraints can be imposed (in particular deadlines).

Given a set of happened events (i.e. an event log or history), expectations are generated by he operational counterpart of the $\mathcal{S}$CIFF language, namely the $\mathcal{S}$CIFF Proof Procedure [7]. The most distinctive feature of this proof procedure is the ability to check that the generated expectations, considered as a particular class of abducibles, are *fulfilled* by the actual participants behaviour. If a participant does not behave as expected w.r.t. the model, the proof procedure detects and raises as soon as possible a violation.

Our approach is suitable for modeling the careflow aspects of a clinical practice guideline, especially when the execution order and the appropriateness of the health services should be strongly enforced (like for example screening protocols). In this context, we are interested in detecting two different types of violation. The first one is raised when a participant does not act as expected by the careflow model (i.e., an expectation is not fulfilled by a corresponding happened event); the second one is raised when a participant performs activities not expected by the model (i.e., a happened event is not explicitly expected). When a violation is detected, two possible hypothesis could be given in order to explain such a violation: either the participants exhibited a wrong behavior (w.r.t. the careflow model), or the model itself has not been properly defined (hence it does not fit well with the real guideline's execution). Assuming the latter hypothesis, violations are a useful to understand *how* and *where* the careflow model specification lacks.

4 Translation Algorithm

Intuitively, a careflow model specifies that, when an activity block is performed, other activities should be performed in the right order and with the right attributes. From the $\mathcal{S}$CIFF viewpoint, this is equivalent to specify an IC that relates the happened event with the future ones. Given an activity block A, the part of the diagram next to A is considered as a description of the possible behaviours which the participants has to exhibit. Therefore, the algorithm generates an IC whose body contains the happening of e_A and whose head is determined by the consequent diagram part. The notation e_A represents the event to which a generic block A has been ontologically mapped[1]; the properties of this event, namely its name and its attributes, are respectively determined by the name of the ontological activity and the set of formal participants associated to A. Leaving A and going forward in the graphical model, for each branch a new activity block will be detected (sooner or later), and will be mapped to an expectation about the future participants behaviour. Afterward, these blocks will be considered (recursively) as new starting points by the algorithm.

Note that start, return and end blocks are mapped to events too: even if they do not really represent a concrete working step during the process application, they are used as terminal points that identify the start and the conclusion of a (sub)process. In the following, we will refer to the blocks which are mapped to events as *event-blocks*.

The algorithm visits a GOSpeL diagram partitioning it into special sub-sets (called *minimal windows*) and translating each sub-set to a Social Integrity Constraint. In order to define a minimal window, we introduce some other concepts: *precursors* and *successors* sets, *path*, *window* and window's *source* and *fringe*.

Definition 1 (Precursors and Successors Sets). *Given a block b:*

- *Suc_b is the set of blocks to which b is directly connected through its outgoing relations (successors set);*

[1] We adopt an atomic model for simple activities.

- Pre_b *is the set of blocks to which* b *is directly connected through its incoming relations (precursors set).*

Definition 2 (Path). *A path* $P(s,d)$ *is a sequence of blocks through which block* s *and block* d *are connected, following the order relations. Defining the sequence as* $b_0 = s, b_1, \ldots, b_{n-1}, b_n = d$*, we have:*

$$b_j \in Suc_{b_{j-1}} \wedge b_j \in Pre_{b_{j+1}} \forall j = 1, \ldots, n-1$$

Definition 3 (Window). *A subset* $\mathcal{W}$ *of GOSpeL blocks is a* window *if it is connected, i.e.*

$$\forall b_1 \in \mathcal{W} \exists b_2 \in \mathcal{W} s.t. \exists P(b_1, b_2) \in \mathcal{W} \vee \exists P(b_2, b_1) \in \mathcal{W}$$

Note that $P(s,d) \in \mathcal{W}$ iff all the blocks of the sequence belong to $\mathcal{W}$.

Definition 4 (Window Source and Fringe). *The* source *and the* fringe *of a window* $\mathcal{W}$ *are respectively:*

$$S_{\mathcal{W}} = \{b \in \mathcal{W} \mid \nexists b' \in \mathcal{W} \; s.t. \; \exists P(b', b)\}$$

$$F_{\mathcal{W}} = \{b \in \mathcal{W} \mid \nexists b' \in \mathcal{W} \; s.t. \; \exists P(b, b')\}$$

Definition 5 (Minimal Window). *A window* $\mathcal{W}$ *is* minimal *iff* $\forall b \in \mathcal{W}$ *the following properties hold:*

1. *if* $b \in S_{\mathcal{W}}$ *then* b *is an event-block;*
2. *else if* $b \in F_{\mathcal{W}}$ *then* b *is an event-block;*
3. *else* b *is not an event-block (i.e. it is a split or merge);*
4. *if* b *is a split-block then* $Suc_b \in \mathcal{W}$*;*
5. *if* b *is a merge-block then* $Pre_b \in \mathcal{W}$*.*

Properties 4 and 5 of Def. 5 ensure that when a split-block (a merge-block respectively) belongs to the minimal window, all the branches which diverge from (converge to, resp.) it are included. Note also that, for a well-formed flow-chart, it is impossible to have a window that contains a split-block followed by a merge one (each path that connects two blocks of this type must include at least an activity-block between them).

4.1 Mapping of a Minimal Window to an IC

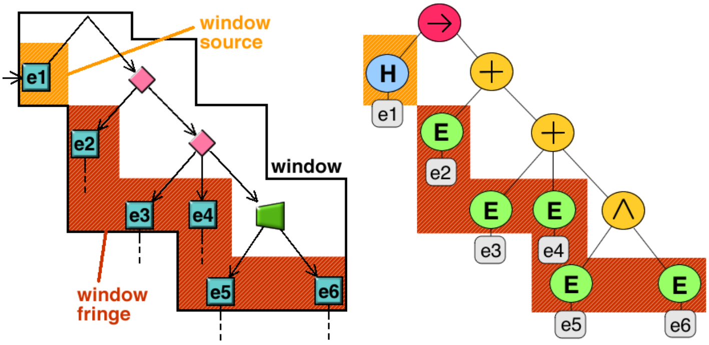
Figure 1(a) shows a minimal window and Figure 1(b) its translation. It's easy to see a tight similarity between the minimal window and the abstract parse tree of the corresponding IC.

The translation procedure of a minimal window $\mathcal{W}$, named in the following $GENERATE\_IC$, operates as follows [2,3]:

1. $\forall b \in S_{\mathcal{W}}$ generates $\mathbf{H}(e_b, T_b)$ (if b is a macroblock, its completion point is chosen);
2. creates a body of a rule by composing the happened events in a way that depends on the merge-blocks in $\mathcal{W}$;

[2] Remember that $S_{\mathcal{W}}$ and $F_{\mathcal{W}}$ contain only event-blocks (Property 1 and 2 of Definition 5).

[3] For the sake of clarity, we make the assumption that each ontological activity is associated at most to one activity block. The general algorithm does not require to state this assumption.



(a) (b)

Fig. 1. Minimal window and abstract parse tree of its translation

3. $\forall b \in F_{\mathcal{W}}$ generates $\mathbf{E}(e_b, T_b)$ (if b is a macroblock, its start point is chosen);
4. creates a head composing the expectations in a way that depends on the split-blocks in $\mathcal{W}$.

Let us consider for example a parallel join, the only merge-block defined in GOSpeL: due to its synchronization semantics the generated IC must trigger (i.e., must have a body that becomes true) only when all the previous events happen. Therefore, in presence of a parallel join the body will contain a conjunction of the happened events generated during the first step; Property 5 of Definition 5 ensures that all the previous events are considered for the synchronization.

Similarly, split blocks determine how the expectations generated in the third step are composed; as Figure 1 suggests, the parallel fork is mapped to a logical conjunction among the expectations found on each outgoing branch, whereas the semantics of deferred choice imposes mutual exclusion between branches, generating a rule that waits for one among several events. Furthermore, the behaviour of an ex-or blocks is the same as the deferred choice one, despite the fact that each alternative is associated to its logical condition (i.e., it can be chosen iff the associated guard is evaluated to true).

4.2 General Algorithm

Giving the start block $Start$ of a GOSpeL model, the translation algorithm operates splitting the whole diagram into a set of minimal windows and mapping each window to a an IC:

```
$ics = \emptyset, visited = \emptyset, fringe \leftarrow Start$
while $fringe \neq \emptyset$ do
    $cur \leftarrow REMOVE\_ONE(fringe)$
    $\mathcal{W} \leftarrow CONSTRUCT\_MINIMAL\_WINDOW(cur)$
    $ics \leftarrow ics \cup GENERATE\_IC(\mathcal{W})$
    $visited \leftarrow visited \cup S_{\mathcal{W}}$
    $fringe \leftarrow [fringe \cup F_{\mathcal{W}}] - visited$
end while
```

The $fringe$ set, which initially contains only $Start$, represents dynamically the frontier of the already covered part. At each iteration step, one element is extracted from $fringe$, say, cur. At line 4, the minimal window $\mathcal{W}$ s.t. $cur \in S_{\mathcal{W}}$ is founded. Operationally, $\mathcal{W}$ is constructed starting from cur and visiting the diagram partially forward and partially backward (when a merge block is encountered, Property 5 of Definition 5 says that all its previous branches should be included). The mapping of cur is then handled by the $GENERATE\_IC$ procedure, which has been described in the previous paragraph. Finally, the $visited$ and $fringe$ sets are updated to avoid repetition: remember indeed that in GOSpeL different alternatives may converge to a single path. Figure 2 shows how a fragment of a simple diagram is partitioned into minimal windows.

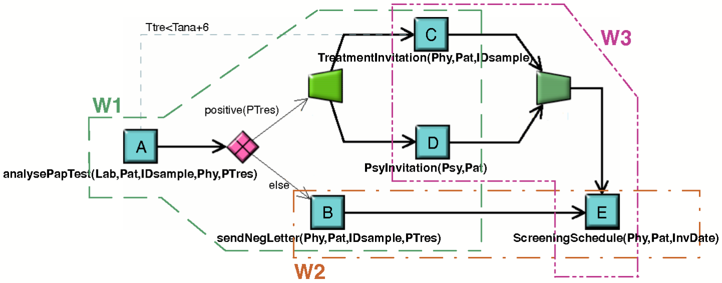

Ttre<Tana+6
W1
W3
W2
A
B
C
D
E
positive(PTres)
else
analysePapTest(Lab,Pat,IDsample,Phy,PTres)
TreatmentInvitation(Phy,Pat,IDsample)
PsyInvitation(Psy,Pat)
sendNegLetter(Phy,Pat,IDsample,PTres)
ScreeningSchedule(Phy,Pat,InvDate)

Fig. 2. Example of a GOSpeL diagram fragment

5 A Case Study

As a case study for exploiting the potentialities of our approach we choose the cervical cancer screening guideline proposed by the sanitary organization of the Emilia Romagna region of Italy [8]. Cervical cancer is a disease where malignant (cancer) cells grow in the tissues of the cervix. The screening program proposes several tests in order to early detect and treat cervical cancer.

For the sake of space, we describe in this section its application to the careflow model of a simplified cervical cancer screening protocol. In this careflow, a lab Lab analyses a pap-test $IDsample$ of patient Pat and sends a report $PTres$, containing a set of signs on the sample, to the screening physician Phy. Phy evaluates $PTres$ and classifies $IDsample$ as positive (cancer evidence found) or negative (normal). If positive, the protocol prescribes that Pat should be invited, in parallel, for the cancer treatment and for a psychological consult. Note that the treatment invitation should be sent to the patient within a deadline of six days. In case of a negative evaluation a letter should be sent to Pat reporting that the pap-test is normal.

The positive and negative flows converge in a single one which proposes as activity the scheduling of the next pap-test.

The GOSpeL model of this careflow is composed by an extension of the base ontology, which contains entities and activities specific of the screening domain, and by the graphical model shown in Figure 2. This model is then translated, by the algorithm described in Section 4, in a set of ICs starting from block A with $fringe$ set to $\{A\}$. At the first iteration step the algorithm extracts A and, launching a visit from it, individuates $\mathcal{W}_1$, which has $S_{\mathcal{W}_1} = \{A\}$ and $F_{\mathcal{W}_1} = \{B, C, D\}$. The following IC is produced:

$$\begin{aligned}
&\mathbf{H}(analysePapTest(Lab, Pat, IDsample, Phy, PTres), T_{ana})\\
\rightarrow &positive(PTres)\\
&\wedge \mathbf{E}(treatmentInvitation(Phy, Pat, IDsample), T_{tre})\\
&\wedge T_{tre} > T_{ana} \wedge T_{tre} < T_{ana} + 6\\
&\wedge \mathbf{E}(psyInvitation(Psy, Pat), T_{psy}) \wedge T_{psy} > T_{ana}\\
\vee &not(positive(PTres))\\
&\wedge \mathbf{E}(sendNegLetter(Phy, Pat, IDsample, PTres), T_{sen}) \wedge T_{sen} > T_{ana}
\end{aligned} \tag{2}$$

Note that the temporal constraint between A and C is inserted as a CLP constraint over T_{tre} and T_{ana}. Other time constraints are automatically generated due to the partial order imposed by order relations. The exclusive choice condition is mapped to the evaluation of the predicate *positive/1*, contained in the SOKB. A pap-test is positive if almost one cervical cancer type can be detected. Since each cancer type is characterized by a specific set of laboratory results, the predicate *positive/1* verifies if almost one of three possible cancer types has more than an half of its supporting signs in $PTres$. This is a trivial description used only in order to exploit the reasoning capabilities of the SOKB. Now the algorithm proceeds updating the $fringe$ set, which becomes $fringe = \{B, C, D\}$. Supposing B is extracted, the algorithm finds window $\mathcal{W}_2$, whose translation is straightforward. After having translated $\mathcal{W}_2$ the fringe contains C, D and E. If either block C or D are extracted, due to the presence of a parallel join the algorithm finds a window which has $S_{\mathcal{W}_2} = \{C, D\}$ and $F_{\mathcal{W}_2} = \{E\}$, and generates the IC (3). The final set of ICs for the cervical cancer screening example is composed by three rules.

$$\begin{aligned}
&\mathbf{H}(treatmentInvitation(Phy, Pat, IDsample), T_{tre})\\
&\wedge \mathbf{H}(psyInvitation(Psy, Pat), T_{psy})\\
\rightarrow &\mathbf{E}(screeningSchedule(Phy, Pat, InvDate), T_{scr}) \wedge T_{scr} > T_{tre} \wedge T_{scr} > T_{psy}
\end{aligned} \tag{3}$$

Given this set of ICs, the $\mathcal{S}$CIFF proof procedure is used by the SOCS-SI [9] tool for verifying the conformance. Let us consider for example a simple execution of the above careflow process represented by a set of happened events:

1. $\mathbf{H}(analysePapTest(lab, pat, 123, phy, [res_1, \ldots, res_n]), 5)$
2. $\mathbf{H}(psyInvitation(psy, pat), 7)$
3. $\mathbf{H}(treatmentInvitation(phy, pat, 123), 20)$
4. $\mathbf{H}(screeningSchedule(phy, pat, 15apr2007), 30)$

When the pap-test analysis is passed to the proof procedure, the first IC triggers and, supposing that the predicate $positive([res_1, \ldots, res_n])$ succeeds, we have two pending expectations: $\mathbf{E}(treatmentInvitation(phy, pat, 123), T_{tre}) \wedge T_{tre} \in [6, 11]$ and $\mathbf{E}(psyInvitation(Psy, pat), T_{psy}) \wedge T_{psy} > 5$. Now we have that the second happened event fulfills the second expectation, grounding Psy to psy and T_{psy} to 20, whereas the treatment invitation event matches with the first one. Unfortunately, the match implies that T_{tre} unifies with 30, which does not satisfy the deadline and causes therefore a violation to be raised. The execution is then classified as non conformant.

This conformance verification approach has been evaluated by using the careflow model of the cervical cancer screening process [8] and a database containing 1950 careflow executions. Some of them, representing incorrect behaviours, were introduced in this database, in order to deeply test our approach and our tools. Each execution contains several events: from the minimum of one (the screening invitation followed by no response) to the maximum of 18 (the whole careflow). The total time occurred to verify the conformance of the 1950 executions w.r.t. the careflow model was 12 minutes (average time of 369msec for each execution). 1091 executions resulted to be not conformant w.r.t. the formalization we have initially proposed. These results were analyzed by a screening expert which confirmed all the conformant classifications and proposed some changes to the careflow model in order to consider as conformant some particular cases, not allowed by the initial model. Using this revised model, we avoided false non conformant classifications, reducing the number of executions classified as non conformant to 44: this result agrees indeed with the "wrong behaviour" executions we introduced in the database. The conformance results were considered useful by the screening expert for the quality evaluation of the careflow process and its revision.

6 Related Works

Several medical guidelines support systems have been proposed to represent and manage clinical guidelines but, for the sake of space, we limit ourselves to only three: GLARE [10], PROforma [11] and NewGuide [12]. GLARE [10] is a system for acquiring, representing and executing clinical guidelines. It provides consistency checks, advanced temporal reasoning techniques, what-if functionalities and guideline properties evaluation. PROforma [11] is a formal language capable to represent a clinical guideline in terms of a network of tasks and data items. NewGuide [12] puts together medical knowledge formalization techniques and workflow management systems (named Careflow Management Systems CfMS). The system supports the definition (in a language similar to Petri Nets), execution and monitoring of guidelines and careflows.

Comparing GLARE, PROforma and NewGuide with our approach we notice that our approach can be considered complementary w.r.t the ones proposed by GLARE and PROforma, since they do not tackle conformance verification issues on careflow execution traces. With respect to NewGuide, we think that our approach may be useful to add verification functionalities to the CfMS administration and monitoring tools [12].

7 Conclusions

In this work we have described a solution for the conformance verification of careflow process executions. We have shown how a careflow model, defined through the GOSpeL graphical language, could be automatically translated to the $\mathcal{S}$CIFF language[4], based on computational logic and abductive logic programming, and how this formalization is then used by the proof-theoretic counterpart of the $\mathcal{S}$CIFF language to verify the conformance of a given careflow process execution w.r.t. the model. The feasibility of this approach has been tested on a cervical cancer screening protocol.

We plan to investigate in future work whether our approach can be extended to the workflow patterns discussed in other guideline support systems, like in [13]. Another ongoing work is about the proof of "high level" properties on the formalized guideline specification by using an extension of the $\mathcal{S}$CIFF proof procedure (named g-$\mathcal{S}$CIFF). For instance, given the IC_S representation of the above guideline fragment, we can ask to g-$\mathcal{S}$CIFF if a history exists s.t. a treatment invitation is sent to the patient. If this is the case, g-$\mathcal{S}$CIFF will produce a successful proof, generating the corresponding history.

Acknowledgments. This work has been partially supported by NOEMALIFE under the "SPRING" regional PRRITT project, by the PRIN 2005 project "Specification and verification of agent interaction protocols" and by the FIRB project "TOCAI.IT". We would like to thank reviewers and participants of the ECAI 2006 workshop on AI techniques in healthcare - evidence-based guidelines and protocols - for the fruitful discussions on a preliminary version of this paper.

References

1. Muir, G.: Evidence-based Healthcare. Churchill Livingston, London (1997)
2. Careflow management systems
   http://www.openclinical.org/briefingpaperStefanelli.html
3. Chesani, F., Matteis, P.D., Mello, P., Montali, M., Storari, S.: A framework for defining and verifying clinical guidelines: A case study on cancer screening. In: Esposito, F., Raś, Z.W., Malerba, D., Semeraro, G. (eds.) ISMIS 2006. LNCS (LNAI), vol. 4203, pp. 338–343. Springer, Heidelberg (2006)
4. Alberti, M., Gavanelli, M., Lamma, E., Mello, P., Torroni, P.: Specification and verification of agent interactions using social integrity constraints. ENTCS 85(2) (2003)
5. Societies of computees (SOCS) Available at:
   http://lia.deis.unibo.it/Research/SOCS/
6. Jaffar, J., Maher, M.: Constraint logic programming: a survey. Journal of Logic Programming 19-20, 503–582 (1994)
7. The SCIFF abductive proof procedure, Available at
   http://lia.deis.unibo.it/Research/sciff/
8. Cervical cancer screening in emilia romagna (italy), Available at:
   http://www.regione.emilia-romagna.it/screening/
9. SOCS-SI web site. Available at:
   http://www.lia.deis.unibo.it/research/socs_si/socs_si.shtml
10. Terenziani, P., Montani, S., Bottrighi, A., Torchio, M., Molino, G., Correndo, G.: Applying artificial intelligence to clinical guidelines: The GLARE approach. In: AI*IA. vol. 101, pp. 536–547 (2003)
11. Fox, J., Johns, N., Rahmanzadeh, A.: Disseminating medical knowledge-the proforma approach. Artificial Intelligence in Medicine 14, 157–181 (1998)
12. Ciccarese, P., Caffi, E., Boiocchi, L., Quaglini, S., Stefanelli, M.: A guideline management system. In: MEDINFO 2004, pp. 28–32. IOS Press, Amsterdam (2004)
13. Mulyar, N.A., v.d. Aalst, W.M.P., Peleg, M.: A pattern-based analysis of clinical computer-interpretable guideline modelling languages. BPMcenter.org Technical Note (2006)

Induction of Partial Orders to Predict Patient Evolutions in Medicine

John A. Bohada, David Riaño, and Francis Real

Research Group on Artificial Intelligence
Departament of Computer Sciences and Mathematics, Rovira i Virgili University
Av. Països Catalans 26, 43007 Tarragona, Spain
{john.bohada,david.riano,francis.real}@urv.net

Abstract. In medicine, prognosis is the task of predicting the probable course and outcome of a disease. Questions like, is a patient going to improve?, what is his/her chance of recovery?, and how likely a relapse is? are common and they rely on the concept of state. The feasible states of a disease define a partial order structure with extreme states those of 'cure' and 'death'; improving, recovering, and survival meaning particular transitions between states of the partial order. In spite of this, it is not usual in medicine to find an explicit representation either of the states or of the states partial order for many diseases. On the contrary, the variables (e.g. signs and symptoms) related to a disease and their normality and abnormality values are broadly agreed. Here, an inductive algorithm is introduced that generates partial orders from a data matrix containing information about the patient-professional encounters, and the normality functions of each one of these disease variables.

1 Introduction

In medicine, prognosis is the process by which the probable course and outcome of a disease is predicted. Statistics and Artificial Intelligence have traditionally faced this process with several *methodologies* as survival analysis, logistic regression, Bayesian Networks, Artificial Neural Networks, Genetic Algorithms, and Decision Trees as [4] and [3] report. All these methodologies have been applied to predict *medical facts* as survival, relapse, improvement, worsening, or death. These predictions depend on whether there is a temporal *restriction* related to the prediction or not. Temporal restrictions may be represented as a single point (e.g. probability of suffering a relapse “after one year”) or as multiple independent points in time [4] (e.g. probability of getting an improvement “within the next three months”). In [3], *prognostic models* are classified into those that predict on populations (e.g. patients that are in a similar condition) and those others that predict on individuals. An additional feature of the above methodologies is whether they are able to predict only one fact (e.g. survival) or whether they are able to predict several facts simultaneously.

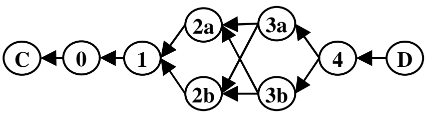
A feasible approach to obtain predictions on several facts simultaneously is based on the concept of *patient condition*, which represents the state of the patient concerning a disease. Thus, finding out the probability of a patient to cure, to

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 489–499, 2007.

improve, to worsen, or to die is equivalent to calculate how likely it is that this patient evolves from his current condition to a condition representing cure, a better than the current condition, an equivalent to the current condition, a worse condition, or the death condition, respectively.

All the possible patient conditions (i.e. states) of a disease define an order relation that represents the pair-wise comparison of the *severity* of the possible conditions in the disease. So, for instance in breast cancer, stage 4 (patients with metastasis) represents a patient condition that is worse than stage 1 (where the tumour is less than 2 cm across and it is not spread). Unfortunately, the severities of two patient conditions are not always comparable or, if they are comparable, it is not always possible to establish one as clearly better than the other one. Therefore, the relationships among the patient conditions of a disease in health-care are frequently represented with *partial orders* which for complex diseases as cancer they are created after an agreement between experts. However, the so created partial orders are not necessarily designed to represent conditions and relationships from a point of view of the severity of the disease but, for instance, to represent the relationships among these conditions from a practical point of view like the sort of recommended treatment is. This can foster differences between what the theoretical model represents (i.e. the expert-based partial order or *standard partial order*) and what is really observed at the health-care centres (i.e. the experience-based partial order). For example, for the data of the SEER repository [7] describing real breast cancer cases, it is observed that 15% of these cases are in a condition whose severity does not correspond to the severity of the stage indicated by the TNM Staging System [8] in Fig. 1.

The reason for that is that the degree of severity of a particular patient condition is not necessarily based on whether this patient fulfils a set of facts or not, but on the combination of the degrees of severity of each one of the variables that define the state of a patient in a particular disease. For instance, it does not seem very wise to admit patients with breast cancers of 2.0 cm in stage 2 (i.e. severity 2), and at the same time do not consider the possibility of a patient with a 2.1 cm tumour to be in stages with severities below or equal to 2 just because the definition of stage 2 in breast cancer sets the size upper limit in 2 cm. Following with the example, it could be the case that the first patient with a 2 cm tumour has other complications affecting the seriousness of his disease, making his condition more severe than the one of the second patient, and causing the prognostic of the first patient not to be very accurate.



TNM BREAST CANCER CONDITIONS

Stage C: no breast cancer observed (Cure). **Stage 1:** The tumour size <2 cm; armpit lymph nodes not affected; cancer not spread. **Stage 2a:** no cells in lymph nodes; cancer in outer covering of the bowel. **Stage 2b:** cancer in outer covering of bowel wall & in nest tissues/organs, lymph nodes not affected; cancer not spread. **Stage 3a:** cancer in inner layer of bowel wall or in the muscle layer; 1-to-3 nearby lymph nodes contain cancer cells. **Stage 3b:** cancer through the bowel wall or in surrounding body tissues/organs; 1-to-3 nearby lymph nodes with cancer cells. **Stage 4:** any size; armpit lymph nodes can be affected; metastasis to other parts of the body. **Stage D:** The patient died (Death).

Fig. 1. TNM Staging System for Breast Cancer

In order to support the correct joint analysis of the condition of a patient with respect to both the standard partial order and the experience-based partial order, it is required to develop algorithms to derive partial orders from the patient records stored in hospital databases. The purpose of this is twofold: on the one hand, these algorithms can be used to generate new health-care knowledge on the feasible stages of a particular disease, and on the other hand, they can be combined with probability theory to increase the accuracy of prognosis on the evolution of a patient.

This paper describes an algorithm to induce partial orders on the patient conditions of a disease. The induction process takes the data of the patients that are registered in the hospital databases and that are described in terms of the variables that condition the health state of the patient in the target disease, and produces a partial order that, together with a state-transition diagram that represent the changes of condition of the patients in the healthcare centre, is able to predict the evolution of new patients.

The rest of the paper has four sections. Section 2 formalises the problem and proposes the structures that the algorithm in section 3 uses to induce partial orders on the feasible patient conditions of a disease. Section 4 describes the tests and the results of these algorithms on three sorts of cancer. The conclusions of the work are exposed in section 5.

2 Condition-Based Prognosis

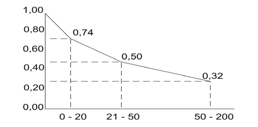
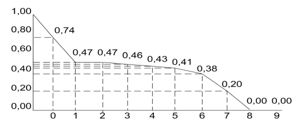
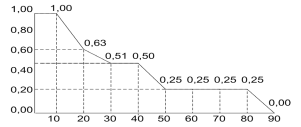
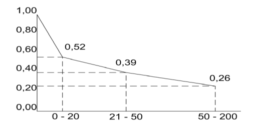
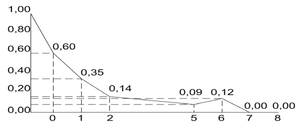
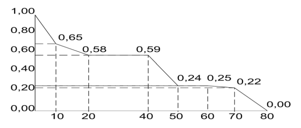
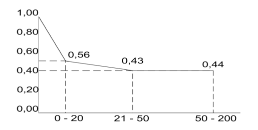
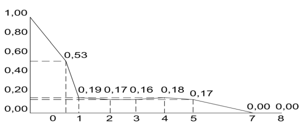
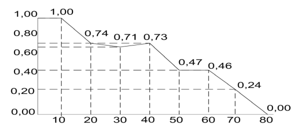
In the process of making a prognosis about the evolution of the health of a patient within a probabilistic framework, there are three main questions to be answered: what are the possible conditions of a patient in the selected disease?, what sort of order there is to compare the seriousness of these conditions?, and how the past evolutions registered in the hospital databases can be used to define a probabilistic model to support the prognostic process?

2.1 Detecting Disease Conditions

For each particular disease $\mathbb{D}$, there is a set of descriptive variables V=$\{v_1, ..., v_k\}$ with respective domains $Dom(v_i)$; i=1, ..., k. Each variable v_i represents a property of the disease that is relevant to understand the condition of the patients suffering from that disease. Each v_i defines a *severity* function s_i: $Dom(v_i)\rightarrow[0,1]$ that provides the degree of seriousness of each one of the values that the variable can take. That is to say, $s_i(v)$ is a value between zero and one representing the severity of the condition of any patient for which v_i takes the value v, zero being the lowest severity (i.e. null), and one being the highest one. *Slightness* is defined as the opposite of severity, i.e. $\mu_i(v)$=1- $s_i(v)$. For the sake of optimism, the rest of the paper will be based on the concept of slightness rather than on severities. So, Table 1 contains the slightness functions for the variables of tumour size (T), nodes (N) and metastasis (M) in the breast, lung and uterus cancer. These functions are derived from the information contained in the SEER database [7] and may vary from other sources of information.

Given a set of variables V, the condition of a patient p (or *patient condition* c_p) can be formally described as an element of the set $Dom(v_1)\times Dom(v_2)\times\ldots\times Dom(v_k)$ (i.e. c_p=(a_1, …, a_k), a_i being the value p has for variable v_i), and the *global slightness* of c_p

Table 1. Slightness functions for the variables *T*, *N* and *M* in the domains of Breast Cancer, Lung Cancer, and Uterus Cancer

	Tumour size (T)	Nodes (N)	Metastasis (M)
BREAST CANCER			
LUNG CANCER			
UTERUS CANCER			

in the disease D as a combination of all the slightness functions of the descriptive variables. Many sorts of combinations exist [1], though here only the arithmetic mean is used. So, $\mu(c_p)=1/k \cdot \Sigma_i\ \mu_i(a_i)$ is the function to calculate the global slightness of any patient condition with values a_1, …, a_k in the variables of V. This combination is possible since a correlation analysis of the data in the SEER database shows that T, N and M are mutually independent variables. Although they are not considered here, alternative combination functions should be taken if the variables to combine are not independent.

A patient condition of a disease $\mathbb{D}$ (or *disease condition* C) is defined as a restriction on the domains of the variables of that disease. So, any disease condition can be formalised as C=(D$_1$, ..., D$_k$) with D$_i \subseteq$ *Dom*(v$_i$), i=1, ..., *k*, and represents a common state of a set of patients suffering from $\mathbb{D}$. The set of all the disease conditions C$_1$, …, and C$_n$ of a disease $\mathbb{D}$ contains the alternative states in which a patient of that disease can be.

For some diseases the set of disease conditions C$_i$ are fixed and well defined, like in cancers where the *Tumour Node Metastasis Staging System* (TNM) [8] was created by the American Joint Committee on Cancer (AJCC) to describe the alternative conditions of diverse cancers; for example, the stages 0, 1, 2a, 2b, 3a, 3b, and 4 in breast cancer that Fig. 1 extends with the extreme conditions *cure* (left side C node) and *death* (right side D node).

In other diseases where there in not an agreed criterion on the set of conditions, these can be obtained from the application of a non-supervised clustering algorithm on a representative sample of *patient conditions* described in terms of the set of variables V. Two alternative sorts of clustering algorithms can be applied: data clustering and conceptual clustering. Data clustering algorithms like *kMEANS* [5] obtain clusters of similar patient conditions that are dissimilar to the patient conditions in other clusters. On the contrary, conceptual clustering algorithms like *COBWEB* [2] obtain clusters as expressions describing the patient conditions contained in the cluster, in terms of the variables in V.

The application of a clustering algorithm can be made directly on the values of the variables in V (i.e. patient respective values a_1, …, a_k) or, alternatively, on the values of the slightness functions of the variables in V (i.e. values $\mu_1(a_1)$, …, $\mu_k(a_k)$). Whereas the first option puts patient conditions with similar descriptions in the same cluster, the second group of algorithms gathers patient conditions with similar slightness values in the same cluster.

2.2 Using Partial Orders to Sort the Seriousness of the Disease

The global slightness function μ defines a complete order relation among the patient conditions that can be described in terms the variables in V. So, for any particular disease, if c_i and c_j represent two patient conditions and $\mu(c_i)>\mu(c_j)$, we interpret that c_i is better than c_j. Nevertheless, this sort of order relation cannot be extended to the comparison of disease conditions where two conditions C_i and C_j of the same disease can not only represent one a worse state than the other, but also incomparable states from the point of view of their respective slightness. This implies that, for any disease D, the order relation of the feasible disease conditions is not necessarily complete.

Formally, given a set of elements A, a *partial order* $P \subseteq A \times A$ on these elements is a binary relation such that P is reflexive (i.e. $e_i \in A \Rightarrow (e_i, e_i) \in P$), anti-symmetric (i.e. $(e_i, e_j) \in P$ and $(e_j, e_i) \in P \Rightarrow e_i = e_j$), and transitive ($(e_i, e_j) \in P$ and $(e_j, e_k) \in P \Rightarrow (e_i, e_k) \in P$).

Partial orders are typically represented as directed acyclic graphs where all the edges that are deducible by transitivity (i.e. *weak* relations) are omitted.

A set of disease conditions $\{C_1, ..., C_n\}$ on a disease D defines a partial order. This partial order can be used to know whether one condition is better or worse than other condition, or if they cannot be compared. For example, Fig. 1 depicts a directed acyclic graph that represents the standard partial order of the breast cancer conditions according to the TNM staging system [8]. It shows, for instance, that a patient in stage 2a is healthier than one patient in stage 3a or 3b (direct edge connection), or 4 (connected by edge transitivity), and not comparable in terms of slightness to patients in stage 2b.

The difference between two partial orders P_1 and P_2 can be measured in terms of the cardinality of the set $(P_1 \cup P_2) - (P_1 \cap P_2)$.

2.3 Using State-Transition Diagrams to Represent the Cases in Hospital DBs

In the previous section we showed how the conditions of a disease define a partial order of their respective slightness. This conceptual structure, however, is unable to represent the evolutions of patients in time which are based on patient improvements, worsenings and stable periods. *State-Transition Diagrams* are directed graphs that model behaviours in terms of states, transitions and actions. Here, states stand for the conditions of a disease, transitions are the evolutions of the observed patients as their conditions change in time, and actions remain unused. Formally speaking, if $\mathbb{C}$ is a set of disease conditions of a disease $\mathbb{D}$, a state-transition diagram is a pair ($\mathbb{C}$, t) such that t: $\mathbb{C} \times \mathbb{C} \rightarrow \mathbb{N}$ is the transition function that, for each couple of disease conditions

C_i and C_j in $\mathbb{C}$, $t(C_i,C_j)$ is the number of patients whose conditions evolve directly from C_i to C_j. The *inflow* and the *outflow* of a disease condition C_i can be calculated with the functions $in(C_i)=\Sigma_j\, t(C_j,C_i)$ and $out(C_i)=\Sigma_j\, t(C_i,C_j)$, respectively.

If this model is used to represent the evolutions of a set of patients across the feasible conditions of a disease, it must be extended with the *admission* and the *discharge functions* a: $\mathbb{C} \rightarrow \mathbb{N}$ and d: $\mathbb{C} \rightarrow \mathbb{N}$ such that for any condition C_i, $a(C_i)$ is the number of patients arriving in condition C_i, and $d(C_i)$ the number of patients leaving from (or still remaining in) condition C_i. See that, for any disease condition C_i, $a(C_i)+in(C_i)$ must be equal to $out(C_i)+d(C_i)$. Then, if $n_i=a(C_i)+in(C_i)$ represents the number of times any patient has been in condition C_i, and $n_t=\Sigma_i\Sigma_j\, t(C_i,C_j)$ the number of changes of disease condition of all the patients registered in a hospital database, the probability of a patient to be in condition C_i is $p(C_i)= n_i/n_t$, the probability of a patient p in condition C_i to evolve to C_j in one transition is $p(C_i,C_j)= t(C_i,C_j)\ /\ n_i$, and the probability of finding a patient that evolves from C_i to C_j is $t(C_i,C_j)\ /\ n_t$.

The above function $p(C_i,C_j)$ can be used to compute the probability of a patient to evolve from one set of disease conditions $\mathcal{A} \subseteq \{C_1, \ldots, C_n\}$ to another set of disease conditions $\mathcal{B} \subseteq \{C_1, \ldots, C_n\}$ in one step as $Pr(\mathcal{A}, \mathcal{B}) = \Sigma_{Ci\in\mathcal{A}}\, \Sigma_{Cj\in\mathcal{B}}\, p(C_i,C_j)$. In its turn, this function, together with a partial order P on the disease conditions, can be used to make prognoses on the likelihood a patient gets cured, improves, worsens, dies, or survives. See equations 1 to 5, respectively where *Condition(p)* represents the current condition of the patient, *cure* is the condition of a healthy patient, and *death* is the condition representing a deceased patient.

$$Pr(p \text{ cures}) = Pr(\{Condition(p)\},\{\text{cure}\}) \tag{1}$$
$$Pr(p \text{ improves}) = Pr(\{Condition(p)\},\{C\text{: } (Condition(p),C)\in P\}) \tag{2}$$
$$Pr(p \text{ worsens}) = Pr(\{Condition(p)\},\{C\text{: } (C, Condition(p))\in P\}) \tag{3}$$
$$Pr(p \text{ dies}) = Pr(\{Condition(p)\},\{\text{death}\}) \tag{4}$$
$$Pr(p \text{ survives}) = 1\text{-}Pr(p \text{ dies}) \tag{5}$$

3 Induction of Partial Orders

Condition-Based Prognosis as it was introduced in section 2 is a three step process that starts with the determination of the conditions of a disease (here, we will consider the set of conditions already available). Once the disease conditions are fixed, a second step takes the data of the evolutions of patients in a health-care centre to induce both a partial order on these conditions, and also a state-transition diagram that contains the probabilities $p(C_i,C_j)$ of evolving from any disease condition C_i to any other disease condition C_j in the context of the selected health-care centre. After that, a third step can be applied that consists on the utilisation of both structures to predict the evolution of new patients: the partial order provides the semantic meaning of what “cure”, “improve”, “worsen”, “die”, or “survive” means in the context of the patient current medical condition, and the state-transition diagram supplies the probabilities needed to compute the final prognostic value. This section describes the procedures to carry out the second and the third steps.

3.1 The Data Model

The two main structures used in condition-based prognosis (i.e. partial order and state-transition diagram) are generated from the same database. This database contains the data about the evolutions of the patient conditions in a health-care centre. The basic data structure is the episode of care. An *episode of care* (EOC) contains all the medical information about the treatment of one patient between the date of admission and the date of discharge. In our approach, an EOC is represented as a sequence of patient-professional *encounters* in which the professional observes the condition of the patient and proposes a course of action. Formally, if V=$\{v_1, ..., v_k\}$ is a set of descriptive variables of the patient conditions in a disease $\mathbb{D}$ and $\mathbb{A}=\{a_1, .., a_p\}$ is a set of medical actions, then an encounter e is a pair (c, a) such that c is a patient condition (i.e. $c \in Dom(v_1) \times Dom(v_2) \times \ldots \times Dom(v_k)$) and a is a subset of actions in A; an EOC is a sequence $e_1, ..., e_q$ of encounters, and the database is a list of EOCs.

3.2 The Statistical Model

According to the data structure described above, for any pair of disease conditions (C_i, C_j), we can apply a statistical procedure to determine, in a first stage, whether there is an order relation between C_i and C_j and, if there is one, in a second stage, decide which of the two conditions represents a better state of the disease from a health point of view (i.e. the order of the relation between C_i and C_j). Once all the pairs of disease conditions are considered, a *statistically significant partial order* on these conditions is obtained. Here, the above mentioned two stages are implemented as statistical hypothesis Student's t-tests.

In the study of a disease D, with $\{C_1, \ldots, C_n\}$ the set of all possible conditions of D, and provided a database containing a representative sample of encounters of all the patients that have been treated of that D, the description of the state of the patient in each encounter e_k in terms of the variables in V defines a patient condition c_k with a slightness value $\mu(c_k)$ –or $\hat{\mu}(c_k)$ in statistics notation. Simultaneously, this patient condition c_k classifies the encounter in one of the disease conditions $C_1, \ldots, C_n$.

Let us call E_k the set of the encounters in the database that are classified in C_k, and $S_k=\{\mu(c_j): e_j \in E_k\}$ the set of μ-values of their patient conditions. Then, for any pair of disease conditions C_i and C_j, the respective sets S_i and S_j are the two independent samples of a Student's t-test with null hypothesis the means of the slightness values of the elements in C_i and the elements in C_j are equal, provided that the underlying distributions are normal.

Only if the null hypothesis is rejected, C_i and C_j have an order relation whose sense is evaluated with a new Student's t-test with null hypothesis the means of the slightness values of the elements are larger in C_i than in C_j. Both t-tests are based on the t-value (6) where μ's, σ's and n's represent the mean, standard deviation, and number of elements of the samples, respectively.

$$\beta = \frac{\overline{\mu_i} - \overline{\mu_j}}{\sqrt{\frac{\sigma_i^2}{n_i - 1} + \frac{\sigma_j^2}{n_j - 1}}} \tag{6}$$

3.3 The Algorithm

An algorithm to induce partial orders under the previously described statistical model is introduced in this section. This algorithm realizes the induction process according to the data and the statistical models of sections 3.1 and 3.2, respectively. The final result of the algorithm is a partial order that explains the slightness degree of a disease in terms of the improvement or worsening between the conditions of a disease.

```
Algorithm MakePartialOrder (C, data, $\alpha$)
{Let C = {C1,…,Cn} be a set of conditions on a disease D}
{Let data    = {$EOC_1$, …, $EOC_k$} be a list of episodes of care of D}
{     $EOC_i$   = {$e_{i1}$, …, $e_{ik_i}$} the list of encounters in $EOC_i$, i=1..k}
{Let $\alpha$ the statistical significance of the test -e.g. 0.01}
   j: float
   PO = ∅; {empty partial order on the set of disease conditions C}
   For any pair of conditions (Ci, Cj) in C×C
       Ei = {$e_{xy} \in \cup_z EOC_z$: Ci is the condition of the patient in encounter $e_{xy}$}
       Ej = {$e_{xy} \in \cup_z EOC_z$: Cj is the condition of the patient in encounter $e_{xy}$}
       Si = {$\mu(c_x)$: $c_x$ is the condition of the patient in $e_x$, for all $e_x \in$Ei}
       Sj = {$\mu(c_x)$: $c_x$ is the condition of the patient in $e_x$, for all $e_x \in$Ej}
       Calculate the t-value $\beta$ according to equation 6
       If $|\beta| < t_{\alpha/2}$ (first hypothesis test indicates Ci and Cj are related) then
         If $\beta > t_\alpha$ (second hypothesis test indicates Ci is better than Cj) then
           Insert (Ci,Cj) in PO;
         else
           Insert (Cj,Ci) in PO;
       End If; End If;
   End For;
   Write the order relation PO;
End Algorithm.
```

4 Experiments

In order to induce partial orders, we used the databases on the diseases Breast Cancer (55939 encounters), Lung Cancer (19491 encounters) and Uterus Cancer (705 encounters) obtained from the SEER Cancer Incidence Public-Use Database [7]. These databases contain information on patient conditions based on three variables: Tumour Size, Lymph Nodes, and Metastasis classified according to the TNM System [8]. Data with unknown or missing values are removed from the databases. The distribution of these data according to each disease condition is described in Table 2.

Table 2. Distribution of episodes according to each disease condition

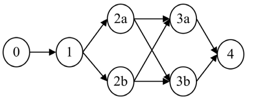
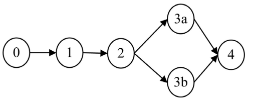
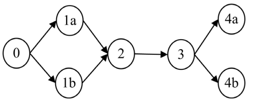
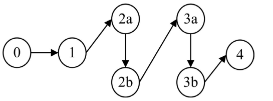
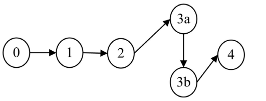
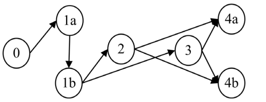
Cancer Disease	Disease conditions									Total
	0	1a	1b	2a	2b	3a	3b	4a	4b	
Breast	7073	25566		13387	6550	1456	940	967		55939
Lung	11	7298		1338		2629	3022	5193		19491
Uterus	51	242	203	79		45		5	80	705

Two sorts of tests have been performed on these databases: one that is used to compare the difference between the standard partial orders which are proposed by the TNM Staging System [8], and the experience-based partial orders obtained by the inductive algorithm introduced in section 3.3 when it is applied on the proposed databases. The second test is about how these differences affect the process of prediction on the facts of cure, improvement, worsening, death, and survival in breast, lung, and uterus cancers.

4.1 Results on the Induction of Partial Orders

Table 3 shows both the standard partial orders [8] and the partial orders the proposed algorithm induces form the three databases. The distances between the standard and the induced partial orders are 2, 1 and 2, respectively. These differences are caused either by the detection of new relations that were not present in the standard partial order or by the elimination of relations that do not achieve the statistical significance level required to be part of the experience-based partial order. So in breast cancer, the relations 2a-2b and 3a-3b are statistically justified though they were not in the standard partial order. A similar case is observed in lung cancer with the relation 3a-3b, and in uterus cancer with relation 1a-1b. In this last domain, the SEER database does not provide enough evidence to keep the standard order relation between stages 2 and 3 in the experience-based partial order.

Table 3. Partial orders induced

	Breast Cancer	**Lung Cancer**	**Uterus Cancer**
Standard Partial Order			
Experience-Based Partial Order			

These single differences between standard and experience-based partial orders are cause of new differences when the transitivity property is applied, and the final differences increase to 3%, 2%, and 10% of the total number of binary relations, this meaning that 3, 2, and 10 out of 100 comparisons get different responses whether the standard or the experience-based partial orders are queried.

4.2 Results on the Condition-Based Prognosis

Equations 1 to 5 in section 2.3 are used to calculate the probabilities of improvement, worsening, cure, death and survival in Breast, Lung and Uterus cancers for both, the standard partial order, and the experience-based partial order the algorithm in section 3.3 obtains for the data of the SEER repository [7], representing real patients.

	BREAST CANCER							**STND**		**EXP.B**	
	0	**1**	**2a**	**2b**	**3a**	**3b**	**4**	**I**	**W**	**I**	**W**
0	1	0	0	0	0	0	0	*0*	*0*	*0*	*0*
1	0.6	0.2	0.1	0.1	0	0	0	*0.75*	*0.25*	*0.75*	*0.25*
2a	0.3	0.2	0.1	0.2	0.1	0.1	0	*0.64*	*0.36*	*0.55*	*0.44*
2b	0.1	0.1	0.2	0.1	0.1	0.2	0.2	*0.36*	*0.64*	*0.44*	*0.55*
3a	0	0	0.1	0.1	0.3	0.2	0.3	*0.39*	*0.61*	*0.29*	*0.71*
3b	0	0	0	0.2	0.3	0.1	0.4	*0.25*	*0.65*	*0.55*	*0.44*
4	0	0	0	0	0	0	1	*0*	*0*	*0*	*0*

	LUNG CANCER						**STND**		**EXP.B**	
	0	**1**	**2**	**3a**	**3b**	**4**	**I**	**W**	**I**	**W**
0	1	0	0	0	0	0	*0*	*0*	*0*	*0*
1	0.8	0.1	0.1	0	0	0	*0.88*	*0.11*	*0.88*	*0.11*
2	0.1	0.3	0.1	0.2	0.2	0.1	*0.44*	*0.55*	*0.44*	*0.55*
3a	0	0.1	0.2	0.1	0.2	0.4	*0.39*	*0.61*	*0.33*	*0.66*
3b	0	0	0.1	0.3	0.1	0.5	*0.24*	*0.76*	*0.44*	*0.55*
4	0	0	0	0	0	1	*0*	*0*	*0*	*0*

	UTERUS CANCER							**STND**		**EXP.B**	
	0	**1a**	**1b**	**2**	**3**	**4a**	**4b**	**I**	**W**	**I**	**W**
0	1	0	0	0	0	0	0	*0*	*0*	*0*	*0*
1a	0.4	0.3	0.1	0.1	0.1	0	0	*0.68*	*0.32*	*0.57*	*0.43*
1b	0.4	0.4	0.1	0	0.1	0	0	*0.75*	*0.25*	*0.88*	*0.11*
2	0.1	0.1	0.1	0.1	0.2	0.2	0	*0.33*	*0.67*	*0.58*	*0.42*
3	0	0	0.1	0.1	0.1	0.2	0.5	*0.22*	*0.78*	*0.32*	*0.68*
4a	0	0	0	0.1	0.1	0.2	0.6	*0.25*	*0*	*0.25*	*0*
4b	0	0	0	0	0	0	1	*0*	*0*	*0*	*0*

Fig. 2. Probabilities of evolution among disease conditions in breast, lung, and uterus cancers

In order to analyse the differences between the prediction values obtained with the utilisation of either the standard or the experience-based partial orders, the probabilities $p(C_i,C_j)$ that are obtained from the real evolution of a set of patients, are used to define a matrix of patient evolutions. Fig. 2 shows the probability matrices employed to analyse these differences in the cases of breast, lung, and uterus cancers.

The probabilities of cure, death, and survival are identical for the standard and the experience-based partial orders, as expected, since the conditions of *cure* and *death* are the same in both partial orders. However, the predictions on improvement (I) and worsening (W) differ if we use one or the other partial orders, as the numbers in grey indicates. Some of these differences cause the prognostic with the standard partial order to provide excessive "hope" (e.g. in uterus cancer, patients in stage 1a are given 68% of improvement, whereas the experience says that only 57% will improve), or excessive "despair" (e.g. in uterus cancer, patients in stage 2 get 67% of worsening, when reality shows that it is only 42%).

5 Conclusions

In this paper, we have introduced a method to induce partial orders for patient conditions in a disease, which is part of a broader work in the area of machine learning to support healthcare activities [6]. Here, the partial orders which are built from real experiences happened in health-care centres show the gap there is between the criteria to assess the patient condition proposed by medical experts (standard partial order), and the criteria coming out of the medical daily situations (experience-based partial order).

From the tests described in the previous section, we can conclude there are clear structural differences between the standard partial orders proposed by the physicians and those others that are induced from the data of the SEER repository about real patients. A direct implication of these differences is that the prognosis about the evolution of patients may change drastically. This effect has been confirmed with the results of the tests performed which may drive the physician to incorrect predictions of patient future improvements and worsenings.

Acknowledgment. This work has been partially funded by the European IST Project K4CARE (IST-026968) and the Spanish Project HYGIA (TIN2006-15453-c04).

References

1. Figueira, J., Greco, S., Ehrgott, M. (ed.): Multiple Criteria Decision Analysis. State of the Art Surveys. Springer's International Series, New York (2005)
2. Fisher, D.: Knowledge acquisition via incremental conceptual clustering. Machine Learning 2, 139–172 (1987)
3. Lucas, P., Ameen, A.-H. (ed.): Prognostic Methods in Medicine. Artificial Intelligence in Medicine vol. 15, pp. 105–119 (1999)
4. Machado, O.L.: Methodological Review: Modelling Medical Prognosis: Survival Analysis Techniques. Journal of Biomedical Informatics 34, 428–439 (2001)

5. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Procs of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1st edn., pp. 281–297. University of California Press, Berkeley (1967)
6. Riaño, D., Bohada, J.A., Welzer, T.: The DTP model: Integration of intelligent techniques for the decision support in Healthcare Assistance. EIS2004 (2004)
7. SEER Cancer Statistics Review. Surveillance, Epidemiology, and End Results (SEER) program public-use data (1973-2003). National Cancer Institute, Surveillance Research program, Cancer Statistics Branch, released April 2006, based on the (November 2005 submission) http://www.seer.cancer.gov
8. Sobin, L.H., Wittekind, C.: TNM Classification of Malignant Tumours, 6th edn. John Wiley & Sons, New Jersey (2002)

Interacting Agents for the Risk Assessment of Allergies in Newborn Babies

Giorgio Leonardi[1], Silvana Quaglini[1], Mara de Amici[2],
Mario Stefanelli[1], Cristina Torre[3], and Giorgio Ciprandi[4]

[1] Department of Computer Science and Systems, Università di Pavia, Italy
[2] Policlinico S. Matteo, Pavia, Italy
[3] Department of Pediatrics, IRCCS Policlinico S. Matteo, Pavia, Italy
[4] Department of Internal Medicine, AOU San Martino, Genova, Italy

Abstract. Allergic diseases are increasing all over the world. Therefore, the risk assessment of allergy in newborns is a key issue for prevention purposes. The risk can be assessed at the birth by combining information about familiarity with results of blood examination. Then, the individual must be monitored, particularly in the fist months of life, in order to better define the type of allergy and the risk. The monitoring is carried on by different professionals (agents), therefore the communication and collaboration between these agents must be supported in order to obtain the best treatment strategy for the baby. This paper presents a new project which allows the cooperation between the agents involved in the risk assessment of allergy in newborn babies, and presents the main technologies which will be used to develop it.

1 Introduction

All over the world there is evidence that allergic diseases are increasing, and most of the allergies develop within the first decade of life. Among possible reasons for that, scientific debate in the literature reports air pollution, lifestyle, medication abuse, vaccinations, food additives, but weak evidence exist for definite answers. Diagnosis of allergy or of atopy (i.e. the tendency to develop immediate hypersensitivity to allergens) is a real challenge. Although familiarity is important, it is often extremely difficult to collect a detailed and focused family clinical history and to establish relationships between reported disturbances and allergies and interactions with other diseases. That's why more specific diagnostic tests must be performed. The two main types of diagnostic procedures are skin tests (in vivo) and blood tests (in vitro measure of immunoglobulin IgE), both measuring reaction of the patient's antibodies to a set of allergens. Even if skin testing has been for many years the most common screening method for allergy evaluation, blood tests are much more comfortable and quick, thus they are becoming more and more prescribed, mainly by non-allergists (pediatrician and general practitioners). The problem addressed in this paper is that allergens are hundreds, and the tendency of non-specialists is to test the biggest possible number of them. This trend caused a dramatic increase of allergy-related healthcare costs.

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 500–505, 2007.

Moreover, it does not necessarily increase the correctness of diagnosis, because of multiplication of false positive results. We propose a project for developing a network among neonatology departments, institutional healthcare agencies, pediatricians, General Practitioners (GPs), ambulatories and families, in order to suggest the best diagnostic and follow-up procedures.

2 Architecture of the System

The project described involves several actors (or agents) belonging to different organizations. The aim is to create a communication channel which permits the cooperation of the agents and the exchange of information and documents.

2.1 Serviceflow Management System

A Serviceflow Management System (SMS) is a system able to manage the overall care delivery process by establishing a tight link between different organizational units and professionals. The design of such a system is complex because there are several requirements to consider [2], and the agents involved in the care process must agree on the choice of IT support, on the communication protocols and on the timings and modalities of the service delivery. The advantage of the Serviceflow approach is twofold: from the patient's point of view, the coordination and synchronization of all the services are assured; from the organizations' point of view it permits to separate the definition of the activities provided from the method of providing them. The concept of Serviceflow relies on the concept of Service Point (SP): a SP is a "place" where the consumer and the provider of the service meet. Here, based on previous agreements, the consumer asks for a service and waits the producer to fulfil it. The SPs permit to coordinate the activities of the organizations ensuring their autonomy. SPs are characterized by pre- and post-conditions [3] that are respectively input and output parameters to be verified respectively to start the service and to determine the SP's success. These conditions enable the right service at the right time. By means of the SPs it is possible to coordinate the work activities, to monitor the work during its execution, to validate the executed activities and to manage dynamic process changes and exceptions.

A Serviceflow "is the successive interrelation of a number of Service Points" [2]. The execution of a Serviceflow, exploiting the agreements and the pre- and post-conditions, generates the correct flow of services for the patient.

2.2 Architecture

The architecture of the system is the result of our previous studies [5]. This architecture, shown in Fig. 1, is composed three different levels.

Each organization (Organizational Units level) manages its activities with private processes, (i.e. applications and/or workflow systems [4]). Fragments of these processes are published as Service Processes (SPRs) at the Service level,

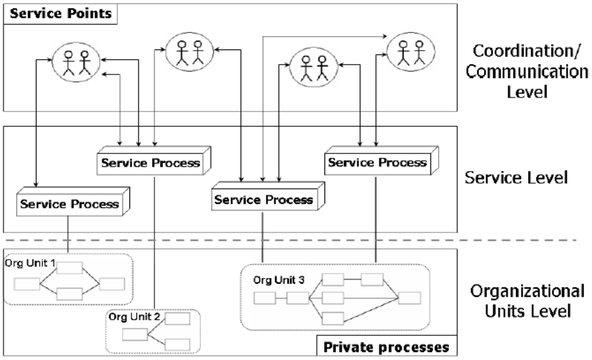


Fig. 1. Architecture of the system

acting as a public interface. A SPR is a public abstraction of an activity carried out by an organizational unit. A SPr definition contains also the conditions to be respected together with information about the provider. At the Coordination/Communication level [1], there are the SPs, which coordinate and synchronize any interaction exploiting the SPRs defined. As highlighted by the dotted line in Fig. 1, the model clearly separates the organization offer from its implementation in order to meet the privacy needs and the implementation choices of every organization unit.

On the basis of this general architecture, we designed a system which manages the network of agents involved in the assessment and treatment of allergic or atopy diseases of babies.

3 Application

The goal of this project is the creation of a collaborative network involving all the agents working on the diagnosis and the setup of the follow-up procedures for the babies considered at risk of breaking out allergies during the first ten years of life. This net is composed by: neonatology units, allergologic units, health-care institutional units, pediatricians, general practitioners and families of the babies. This application mainly focuses on the diagnostic procedures to assess the presence of allergies in the baby, or the possibility to break out allergies during the first years of life. The diagnostic strategy is carried out in three phases: (1) at the neonatology unit, a first evaluation is carried out by studying the familiarity and number of total IgE. If the risk is high, the baby is labeled as "positive", otherwise the label is "negative"; (2) if the label is "positive", the pediatrician carries out further investigations, monitoring the baby with appropriate tests until he/she turns 4; (3) when the baby turns 4, the diagnostic strategy implies the use of different tests and examinations, carried out by the GP when the baby turns 6.

Tests and examinations are carried out on a regular basis, according to predefined guidelines. During the phases presented above, The baby meets mainly three different agents: a newborn specialist in the first, the pediatrician in the second and the GP in the third. This suggests to define every phase as a SPs of a Serviceflow at the *Coordination/Communication* level of the architecture shown in fig. 1. An additional SP permits the contacts between an agent at the neonatology unit and the provider of the data base containing all the baby's clinical information. These contacts allow the agents at the neonatology unit to receive results and outcomes of the tests and exams carried out and use these data to correct the initial labeling strategies in order to reduce the number of false positives and false negative cases. The interrelation of these SPs generates the Serviceflow shown in Fig. 2.

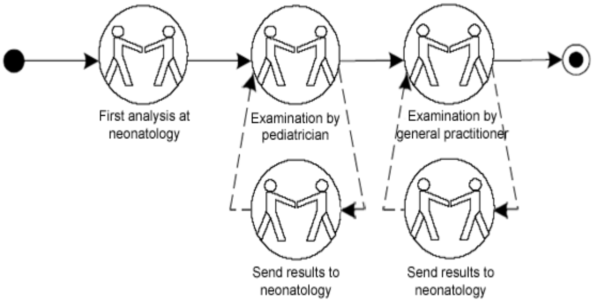

First analysis at neonatology
Examination by pediatrician
Examination by general practitioner
Send results to neonatology
Send results to neonatology

Fig. 2. Serviceflow for the risk assessment

Every phase described above is guided by a different guideline. Every guideline will be executed when the related SP is activated. Using this technique, we can offer the patient the proper services depending on his/her age and label, and execute the services with the proper guideline. The rest of this section describes one of the guidelines formalized with the tool *Guide* [6]. The implementation of these guidelines will compose the *Organizational Units* level of the architecture in fig. 1.

- *Guideline associated with the "examination by pediatrician" SP*

This guideline describes the management of newborns and babies under 4 years of age, carried out by the pediatrician, after the first analysis at the neonatology.

If the baby is symptomatic, the pediatrician checks her/his label: if it is negative or unknown, the test "Phadiatop level I" is carried out and his label may change; otherwise a "level II" test is performed.

If the baby is asymptomatic: (1) if the label is positive and one year has passed from the last test, a "Phadiatop level I" is carried out again. This allows performing the test every year on the babies labeled positive; (2) if the label is unknown and age is under 12 months, a label is assigned on the basis of the results of a a "Phadiatop level I"; (3) if the label is negative, or the age is over 12 months, the guideline stops.

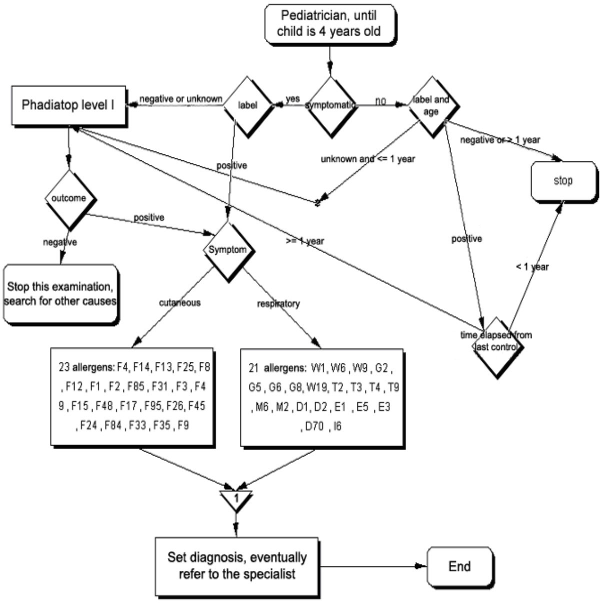

Pediatrician, until child is 4 years old
Phadiatop level I
negative or unknown
label
yes
symptomatic
no
label and age
negative or > 1 year
unknown and <= 1 year
stop
positive
outcome
positive
negative
Symptom
>= 1 year
positive
< 1 year
Stop this examination, search for other causes
cutaneous
respiratory
time elapsed from last control
23 allergens: F4, F14, F13, F25, F8 , F12 , F1 , F2 , F85 , F31 , F3 , F49 , F15 , F48 , F17 , F95, F26, F45 , F24 , F84 , F33 , F35 , F9
21 allergens: W1, W6 , W9 , G2 , G5 , G6 , G8, W19, T2 , T3 , T4 , T9 , M6 , M2 , D1, D2 , E1 , E5 , E3 , D70 , I6
1
Set diagnosis, eventually refer to the specialist
End

Fig. 3. Guideline 2: examination by pediatrician

The system proposed in this project manages the contacts between the baby and the agents involved in his/her care process, the exchange of information and documents between the agents and the coordination of the diagnostic activities in every phase. The contacts between the baby and the agents and the exchange of information are provided through the Serviceflow shown in Fig. 2, while the diagnostic activities are driven by guidelines like the one shown in Fig. 3. This system enforces the cooperation between the agents involved and the application of the protocols (guidelines) defined in order to offer a better diagnosis and treatment services without wasting resources in unuseful tests.

4 Conclusion

The challenge of the project we propose is to demonstrate that a Serviceflow Management System, associated to specific workflows based on clinical practice guidelines developed by expert allergologists, may be of great benefit for the national healthcare system, improving diagnostic procedures and follow-up and decreasing diagnostic tests expenditures. We are aware of the several issues that must be solved to develop such a system, first of all the integration with existing information systems that are currently used at the different levels (hospital units, healthcare agencies, regional level, etc), but also we are confident that the proposed architecture could harmonize those systems and foster collaboration.

References

1. Perrin, O., Godart, C.: A model to support collaborative work in virtual enterprises. Data Knowledge Engineering 50(1), 63–86 (2004)
2. Wetzel, I., Klischewski, R.: Serviceflow beyond Workflow? Concepts and Architectures for Supporting Inter-Organizational Service Processes, Advanced Information Systems Engineering. In: Proceedings 14th CAiSE, Springer Lecture Notes in Computer Science, Berlin, pp. 500–515 (2002)
3. Klischewski, R., Wetzel, I., Baharami, A.: Modeling Serviceflow. Information Systems Technology and its Applications. In: Proceedings ISTA, German Informatics Society, Bonn, June 2001, pp. 261–272 (2001)
4. Panzarasa, S., Maddè, S., Quaglini, S., Pistarini, C., Stefanelli, M.: Evidence-based careflow management systems: the case of post-stroke rehabilitation. Journal of Biomedical Informatics 35(2), 123–139 (2002)
5. Leonardi, G., Panzarasa, S., Quaglini, S., Stefanelli, M., van der Aalst, W.M.P.: Interacting agents through a web-based health serviceflow management system, Journal of Biomedical Informatics (inpress, 2007) doi:10.1016/j.jbi.2006.12.002
6. Ciccarese, P., Caffi, E., Boiocchi, L., Quaglini, S., Stefanelli, M.: A guideline management system. In: Medinfo 2004, pp. 28–32 (2004)

Author Index