Proceedings of the 1st

International Workshop on Artificial Intelligence and NetMedicine

Montpellier, France

27 August 2012

In conjunction with

ECAI 2012

Proceedings of the 1st

International Workshop on Artificial Intelligence and NetMedicine

Montpellier, France

27 August 2012

In conjunction with

ECAI 2012

1st International Workshop on Artificial Intelligence and NetMedicine

NetMed 2012

Table of Contents

Message from the Program Chairs

Medical telereporting and second-opinion over the Internet are nowadays cost-effective and widely adopted practices. Physicians and general practitioners make daily use of tele-consultation over the WEB, VOIP, chat and video-conferencing.

Social networking favors the constitution of large communities of members sharing similar medical interest, so that TeleMedicine is rapidly turning into what we call "**NetMedicine**", which simply denotes every Health-related activity which is carried on through the Internet.

Since its inception and along all its history, Artificial Intelligence served the Medicine, under both its souls, the logicistic and the connessionistic ones. But in the current digitally networked and hyperlinked e-Health scenario, Artificial Intelligence has to play also new important roles. Today we urge intelligent software to semantically interpret and filter diagnostic data, automatically classify and convey medical information, virtualize nurses and hospital lanes to reduce the costs of healthcare, etc.

The **International Workshop on Artificial Intelligence and NetMedicine** (NetMed) aims at bringing together scholars and practitioners active in Artificial Intelligence driven Health Informatics, to present and discuss their research, share their knowledge and experiences, define key research challenges and explore possible international collaborations to advance the intelligent practice of Medicine over the Internet.

The NetMed Workshop collects original contributions on research and application aspects of Artificial Intelligence driven e-Health. In particular, areas of interest include:

- Tele-Health and Telemonitoring over the Internet
  - Collaborative care and communication
- Intelligent devices and instruments
  - Ontology modeling and reasoning in Health Information Engineering and Systems
  - SNOMED CT
- Patient care, monitoring and diagnosis
  - AI-based clinical decision making
  - Clinical Evidence-Based decision support systems
- Architectures of Electronic Health Records
- AI in medical education
- Medical knowledge engineering
  - Medical data mining
- Modelling and simulation
- Implementation and case studies
- Intelligent Visualization in Medicine
- Intelligent Medical Information Systems

    - Intelligent health records
- Automated Reasoning and Metareasoning in Medicine
- Philosophical, Ethical, and Social issues of AI in Medicine
- Extending quality healthcare to rural communities
- Health Informatics in the developing world

We would like to thank the ECAI organization for having allowed us to organize this event. We would like to thank all the authors for having submitted their work to the workshop for selection, the Program Committee members for their effort in reviewing the papers, the presenters for ensuring interesting sessions, and the attendees for participating into this event.

We hope that interesting ideas and discussions will come out of the presentations, demos and the questions that will alternate along the day. We hope you will find this day interesting and enjoyable.

Aldo Franco Dragoni, Università Politecnica delle Marche, Italy
Roberto Posenato, Università degli Studi di Verona, Italy
NetMed 2012 Program Chairs

Keynote speaker

Artificial-intelligence-augmented clinical medicine

Klaus-Peter Adlassnig
Section for Medical Expert and Knowledge-Based Systems
Center for Medical Statistics, Informatics, and Intelligent Systems
Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria
klaus-peter.adlassnig@meduniwien.ac.at

Abstract

Background
Nowadays, clinical decision making is increasingly based on a large amount of patient medical data, on continuously growing medical knowledge, and on extended best clinical practice guidelines.

Clinical decision support
There is evidence that clinical decision support systems can significantly improve quality of care in, eventually, all areas of clinical medicine [1]. Technically, suitable means to formally represent clinical knowledge and to connect decision support algorithms with patient data sources in a seamless way are prerequisites for successful clinical decision support applications.

Clinical decision support server and Arden Syntax
Arden Syntax, as an internationally standardized formal language for medical knowledge representation and processing [2–4], was implemented as a clinical decision support server and equipped with service-oriented interoperability [5]. This technical solution has already been proven to be deployable in connection with hospital and intensive care information systems and practicable useful in a number of clinical areas [6]. Telemedical and mHealth systems also participate in this technological advance [7].

Routinely-used, fully automated, knowledge-based system for detection and continuous monitoring of ICU-acquired infections
An example for extended clinical decision support in infection control is given by Moni/Surveillance-ICU, a system for the early recognition and the automated monitoring of hospital-acquired infections in intensive care units with adult patients [8–11]. This knowledge-based system includes concepts of fuzziness to formally represent medical linguistic terms. The European Centre for Disease Prevention and Control (ECDC) criteria

for hospital-acquired infections [12] form the basis of its knowledge base; results are given in form of degrees indicating to which extent the ECDC definitions are fulfilled by the patient data taken into account.

Artificial-intelligence-augmented clinical medicine

Today, clinical decision support technology becomes integrated in or connected with various health care information systems such as hospital, laboratory, and intensive care information systems, electronic health record, telemedicine, and web-based systems. Thus, many forms of clinical decision support in the diagnostic and therapeutic process render possible, for instance, clinical reminders, alerts, recommendations, support in differential diagnosis, therapy selection, and patient management according to guidelines and protocols. In this context, Arden Syntax, or its extended form Fuzzy Arden Syntax [13, 14], seems highly suitable for developing clinically useful decision support systems. Soon, a new type of proactive clinical information systems will become available. Through web-services, a globally available medical knowledge grid—adapting its content to the individual parameters of the patient—will eventually emerge.

References:

[1] Kawamoto, K., Houlihan, C.A., Balas, E.A. & Lobach, D.F. (2005) Improving Clinical Practice Using Clinical Decision Support Systems: A Systematic Review of Trials to Identify Features Critical to Success. British Medical Journal 330(7494), 765–768.

[2] Hripscak, G. (1994) Writing Arden Syntax Medical Logic Modules. Computers in Biology and Medicine 24, 331–363.

[3] Health Level 7. The Arden Syntax for Medical Logic Systems, Version 2.7. Ann Arbor, MI: Health Level Seven, Inc., 2008.

[4] Samwald, M., Fehre, K., de Bruin, J. & Adlassnig, K.-P. (2012) The Arden Syntax Standard for Clinical Decision Support: Experiences and Directions. Journal of Biomedical Informatics 45, 711–718.

[5] Fehre, K. & Adlassnig, K.-P. (2011) Service-Oriented Arden-Syntax-Based Clinical Decision Support. In Schreier, G., Hayn, D. & Ammenwerth, E. (Eds.) Tagungsband der eHealth2011 – Health Informatics meets eHealth – von der Wissenschaft zur Anwendung und zurück, Grenzen überwinden – Continuity of Care, 26.–27. Mai 2011, Wien, Österreichische Computer Gesellschaft, Wien, 123–128.

[6] Adlassnig, K.-P. & Rappelsberger, A. (2008) Medical Knowledge Packages and their Integration into Health-Care Information Systems and the World Wide Web. In Andersen S.K., Klein, G.O., Schulz, S., Aarts, J. & Mazzoleni, M.C. (Eds.) eHealth Beyond the Horizon–Get IT There. Proceedings of the 21st International Congress of the European Federation for Medical Informatics (MIE 2008), IOS Press, Amsterdam, 121–126.

[7] Rudigier, S., Brenner, R. & Adlassnig, K.-P. (2010) Expert-System-Based Interpretation of Hepatitis Serology Test Results as App Store iPhone Application. In Schreier, G., Hayn, D.

& Ammenwerth, E. (Eds.) Tagungsband der eHealth2010 – Health Informatics meets eHealth – von der Wissenschaft zur Anwendung und zurück, Der Mensch im Fokus, 6.–7. Mai 2010, Wien, Österreichische Computer Gesellschaft, Wien, 235–240.

[8] Adlassnig, K.-P., Blacky, A. & Koller, W. (2008) Fuzzy-Based Nosocomial Infection Control. In Nikravesh, M., Kacprzyk, J., and Zadeh, L.A. (Eds.) Forging New Frontiers: Fuzzy Pioneers II – Studies in Fuzziness and Soft Computing vol. 218, Springer, Berlin, 343–350.

[9] Adlassnig, K.-P., Blacky, A. & Koller, W. (2009) Artificial-Intelligence-Based Hospital-Acquired Infection Control. In Bushko, R.G. (Ed.) Strategy for the Future of Health, Studies in Health Technology and Informatics 149, IOS Press, Amsterdam, 103–110.

[10] Blacky, A., Mandl, H., Adlassnig, K.-P. & Koller, W. (2011) Fully Automated Surveillance of Healthcare-Associated Infections with MONI-ICU – A Breakthrough in Clinical Infection Surveillance. Applied Clinical Informatics 2(3), 365–372.

[11] De Bruin, J.S., Adlassnig, K.-P., Blacky, A., Mandl, H., Fehre, K. & Koller, W. (2012) Effectiveness of an Automated Surveillance System for Intensive Care Unit-Acquired Infections. Journal of the American Medical Informatics Association, doi:10.1136/amiajnl-2012-000898.

[12] European Centre for Disease Prevention and Control (ECDC). Healthcare-associated Infections Surveillance Network (HAI-Net). http://ecdc.europa.eu/en/activities/surveillance/HAI/Pages/default.aspx.

[13] Vetterlein, T., Mandl, H. & Adlassnig, K.-P. (2010) Fuzzy Arden Syntax: A Fuzzy Programming Language for Medicine. Artificial Intelligence in Medicine 49(1), 1–10.

[14] Vetterlein, T., Mandl, H. & Adlassnig, K.-P. (2010) Processing Gradual Information with Fuzzy Arden Syntax. In Safran, C., Reti, S. & Marin, H. (Eds.) Proceedings of the 13th World Congress on Medical Informatics (MEDINFO 2010), Studies in Health Technology and Informatics 160, IOS Press, Amsterdam, 831–835.

Program Committee

Paolo Giorgini, Università di Trento, Italy
Femida Gwadry-Sridhar, University of Western Ontario, Canada
Roque Marín Morales, University of Murcia, Spain
Niels Peek, University of Amsterdam, The Netherlands
Catherine Garbay, French National Centre for Scientific Research, France
Basilio Sierra, University of the Basque Country, Spain
Hameedullah Kazi, Isra University, Hyderabad, Pakistan
Yuval Shahar, Ben Gurion University, Israel
Allan Tucker, Brunel University, United Kingdom
Paola Mello, Università di Bologna, Italy
Carolyn McGregor, University of Ontario, Canada Research Chair in Health Informatics
Paolo Terenziani, Università di Torino, Italy
Steve Rees, Aalborg University, Denmark.
Evelina Lamma, Università di Ferrara, Italy
Zhiming Liu, Unitenations University, Macao
Paul de Clercq, University of Eindhoven, The Netherlands
Peter Lucas, Radboud University Nijmegen, The Netherlands
Xiaohui Liu, Brunel University, United Kingdom
Rainer Schmidt, Universität Rostock, Germany
Domenico Massimo Pisanelli, Consiglio Nazionale delle Ricerche, Italy
John Fox, Oxford University, United Kingdom
Henrik Eriksson, Linköping University, Sweden
Siriwan Suebnukarn, Thammasat University, Thailand
Constantine D. Spyropoulos, NCSR "Demokritos", Greece
Werner Horn, Medical University of Vienna, Austria
Andrea Claudi, Università Politecnica delle Marche, Italy

Leveraging a social network of peers for NetMedicine: personalizing the selection of web objects for improved health education

John Champaign and Robin Cohen

University of Waterloo {jchampai, rcohen}@uwaterloo.ca

Abstract. In this paper, we present an approach for reasoning about which web objects from an existing repository should be presented to patients who are trying to learn how to manage a particular medical problem. Our approach models the benefits in learning gained by socially connected peers in order to then recommend those objects predicted to offer the best gains in knowledge to the new user. This is achieved in a framework where the past learning gains of peers are modeled and recorded with the objects. In essence, we leverage techniques from the subfield of intelligent tutoring, considering our users as students who are learning to manage their healthcare. While the value of our approach has already been confirmed through simulations of student learning, we move forward in this paper to conduct a study with human users. We demonstrate the effectiveness of our algorithms for selection of web objects for the learning of users, compared to other algorithms which employ a less principled approach for content selection. This is done for the specific healthcare challenge of assisting caregivers of autistic children. We provide compelling evidence for the value of our proposed vision for telehealth: one where peers sharing a medical interest can be successfully leveraged in order to effectively inform new users in the management of their healthcare.

Keywords: social networks with common medical interests, collaborative care, extending quality care

1 Introduction

With an aging population, home healthcare solutions are becoming, by necessity, more prevalent. Caregivers and patients alike face the challenge of making medical decisions in dynamically changing environments, using whatever resources are available in the home. With copious amounts of information (e.g. text, videos, interactive systems) users benefit from methods for effectively focusing on what would be most beneficial to view.

Our research aims to provide important decision-making support in these scenarios by leveraging the learning of peers through a social networking approach. In particular, we propose that peer-based tutoring form the basis of the information imparted to homecare caregivers and patients. Distinct from other

approaches to peer-based intelligent tutoring which assume an active social network of information exchange in real-time (e.g. [1]), we propose a framework that makes use of learning experienced by peers at several points in the past and allow these peers to streamline content that will be shown to future students. In essence, we seek to adopt an approach to learning that respects what McCalla has referred to as the ecological approach [2]: enabling various learning objects (texts, videos, book chapters) and adjusted versions of these objects to be introduced to peers, based on the past experiences of other, similar, students with this media content (or learning objects).

An example scenario helps to motivate our research. Consider a diabetic patient, attempting to manage his disease. Distinct from an approach of simply posting a query to a discussion group and receiving various responses from peers (with varying degrees of reliability), one would treat this problem as one of properly teaching the patient suitable information that may be contained in a variety of online articles or instructional videos. We assume a corpus of these learning objects exists and has been experienced by other peers in the past. Pre- and post-testing of the learning achieved by these peers is conducted (for example, through an exit quiz that results in a level of understanding represented as a grade achieved, before and after the interacting with the learning object). Then, each learning object has stored with it the users who have experienced it, along with the benefit that each user obtained (an increase, or decrease, in grade level achieved).

In determining which learning object to display to a new user, we propose two distinct methods. The first focuses on presenting to new students those learning objects which produced the most benefit to like-minded peers, where the similarity between students is determined on the basis of their overall level of knowledge. This approach is motivated by collaborative filtering techniques, as performed in recommender systems [3]. For example, those learning objects which resulted in a weak understanding for other similar patients would be avoided for the new student. This system allows the object that is best suited to a particular student population to be shown to them.

Our second focus concerns the situation where there may be a particular article in a book (or some other subset of a larger learning object) on managing healthcare, which is of special value. As with our algorithm for recommending learning objects, the determination of which of these smaller articles to present to a peer will be based on the learning that is experienced by others. The object would be added to the corpus and then its overall benefit to peers can be tracked. It is possible that for one population of (perhaps more advanced) students a more targeted, succinct learning object would be preferable, while for another population of students a learning object with additional explanations may be preferable. In addition, one can manage the entire corpus by eventually discarding learning objects that are not of use (garbage collection), resulting in a refined and more valuable corpus on which the learning may proceed.

In all, we believe that home healthcare can be improved by enabling patients and caregivers to learn on the basis of the past learning of their peers, through

Algorithm 1 Pseudocode For Collaborative Learning Algorithm (CLA)

```
Input the current-student-assessment
for each learning object: do
  Initialize currentBenefit to zero
  Initialize sumOfBenefits to zero
  Input all previous interactions between students and this learning object
  for each previous interaction on learning object: do
    similarity = calculateSimilarity(current-student-assessment, interaction-initial-assessment)
    benefit = calculateBenefit(interaction-initial-assessment, interaction-final-assessment)
    sumOfBenefits = sumOfBenefits + similarity * benefit
  end for
  currentBenefit = sumOfBenefits / numberOfPreviousInteraction
  if bestObject.benefit < currentBenefit then bestObject = currentObject
end for
if bestObject.benefit < 0 then bestObject = randomObject
```

judicious choice of material to present to the learners, which evolves over time as the learning experiences of the peer group expand.

1.1 Our Approach

Our proposed algorithm for determining which learning objects to present to students is presented in Algorithm 1. We assume that we are tracking a set of values, *v[j,l]*, representing the benefit of the interaction for user j with learning object *l*. *v[j,l]* is determined by assessing the student before and after the interaction, and the difference in knowledge is the benefit. We also record for each learning object what we refer to as the *interaction history*: the previous interactions of students with that object, in terms of their initial and final assessments.[1] We assume that a student's knowledge is assessed by mapping it to 18 discrete levels: A+, A, A-, ... F+, F, F-, each representing $\frac{1}{18}$th of the range of knowledge. This large-grained assessment was used to represent the uncertainty inherent in assessing student knowledge, and only this large-grained assessment is used to reason about the students' ability in our approach. The CLA did not have access to the fine-grained knowledge values from the simulation.

The anticipated benefit of a specific learning object l, for the active user, a, under consideration would be [2]

$$p[a,l] = \kappa \sum_{j=1}^{n} w(a,j)v(j,l) \tag{1}$$

[1] The algorithm would be run after an initial phase where students are learning through the use of a set of learning objects. These students' experiences would then form the basis for instructing the subsequent students.

[2] Adapted from [3].

where *w(a,j)* reflects the similarity $\in$ (0,1] between each user j and the active user, a, and κ is a normalizing factor. $\frac{1}{|n|}$ was used as the value for κ in this work where n is the number of previous users who have interacted with learning object l. *w(a,j)* was set as $\frac{1}{1+difference}$ where difference is calculated by comparing the initial assessment of j and the current-student-assessment, and assigning an absolute value on the difference of the letter grades assigned. This is in order to obtain a similarity between 0 and 1, with 1 representing identical assessments. So the difference of A+ and B- would be 5 and the difference of D+ and C- would be 1. *v(j,l)* is also computed using a difference. Instead of a sum of the absolute differences between the initial assessments of two users, it is the sum of the difference between initial and final assessments for user j's interactions with learning object l. For example, *v(j,l)* where j is initially assessed as A+ and finally assessed at B- would be -5, while where j is initially assessed at B- and finally assessed at A+ would be 5. This is shown as the calculateBenefit function in Algorithm 1. In the absence of other criteria, a user a will be assigned the learning object l that maximizes *p[a,l]*. If the maximum *p[a,l]* is a negative anticipated benefit, a random learning object will be assigned to the user.

The CLA's value in achieving increases in knowledge to students has been confirmed by a method of simulated student learning[4, 5] achieving performance approaching that of algorithms with perfect knowledge about the students, the learning objects and the learning gains of their interactions. Below we show just one graph of results where the learning of 50 students was simulated over 100 trials with 20 iterations where the mean of the average student knowledge is mapped. Simulated students interacted with learning objects that had varying impact depending on the student's current assessment grade where a total of 100 learning objects were included in the repository. The Raw Ecological curve embodies the CLA Algorithm. The Pilot variation allows 10% of the students to prime the system first and the Simulated Annealing variation included a cooling phase where students first had a chance of being randomly assigned a learning object; both variants are done to address cold-start problems. All three variations show very effective student learning (Figure 1). In Section 3 we present a human evaluation in order to confirm the value of these methods; necessarily we are investigating a smaller sample size (i.e. we cannot easily subject participants to thousands of learning experiences nor easily manage hundreds of participants in one study). But the learning that is accomplished is now matching the ground truth for those students (revealed through performance on assessment quizzes).

2 Human Evaluation

To study the effectiveness of our approach with humans we conducted a preliminary evaluation with participants at the University of Waterloo. We chose as an application domain enabling users to learn about how to care for a child with autism (which may arise as a home healthcare scenario, of interest to projects such hSITE [6], with which we are involved). Our first step was to assemble our repository of learning objects: the material that students would learn from. In

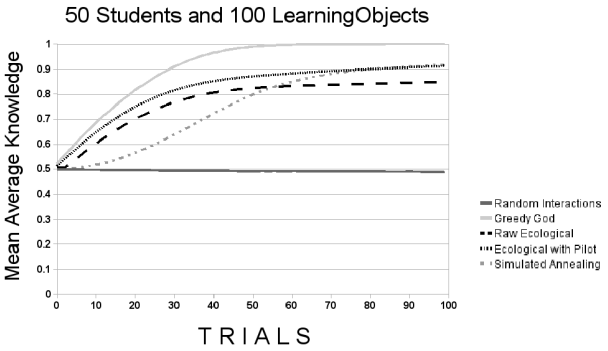


Fig. 1: Comparison of 5 Approaches for Selecting Learning Objects

collaboration with a clinical psychologist specializing in children and autism, we created 20 learning objects (16 text articles and 4 videos) that each took about 5 minutes to experience. Also in collaboration with the psychologist we created a 10-question multiple choice assessment, covering material from the learning objects. This was used to carry out the pre- and post-test assessments which serve to model learning gains in students (and form a component of our algorithm for determining which objects to present to each student). We hypothesized that a group of students using our peer-based technique for selecting learning objects would show greater learning gains than a control group that had learning objects randomly assigned to them. The aim of our study, therefore, was to validate our proposed CLA.

In order to obtain feedback from our participants about corpus division, we also explained our corpus approach to participants and then offered them the opportunity to subdivide each learning object that they were shown, as the learning proceeded.[3] We finally obtained more information during an exit survey where participants responded to questions asking them how they felt about this option of streamlining learning objects. The entire process lasted approximately 1 hour. 23 volunteers participated in our experiment, including graduate students, undergraduate students and staff members at the university. All were at least 18 years old, fluent in English and not an expert in autism spectrum disorders.

2.1 Procedure

Each participant experienced 5 learning objects and was assessed before and after each for a total of 6 assessments, with the first assessment before they experience

[3] These subdivisions were not used by other participants. Getting enough data, with the limited number of participants, to differentiate between original learning objects and streamlined versions would have been problematic.

any learning objects being a measure of the student's initial knowledge before seeing any learning objects. The assessments were the same 10 multiple choice questions each time.[4] The quiz was designed so that each question was covered well by different learning objects in the repository (and more than one learning object served to help a student to respond to that question). After experiencing each learning object, each participant did the assessment quiz and also answered a separate questionnaire allowing the student to propose a streamlining (division) of that learning object. At the end of the experiment each participant was given an exit survey asking them their overall feelings about streamlining and soliciting general feedback.

The first 12 participants were randomly assigned learning objects. They were used both as a control group and to provide training data for our technique[5]. The next 11 participants experienced a curriculum sequence provided by our approach. Participants read hardcopy articles or watched videos on a provided netbook and then a "Wizard of Oz" style study was performed. For our technique, a program was written using the CLA (Algorithm 1) and the answers provided by participants in their pre-test assessments served as the current student assessment; a new recommendation for a learning object was then determined. This sequence continued until the student had experienced five different learning objects. In essence, the first 12 participants served to prime the system for the remaining participants. After this phase, each learning object in the repository had 3 experiences recorded: while the initial control group of students were shown a random set of objects, which objects would be presented to each was determined offline in a way that ensured that each object would be shown to 3 different participants. The net-benefit obtained by each subject in the control group (number of questions correct between pre and post-test) became part of that object's interaction history. For the participants in our experimental group, determining the similarity between the current student and previous peers was measured by comparing the number of questions on the assessment that were answered identically. Only the data collected from the training group was used to make recommendations to the experimental group.[6] No learning objects were shown twice to the same participant.

[4] This was done in part to ensure that we were modeling comparable learning experiences from the participants.

[5] Individual interactions between these students and learning objects were used as training data, while the aggregate learning over the entire session was used as the control group.

[6] Had we followed our proposed approach and continually added data for the program to make recommendations from, the final participants would have been given learning objects based on a richer repository of data and the experimental group would not have been provided with a consistent treatment.

2.2 Results

Curriculum Sequencing We first compared the learning gains of our 11 experimental group participants, namely the post-test (their final assessment) minus the pre-test (their first assessment).

	Mean	s.d.	Mean (without P20)	s.d. (without P20)
Control	1.83	1.27		
Experimental	3.09	2.21	3.4	2.07

Table 1: Comparison of overall learning gains of users in each case

These results can be interpreted that, on average, participants in the control group got 1.83 more questions correct (out of 10) after completing the 5 learning objects and participants in the experimental group got an average of 3.09 more questions correct.

P20 was a participant who did not seem to be taking the experiment seriously, did not read learning objects fully and rushed through the experiment (finishing in about 40 minutes when most participants took about 1 hour). The data was analyzed with and without this participant's data included.

An a priori alpha level of significance of $\alpha = 0.05$ was used for each statistical analysis. The results were statistically reliable at $p > 0.05$ (one-sided, two samples, unequal variance t-test) which was not statistically significant. With participant 20 removed, the results were statistically reliable at $p < 0.05$ (one-sided, two samples, unequal variance t-test) which was statistically significant.

Next, we compared the proportional learning gains of participants. This was to take into consideration the suggestion of Jackson and Graesser [7] that simple learning gains are "biased towards students with low pretest scores because they have more room for improvement". This is measured using [(post-test − pretest)/(10-pretest)][7].

	Mean	s.d.	Mean (without P20)	s.d. (without P20)
Control	0.530	0.452		
Experimental	0.979	1.07	1.08	1.02

Table 2: Comparison of proportional overall learning gains of users

	Mean	s.d.	Mean (without P20)	s.d. (without P20)
Control	0.367	0.253		
Experimental	0.618	0.442	0.68	0.413

Table 3: Comparison of average learning gains of users in each case

The results were statistically reliable at $p > 0.05$ (one-sided, two samples, unequal variance t-test) which was not statistically significant. With participant 20 removed, the results were statistically reliable at $p > 0.05$ (one-sided, two samples, unequal variance t-test) which also was not statistically significant.

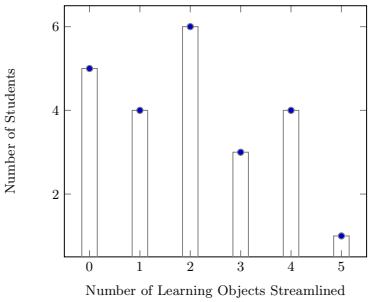
Next, we considered the per-LO learning gains of each student. Here, the change in assessment after assignment of a single learning object, were measured for each learning object experienced and the average computed. This average was then compared for the control and experimental groups.

[7] 10 is the maximum possible score on a 10 question multiple-choice quiz.

The results were statistically reliable at $p > 0.05$ (one-sided, two samples, unequal variance t-test) which was not statistically significant. With participant 20 removed, the results were statistically reliable at $p < 0.05$ (one-sided, two samples, unequal variance t-test) which was statistically significant.

Taken together, our results indicate that students presented with learning objects determined by our algorithm achieved greater learning gains than those who were randomly assigned objects.

Corpus Approach Each participant was invited, after the concept had been explained to them, to streamline learning objects according to the corpus approach.



Question	Mean	s.d.
Q1	0.227	3.35
Q2	0.864	2.949
Q3	3.773	1.232

Table 4: Mean answer values to exit survey questions

In spite of being told it was up to them whether or not to streamline learning objects, only 5 out of 23 participants declined to streamline any objects. On average, participants suggested streamlined versions for 2 of the 5 learning objects they saw.[8]

Each participant was asked 3 questions about the corpus approach during their exit survey:

1. How would you rate the difficulty of creating a new streamlined learning object?
2. How would you rate the difficulty of deciding what content to include in a streamlined version?
3. How would you rate the usefulness of a system offering a user the full version or streamlined version of content like you've seen?

Participants were given a 11 point scale, ranging from -5 to 5 with the labels "difficult" at -5, "neutral" at 0, and "easy" at 5 for Q1 and Q2 and "useless" at -5, "neutral" at 0 and "useful" at 5 for Q3.

For the 23 participants the feedback is provided in Table 4.

Although participants were mostly neutral with respect to creating streamlined versions of learning objects (Q1 and Q2), they were clearly positive about using a system where other students create streamlined learning objects for them. This conforms to research on participatory culture (e.g. [8]) which has shown

[8] In practice, participation may be lower if there isn't a researcher sitting across the table when students are deciding whether or not to streamline; however there was clearly a willingness to engage in this activity.

that consumers usually greatly outnumber contributors. It has been shown to be possible (e.g. [9]) to use incentives to encourage greater participation.

3 Conclusion and Discussion

With an overall aim of enabling effective patient-led health management, we offer here a specific approach for peer-based tutoring that makes use of a rich interaction history to personalize delivery of content for users; this serves to assist caregivers in focusing their attention on the most valuable material and demonstrates the true potential of social recommendation for this critical application area. The human study described in this paper confirms the effectiveness of the approach in achieving knowledge gains; the exit survey also support our proposal for allowing peers to augment the repository through corpus division.

Other work in the area of E-Health that has demonstrated the importance of personalized content delivery and of leveraging social networks as part of that learning (e.g. [10, 11]) focus on promoting healthier lifestyles by encouraging reflection and discussions within the family through the use of a collaborative platform. Our approach is aimed instead at allowing individuals to better understand their health concerns and make informed decisions. [12] proposes personalized delivery of video to users to educate about self-care of fibromyalgia. This work confirms several elements in our approach: including video objects, supporting personalized selection of objects from a corpus. Like us, their user study compared the value of their approach with one that was less personalized. One notable difference is that our tailoring is based on modeling peer-experiences.

Collaborative filtering recommender systems [3] also make use of content selection via modeling similarity of peers. On the surface, it might seem that recommendation techniques could be applied directly in an intelligent tutoring setting. However, whereas most recommender systems endeavour to obtain an increasingly specific understanding of a user, an intelligent tutoring system seeks both to understand a user and to enable change or growth. In contrast to positioning a user within a cluster of similar users, we model a continually evolving community of peers who are operating at a similar level of knowledge.

Previous work on collaborative learning, such as [13], has attempted to use interactions between students and the system to provide a better experience for subsequent students. The authors created a program that would capture user problem solving behaviours in the system. This data was then used to develop a tutor, in what they call "bootstrapping novice data (BND)". The authors admit that the task is non-trivial and reach the conclusion that that analysis must happen at multiple levels of abstraction. In contrast, our approach does not try to model specific user actions. Instead it pragmatically considers the sequence that learning material is experienced and how successful the students were.

Also, in contrast to efforts such as [9], in our approach each student's learning is directed by considering all experiences of previous students, thus allowing for a continuous redirection of possible content. Personalization is maintained throughout, as well. This is achieved by modeling the knowledge levels of each

student and an assessment of their current overall understanding in order to perform matching to like-minded peers, for the selection of learning objects.

Scaling is problematic for many approaches to real-time peer-tutoring (e.g. [1]). Our approach, like many ecological approaches, uses data from past interactions and performance improves as the size of the user base and repository of learning objects increases. A very large social network, therefore, is not a challenge at all, but instead an opportunity to provide highly personalized recommendations to students.

References

1. Vassileva, J.: Toward social learning environments. IEEE Transactions on Learning Technologies **1**(4) (2008) 199–214
2. McCalla, G.: The ecological approach to the design of e-learning environments: Purpose-based capture and use of information about learners. Journal of Interactive Media in Education: Special Issue on the Educational Semantic Web **7** (2004) 1–23
3. Breese, J.S., Heckerman, D., Kadie, C.M.: Empirical analysis of predictive algorithms for collaborative filtering. In Cooper, G.F., Moral, S., eds.: Proceedings of the 14^{th} Conference on Uncertainty in Artificial Intelligence. (1998) 43–52
4. Champaign, J., Cohen, R.: A model for content sequencing in intelligent tutoring systems based on the ecological approach and its validation through simulated students. In: Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference, Daytona Beach, Florida (2010) 486–491
5. Champaign, J., Cohen, R.: Exploring the effects of errors in assessment and time requirements of learning objects in a peer-based intelligent tutoring system. In: 2011 FLAIRS Conference Proceedings, Palm Beach, Florida (2011)
6. Plant, D.: hSITE: healthcare support through information technology enhancements NSERC Strategic Research Network Proposal. (2008)
7. Jackson, G.T., Graesser, A.C.: Content matters: An investigation of feedback categories within an ITS. In: Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work, Amsterdam, The Netherlands, The Netherlands, IOS Press (2007) 127–134
8. Almeida, R.B., Mozafari, B., Cho, J.: On the Evolution of Wikipedia, International Conference on Weblogs and Social Media (2007)
9. Cheng, R., Vassileva, J.: Design and evaluation of an adaptive incentive mechanism for sustained educational online communities. User Model. User-Adapt. Interact. **16**(3-4) (2006) 321–348
10. Colineau, N., Paris, C.: A portal to promote healthy living within families. In: eHealth. (2010) 259–266
11. Colineau, N., Paris, C.: Motivating reflection about health within the family: the use of goal setting and tailored feedback. User Model. User-Adapt. Interact. **21**(4-5) (2011) 341–376
12. Camerini, L., Giacobazzi, M., Boneschi, M., Schulz, P.J., Rubinelli, S.: Design and implementation of a web-based tailored gymnasium to enhance self-management of fibromyalgia. User Model. User-Adapt. Interact. **21**(4-5) (2011) 485–511
13. Harrer, A., McLaren, B.M., Walker, E., Bollen, L., Sewall, J.: Creating cognitive tutors for collaborative learning: steps toward realization. User Modeling and User-Adapted Interaction **16** (2006) 175–209

Towards a Ranking of Likely Diseases in Terms of Precision and Recall

Heiner Oberkampf[1], Sonja Zillner[1], Bernhard Bauer[2], and Matthias Hammon[3]

[1]Corporate Technology Siemens AG, Munich, Germany
[2]Software Methodologies for Distributed Systems, University of Augsburg, Germany
[3]Department of Radiology, University Hospital Erlangen, Germany
{heiner.oberkampf.ext,sonja.zillner}@siemens.com

Abstract. Addressing critics to existing clinical diagnosis systems we propose a weighted information-retrieval approach to provide a context-dependent ranking of likely diseases through matching the patient's symptom information to typical disease symptomatology. The matching is based on a formal model of symptoms, diseases and their relations. Knowledge resources incorporated in our model range from common clinical books and the clinician's individual expertise over the patient's context to established medical ontologies.

Keywords: clinical diagnosis support, ranking diseases, information retrieval, medical ontologies

1 Introduction

Diagnosis systems have a long tradition in computer science and artificial intelligence, especially in the medical domain. There has been done a lot in both theoretical foundation and practical application of diagnosis systems. However, there is still not a full acceptance from clinician's side and in practice these systems are rarely used. This is due to several reasons: many diagnosis systems are constructed to deliver a "complete diagnosis" explaining all observations and symptoms[1]. This leads to diagnosis with big disease-sets like "Lymphoma *and* Colorectal Cancer *and* some Infection *and* Diverticulitis *and* ... can explain the symptoms that were observed at the patient". However, clinicians are not interested in such an unspecific (multi-fault) diagnosis. Instead, they are more interested in a single-fault diagnosis even though a single disease cannot "explain" all symptoms. Further, many systems require the physician to enter symptoms manually. This is time-consuming and doesn't allow to include the full patient's symptom information. Additionally, most of the systems do not allow the clinician to bring in his individual expertise. However, correct weighting of symptoms, i.e. determining the relevance of symptoms in a particular context, largely depends on the clinician's intuition. Weighting symptoms, beyond registering the intensity, is crucial for a good diagnosis system. Another problem

[1] We exchangeably use the terms symptom, sign, finding and observation.

of existing diagnostic systems is that their models require knowledge, which is not broadly available or does not meet the nature of medical knowledge. On the one hand we do not have strict causal relations in medicine, as required in most of the logical approaches. On the other hand there is not enough probabilistic knowledge such that one could apply pure probabilistic approaches like Bayes-nets. We argue that a context-dependent ranking of disease-information is more appropriate than a set of diseases explaining (together) all symptoms. A weighted information-retrieval approach can help clinicians in identifying a diagnosis. We even think that weighting the patient's symptoms and then figuring out the "best match" under the set of diseases is very similar to the decision-making process of clinicians. So we decided to base the ranking of likely diseases on an adapted measure of precision and recall through which we determine how well the patient's symptom information and the typical disease symptomatology match. The factors influencing the ranking were identified in interviews with clinicians. The knowledge on which the ranking is based comes from three different perspectives: a) static medical knowledge that is commonly accessible like found in e.g [1] (disease-symptom relations etc.), b) patient-specific data like symptoms, age, gender, patients-history etc. and c) dynamic judgements of the clinician interacting with the diagnose system.

Medical ontologies are not directly necessary for the ranking of diseases itself, however they are a key-technology for semantically integrated acess of data. For example in the Theseus MEDICO project[2] medical ontologies like RadLex[3], FMA[4] or SNOMED CT[5] are used to annotate unstructured data as images and reports to make them better accessible (see e.g. [2]). Similar in [3] the usefulness of annotating clinical data for enhanced information retrieval is pointed out. In [4] we showed how to use these annotations to collect symptom information from clinical data automatically, such that clinicians do not have to search for symptoms in images and reports and enter them manually. This is done with the help of a Disease-Symptom-Ontology (DiSy) which is linked to established ontologies used for annotations. Automatic collection of symptom information makes also the understanding of symptoms in a temporal context possible. Further, ontologies allow to infer implicit symptom information. In section 2 we give an overview of related work, in section 3 we define the factors influencing the ranking of likely diseases and in section 4 we describe the ranking algorithm and then we show first evaluation results in section 5.

2 Related Work

Diagnosis Systems have a long tradition in Computer Science and Artificial Intelligence, in-particular within the medical domain. There are a variety of formalisms and techniques like set-cover, abductive reasoning, logic approaches,

[2] http://theseus-programm.de/en/920.php
[3] http://www.rsna.org/RadLex.aspx
[4] http://sig.biostr.washington.edu/projects/fm/
[5] http://www.ihtsdo.org/snomed-ct/

Baysian networks, rule-based systems, case-based reasoning etc. Most of the logical approaches aim to explain the whole set of observations, i.e. provide a complete diagnosis. This however often leads to diagnosis consisting of large fault sets and thus unspecific and redundant diagnosis, not helping the clinician. Often diagnosis-algorithms optimize the set of faults contained in a diagnosis under consideration of so called parsimony criteria. In set-cover those are e.g. minimal cardinality, irredundancy, relevance, most probable diagnosis and minimal cost [5]. Early attempts to formalization of model-based diagnostic knowledge were made in [6]. In [7] evidence-functions are used to encode the relation between defects and findings. Even though these formalisms are more expressive than our model, this comes with high computational cost: most diagnosis-algorithms are exponential with the number of possible faults. In the medical domain there are many well known implementations of clinical diagnosis systems (expert systems) from around the 1970th and later like e.g. MYCIN [8], INTERNIST-1 [9], CASNET [10], DXplain [11], CADIAG2[12], PATHFINDER [13]. Since Bayes-nets are theoretically best suited for diagnosis they are successfully implemented in some of the mentioned expert-systems. However statistical approaches like Bayes-nets require knowledge, which isn't broadly available: e.g. the a-priori probabilities for signs and symptoms $P(s)$ are mostly not known. Thus we are not able to compute the conditional probability for a disease, given the symptom $P(d|s) = \frac{P(s|d) \cdot P(d)}{P(s)}$ with the help of Bayes-Theorem. Note that $P(s|\neg d)$ is normally not known as well. It is not difficult to see that even though we cannot calculate $P(d_1|s)$ and $P(d_2|s)$ for two diseases d_1 and d_2, we are able to *compare* them by computing $P(d_1|s)/P(d_2|s)$. Thus, without knowing $P(s)$, *ranking* diseases is possible.

Similar to our approach is the work described in [14], where a quality measure for diagnosis is defined. However the factors influencing the ranking are different (e.g. they assume to have knowledge about the probability of a finding being caused by a certain disease). Further they do not allow the clinicians to interact with the system and change e.g. symptom weights. In [15] a model similar to ours represents the disease-symptom relations using fuzzy labels, however the symptoms itself are not weighted. They also chose an IR approach to rank likely diseases.

3 Factors with Influence on the Ranking

In interviews with clinicians we learned that a variety of factors are included in the diagnostic reasoning process. These factors, influencing the likeliness of diseases, can be divided into the following three groups: 1) disease-specific, 2) disease-symptom-relation and 3) symptom-specific. Further we learned that we need to integrate several perspectives: common clinical knowledge, the patient's context as well as the clinician's expertise and intuition. Especially the integration of the clinician's individual expertise is needed in order to enhance user acceptance and enable a dynamic diagnosis work-flow. Next we provide some formal definitions about the basic concepts used in the following.

Definition 1. *We denote the set of diseases by* $D = \{d_i\}_{i=1,\ldots,m}$ *and the set of symptoms by* $S = \{s_j\}_{j=1,\ldots,n}$*. In the context of some patient we have a disjunction of* S *into three sets: present symptoms* $S_{present}$*, open symptoms* S_{open} *(symptoms, which have not been investigated yet) and absent symptoms* S_{absent}

$$S = S_{present} \dot{\cup} S_{open} \dot{\cup} S_{absent}$$

For some disease $d \in D$ *the set of related symptoms is denoted by* $S(d) \subseteq S$*.*

With respect to the ranking of diseases all present symptoms which are in $S(d)$ *support* d, whereas absent symptoms of $S(d)$ *degrade* d. In figure 1 the related symptoms of d are $S(d) = \{s_1, s_2, s_3, s_4\}$, the symptoms $S(d) \cap S_{present} = \{s_1, s_3\}$ *support* disease d, whereas symptoms $S(d_1) \cap S_{absent} = \{s_2\}$ *degrade* disease d. Since absent symptoms allow clinicians to exclude certain diseases, they are important in the diagnosis process: a patient having a normal amount of leukocytes, erythrocytes and thrombocytes is *very unlikely* to have lymphatic or myeloic leukemia. The higher the number of present symptoms and the lower the number of absent symptoms, which are related to a given disease the higher this disease should be ranked. The amount of open symptoms of $S(d)$ express how certain our judgement of d is.

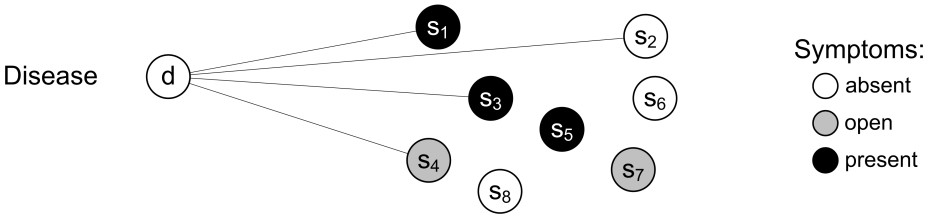


Fig. 1. Matching the typical symptomatology of a disease with the patient's symptoms.

3.1 Disease Specific Factors

The most distinctive diseases-specific factor is the incidence proportion of a disease, which is "the number of new cases within a specified time period divided by the size of the population initially at risk"[6]. With a normalized time period the incidence proportion can be used as the basic probability of a random person to get the disease. For many diseases even age- and sex-specific incidence proportions are available e.g. at Cancer Research UK[7].

Definition 2. *The incidence proportion of a disease* $d \in D$ *normalized to a time period of one year is denoted by* $P(d)$*. For a patient* p *the patient-specific incidence proportion of a disease* $d \in D$ *respecting at least the patient's age and sex, normalized to a time period of one year, is denoted by* $P_p(d)$*.*

[6] Wikipedia
[7] http://info.cancerresearchuk.org/cancerstats

Other patient-specific risk factors with influence on the incidence proportion like former diseases or the patient's life-style (such as smoking etc.), might be included in $P_p(d)$ if they are available.

3.2 Disease-Symptom-Relation

In order to infer likely diseases we need information about their relation to symptoms. In common medical text books, symptoms of diseases are well documented: so $S(d)$ is known. However the strength of the relation is less well and less precisely documented. One would like to have the conditional probability for disease d given symptom s, i.e. $P(d|s)$, but as described above this information is mostly not available. Instead of a precise significance value clinicians consider so called *leading symptoms* of a disease, which have a more significant relation to the respective disease than other symptoms. Further the conditional probability $P(s|d)$ also contributes to the strength of the disease-symptom relation.

Definition 3. *For all symptoms $s \in S$ and all diseases $d \in D$ we define a value $rel(s,d) \in [0,3]$ describing the relatedness of d and s. Two factors contribute to $rel(s,d)$: the conditional probability $P(s|d)$ and a factor for leading symptoms $l(s,d) \in \{1,3\}$ (where $l(s,d) = 3$ iff s is a leading symptom of d, otherwise $l(s,d) = 1$). The relatedness of s and d is then defined as*

$$rel(s,d) := P(s|d) \cdot l(s,d)$$

Setting the factor for leading symptoms to 3 can be understand only as a rule of thumb. To motivate this value, consider e.g. pulmonary embolism: a leading symptom is a suddenly appeared dyspnoea, however other symptoms like thoracic pain, raised D-dimer concentration and expectoration of blood, even though unspecific as single symptoms, in combination the three will make a clinician to think at pulmonary embolism. To give another example consider the three unspecific symptoms fever, night sweats and weight loss. The presence all three is referred to as B-symptomatic – a leading symptom with high prognostic significance for Hodgkin's- and non-Hodgkin's lymphoma.

3.3 Symptom Specific Factors

Weighting symptoms and thus defining their importance and relevance in the patient's context is essential in diagnosis. However this is difficult since the clinician's experience and intuition is involved in several ways. The clinician *sees* which of the different symptoms are the most important in a given situation. Next we define all factors contributing to importance of a certain symptom. Within interviews with clinicians we found that symptoms significantly differ in their default or basic importance and their need to be explained. The default importance factor represents the basic importance value of a particular symptom without consideration of any influencing context factors and is not related to probability. E.g. the symptom "blood in stool" is intrinsically more critical than "feeling powerless".

Definition 4. *For all symptoms $s \in S$ the default importance (default weight) is defined by $\omega_{default}(s) \in (0, 1]$.*

The default importance is defined for all symptoms $s \in S$. In a first attempt we simple formed symptom-groups of similar importance to ease the knowledge gathering. For present symptoms $s \in S_{present}$ we have some more factors like e.g. the intensity. Often the intensity of symptoms is described in general terms like low, medium, high but in the context of the symptom "enlarged lymph nodes" the levels could be "1, 2, many".

Definition 5. *Let $M(s)$ denote the ordered set of intensity descriptions for some symptom s. For all symptoms $s \in S_{present}$ and intensity descriptions $m \in M(s)$ we denote the intensity by $i(s, m) \in (0, 1]$, where $i(s, m) = 1$ if m is the maximal element of $M(s)$, i.e. if the symptom is present with full intensity.*

Clinicians make a distinction between "newly appeared" and "old" symptoms. In their decision-making new symptoms are significantly more important than old ones. Consider e.g. a patient having raised blood-pressure already over some years. If this patient gets to the doctor with acute fever and lymph node enlargements these symptoms are considered more important than the raised blood pressure. On the other hand with respect to pulmonary embolism then a deep vein thrombosis has to be considered, even if the patient might have it already for some time-period. This is why an integrated view on the patient symptoms is so important and considering only new symptoms is not appropriate.

Definition 6. *For all symptoms $s \in S_{present}$ we denote the novelty of s by $n(s) \in \{1, 3\}$, where $n(s) = 3$ if the symptom is new and $n(s) = 1$ otherwise.*

In order to apply the novelty-factor one has to define for each symptoms s time-threshold t up to which the symptom is classified as new. Additionally to those factors, we allow the clinician to influence the importance of some symptoms within a certain range.

Definition 7. *For a symptom $s \in S$ we define a clinicians factor $c(s) \in [0, 2]$ with default value $c(s) = 1$ if no influence from clinician's side was taken.*

Note that the clinician's factor can be used in two ways: for reducing and enhancing the importance. In setting $c(s) = 0$ a present symptom can be disregarded in the ranking. Consider a patient with chronic thoracic pain. A newly appeared cough might be disregarded in the actual diagnosis setting $c(cough) = 0$. However another clinician, suspecting pulmonary embolism, could enhance the importance of the cough setting $c(cough) = 2$. We introduced the clinician's factor in order to give the clinician the possibility to "play" with the system and try different scenarios. The overall importance of a symptom $s \in S$ is defined as:

Definition 8. *The importance (weight) of a symptom $s \in S$ is denoted by $\omega(s)$. For $s \in S_{present}$ we define $\omega(s)$ as*

$$\omega(s) := \omega_{default}(s) \cdot (1 + i(s, m)) \cdot n(s) \cdot c(s) \quad , \forall s \in S_{present}, m \in M(s).$$

Intensity, novelty and clinician's factor are only applied for present symptoms, so for $s \in S \setminus S_{present}$ we define $\omega(s)$ simply as

$$\omega(s) := \omega_{default}(s) \cdot c(s) \quad , \forall s \in S \setminus S_{present}.$$

Note that since $(1 + i(s, m)) \cdot n(s) > 1$ we always have $\omega(s) > \omega_{default}(s)$ for $s \in S_{present}$, so the symptom's importance is always higher, when the symptom is present (even though not new and with low intensity) than when the same symptom is absent. In summary we get three factors the patient-specific incidence of a disease $P_p(d)$, a relatedness value for the disease-symptom relation $rel(s, d)$ and the symptom's weight $\omega(s)$ describing the importance of the symptom. Having those values we obtain a bipartite graph with weighted edges and vertices, where one partition represents the diseases and the other the symptoms (see figure 2). We will refer to this graph as the disease-symptom graph. The symptoms are marked as present, open or absent in dependence of the patient's situation.

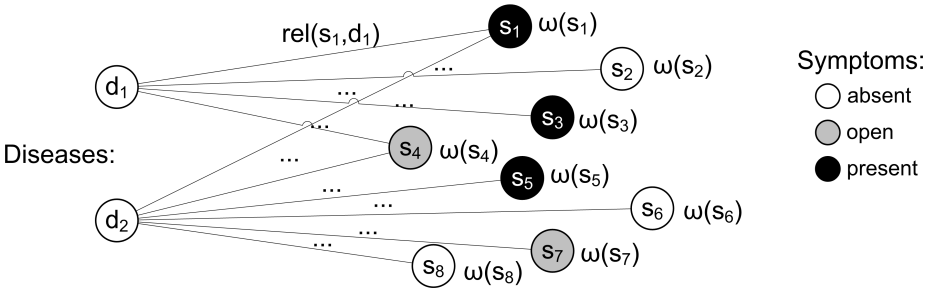


Fig. 2. The disease-symptom graph used for ranking likely diseases based on precision and recall. The weight $\omega(s_i)$ represents the importance of symptom s_i, the strength of the relation is represented by $rel(s_i, d_j)$, exemplary shown for $rel(s_1, d_1)$.

As explained above, present symptoms are *supporting* whereas absent symptoms are *degrading* the related diseases. The stronger the disease-symptom-relation is and the more important a symptom is, the bigger is this effect for the respective disease. This is realized in terms of precision and recall as described in the next section.

4 The Ranking

Based on the disease-symptom-graph (figure 2) we want to generate a ranking of likely diseases, for which we use information about present and absent symptoms. In subsection 4.1 we measure how the patient's symptoms match the symptomatology of a given disease – the basis for the ranking. Additionally in subsection 4.2, we compute and uncertainty factor for each disease, which

measures the relative amount of open symptoms. Further in subsection 4.3 we discuss how to integrate the patient-specific incidence of diseases.

4.1 Matching patient's symptoms to disease symptomatology

In terms of precision and recall the *relevant elements* (symptoms) we would like to meet with a disease, are the present symptoms of the patient $S_{present}$, in figure 2 these are $\{s_1, s_3, s_5\}$. If we consider some disease $d \in D$, then the symptoms that are *retrieved* by this disease are all symptom of d with known status, i.e. $S(d) \cap (S_{present} \cup S_{absent})$ e.g. in figure 2 the symptoms $\{s_1, s_5, s_6, s_8\}$ are *retrieved* by d_2. True positive are then $S(d) \cap S_{present}$ and false positive are $S(d) \cap S_{absent}$. Consequently precision and recall of a disease $d \in D$ with $S(d) \cap S_{present} \neq \emptyset$ are calculated as follows:

$$Precision(d) := \frac{\sum\limits_{s \in S(d) \cap S_{present}} rel(s,d) \cdot w(s)}{\sum\limits_{s \in S(d) \cap (S_{present} \cup S_{absent})} rel(s,d) \cdot w(s)}$$

So $Precision(d)$ is the percentage of potentially relevant disease symptoms that are actually present in the current patient. Since in the formulae for recall (below) we have a sum over all present symptoms $S_{present}$, which might contain also symptoms not in $S(d)$, the relatedness factor $rel(s,d)$ has to be omitted there:

$$Recall(d) := \frac{\sum\limits_{s \in S(d) \cap S_{present}} w(s)}{\sum\limits_{s \in S_{present}} w(s)}$$

So $Recall(d)$ measures the amount of symptoms of the disease d out of those which are present at the patient. A disease perfectly matching the patient's symptoms would have both, a recall and precision of 1. The classical F-measure combines precision and recall:

$$F(d) := 2 \cdot \frac{Precision(d) \cdot Recall(d)}{Precision(d) + Recall(d)}$$

So the $F(d)$ measures, how well the patient's symptoms match the symptomatology of a disease d.

4.2 Uncertainty

In the above calculations of precision and recall of a disease we omitted open symptoms S_{open}. The status of those symptoms is not known simply because they were not inspected yet. Thus the relative amount of open symptoms which are related to a disease can be used as an uncertainty measure for this disease:

$$Uncertainty(d) := \frac{\sum\limits_{s \in S(d) \cap S_{open}} rel(s,d) \cdot w(s)}{\sum\limits_{s \in S(d)} rel(s,d) \cdot w(s)}$$

Note that we use again the weighted relative amount, thus the uncertainty of d is smaller if important symptoms (as e.g. leading symptoms) are already known. The uncertainty factor is attached to the ranked disease to give the clinician an idea of how much the likeliness might change after new examinations.

4.3 Integration of Patient-specific Incidence of Diseases

There are two aspects making the combination of $F(d)$ and $P_p(d)$ difficult. Firstly $P_p(d) \ll F(d)$, i.e. the value of the incidence proportions $P_p(d)$ are very small in comparison with $F(d)$: for a 67 years old male patient p we have $P_p(\text{Hodgkin's lymphoma}) = 4/100000$, but $F(d)$ of top-ranked diseases is typically bigger than $1/2$. Secondly the distribution of the incidence proportions have a much higher variance: for the same patient we have $P_p(\text{colorectal cancer}) = 257/100000$, that is colorectal cancer is about 64-times more often. In experiments however we recognized that the F-measure varies with a factor of at most 10 regarding the top-ranked diseases. Thus simple multiplication of $P_p(d)$ with $F(d)$ we would make $P_p(d)$ predominant. That is why we decided to calculate the overall ranking factor as follows:

$$r(d) := F(d) + \lambda \cdot P_p(d), \quad \lambda \in \mathbb{R}, \lambda > 0$$

For clinicians the matching of symptoms i.e. $F(d)$ is more important than $P_p(d)$ so in a first attempt we chose λ as

$$\lambda = \frac{\frac{1}{|D|} \sum\limits_{d \in D} F(d)}{\max\limits_{d \in D} P_p(d)}$$

in order to limit the influence of $P_p(d)$ up to the average of $F(d)$. However the right value for λ has to be determined in a broad evaluation. Another possibility would be to use only $F(d)$ for the ranking and give the value $P_p(d)$ additionally to the user and highlight (red flag) the most probable diseases under the top-ranked. This is an ongoing discussion.

5 Evaluation

The selection and weighting of the various ranking factors was accomplished in collaboration with our clinical partner. In addition we revealed a first evaluation based on a small dataset encompassing five diseases (Hodgkin's lymphoma, non-Hodgkin's lymphoma, reactive lymphadenitis, colorectal cancer, diverticulitis), about 40 symptoms and ten sample patients. The sample data is captured within an extended version of the Disease-Symptom Ontology described in [4]. With the Jena framework we preformed reasoning and used SPARQL queries to extract the values needed to compute the ranking. In the evaluation we compared for each patient the disease diagnosed by a clinician with the top-ranked disease of our ranking algorithm. For each sample patient the computed top-ranked disease was equal to the user-expert based diagnosed disease.

6 Conclusion and Next Steps

We proposed a ranking algorithm for likely diseases in terms of precision and recall by integrating various clinical knowledge resources. Currently we are planning a bigger quantitative clinical evaluation and fine-tuning of the ranking factors. In this context we will enlarge the set of patients and extend the Disease-Symptom-Ontology to cover more diseases and symptoms. In order to extend the Disease-Symptom-Ontology we plan to incorporate knowledge from online resources such as existing medical ontologies.

References

1. Herold, G.: Herold: Internal Medicine. Gerd Herold, Köln (2011)
2. Seifert, S., Kelm, M., Moeller, M., Mukherjee, S., Cavallaro, A.: Semantic Annotation of Medical Images. Proceedings of SPIE Medical Imaging, February 13-18, San Diego, CA, United States (2010)
3. Opitz, J., Parsia, B., Sattler, U.: Evaluating Modelling Approaches for Medical Image Annotations. CoRR (2010)
4. Oberkampf, H., Zillner, S., Bauer, B., Hammon, M.: Interpreting Patient Data using Medical Background Knowledge. To be published in ICBO proceedings (2012)
5. Reggia, J., Nau, D., Wang, P.: Diagnostic expert systems based on a set covering model. International Journal of ManMachine Studies **19**(5) (1983) 437–460
6. Reiter, R.: A Theory of Diagnosis from First Principles. Artificial Intelligence **32**(1987) (1987) 57–95
7. Lucas, P.: Analysis of notions of diagnosis. Artificial Intelligence **105(1-2)** (1998) 293–341
8. Shortliffe, E.H.: Computer-Based Medical Consultations: MYCIN. Volume 85. American Elsevier (1976)
9. Miller, R.A., Pople, H.E., Myers, J.D.: Internist-1: An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine. New England Journal of Medicine **307**(8) (1982) 468–476
10. Kulikowski, C.A., Weiss, S.M.: Representation of expert knowledge for consultation: The CASNET and EXPERT projects. In Szolovits, P., ed.: Artificial Intelligence in Medicine. Westview Press (1982) 21–55
11. Barnett, G.O., Cimino, J.J., Hupp, J.A., Hoffer, E.P.: DXplain. An evolving diagnostic decision-support system. Jama The Journal Of The American Medical Association **258**(1) (1987) 67–74
12. Adlassng, K.P., Akhavan-Heidari, M.: Cadiag-2/gall: An experimental expert system for the diagnosis of gallbladder and biliary tract diseases. Artificial Intelligence in Medicine **1**(2) (1989) 71 – 77
13. Heckerman, D.E., Horvitz, E.J., Nathwani, B.N.: Update on the Pathfinder Project. In: Proc Annu Symp Comput Appl Med Care. (1989) 203–207
14. Baumeister, J., Seipel, D., Puppe, F.: Incremental Development od Diagnostic Set-Covering Models with Therapy Effects. International Journal of Uncertainty, Fuzzyness and Knowledge-Based Systems **11**(Nov) (2003) 25–49
15. Maio, C.D., Loia, V., Informatica, D., Fenza, G., Linciano, R., Morrone, A.: Fuzzy Knowledge Approach to Automatic Disease Diagnosis. IEEE International Conference On Fuzzy Systems (2011) 2088–2095

Supporting tele-health and AI-based clinical decision making with sensor data fusion and semantic interpretation: The USEFIL case study

Alexander Artikis[1], Panagiotis D. Bamidis[2], Antonis Billis[2], Charalampos Bratsas[2], Christos Frantzidis[2], Vangelis Karkaletsis[1], Manousos Klados[2], Evdokimos Konstantinidis[2], Stasinos Konstantopoulos[1], Dimitris Kosmopoulos[1,3], Homer Papadopoulos[1], Stavros Perantonis[1], Sergios Petridis[1], Constantine S. Spyropoulos[1]

[1] Institute of Informatics and Telecommunications, NCSR 'Demokritos', Ag. Paraskevi 153 10, Athens, Greece
costass@iit.demokritos.gr
[2] Lab of Medical Informatics, Medical School, Aristotle University of Thessaloniki, Greece
bamidis@med.auth.gr
[3] Division of Computer and Information Sciences Rutgers University, New Jersey, U.S.A.

Abstract We propose a three-layered architecture for clinical evidence-based Decision Support Systems. Our architecture allows for off-the-shelf low-cost sensors to be deployed in tele-health environments; counter-balancing low confidence in the sensor data by fusing data from multiple sensors. The relevant data fusion and interpretation layer also forms the interface between sensor data and explicit rules encoding medical knowledge. This achieves a complete separation of the non-medical and medical knowledge, an important step for system adoption.

1 Motivation

Rule-based *Decision Support Systems (DSS)* are routinely used in AI-based clinical decision making, often integrated into larger medical care and tele-health support systems, as well as non-clinical research and laboratory situations.

Although academic research has introduced *machine learning* as an alternative to manually authored rule systems, adoption is marginal at best, mainly due to the paramount importance placed on certified medical personnel's ability to inspect and edit the rules in a human-processable form. In this paper, we present the architecture proposed by the USEFIL project, where clinical evidence-based decision making and low-level data processing are explicitly separated, with a logic-based layer used as the interface between the data processing components that produce *quantitative* clinical data and the rule-based system that operates upon a more *qualitative* representation of clinical evidence (Figure 1).

The advantage is twofold: using machine learning we can exploit the extensive research on fusing measurements from multiple off-the-shelf, low cost sensors

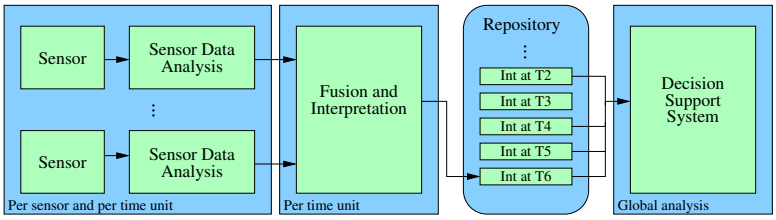


Figure 1. The USEFIL interface between sensor data and DSS.

into a high-confidence representation of the current situation. At the same time, medical knowledge is encoded in human-processable form and the system retains the transparent and predictable decision data flow; that facilitates adoption.

Furthermore, USEFIL addresses the growing application domain of independent living for elderly people. Use of technology in care delivery is still limited, a fact partially attributed to technological shortcomings and partially to user acceptance (29% and 37% of the cases, resp.) [1]. Further evidence shows that identifying and meeting user needs is one of the main difficulties in developing smart home systems [2]. The USEFIL architecture design is driven by the requirement that technology helps provide a comfortable and independent lifestyle. USEFIL carried out requirements collection involving interviews and focus groups for approaching these difficulties, concluding that technology should be *unobtrusive*, avoiding physical contact between the sensors and the user, and also being able to operate on fragmented and unreliable monitoring.

2 Unobtrusive sensor network

The unobtrusive sensor network collects and analyses with intelligent methods data related to physiology, activities of daily life (ADL), emotional status using visual, auditive and other sensors (e.g., activity meter embedded in a wrist watch), exercising the maximum discretion possible. Besides unobtrusiveness, the second ambition is the development of specialized sensors using off-the-shelf low-cost hardware (e.g., webcams as physiometric sensors). Sensor data will be collected via the local wireless network, real-time analysed using machine learning techniques on-site; analysis results (events and measurements) will be securely forwarded to the fusion module and all data (raw and extracted) will be immediately discarded. This avoids the storage and security issues of retaining high-volume *sensitive* information, such as video footage. We believe that this approach adequately addresses concerns with respect to participants privacy.

The video monitoring unit will integrate novel algorithms for health including measuring vital functions such as heart rate, monitoring [3], body temperature, blood oxygen saturation and other important determinants in the puzzle of preventive medicine. One of our prime objectives and major technical chalenges is to explore the extent to which vital signs can be measured without wiring and

without using specialized sensing hardware, developing an unobtrusive and low-cost monitoring system. This is important for accessing a larger user group and for elderly people and chronic disease sufferers who need to constantly monitor their vital signs; as well as for situations, such as involving depressed or elderly patients, where users refuse or forget or are unable to use monitoring systems.

Besides measuring vital functions, audio and visual signals will be analysed for cues regarding emotional state [4] – another major technical chalenge of the USEFIL project. Starting from visual content analysis of spontaneous and unconstrained facial expression and human activity recognition, where current techniques fail in spontaneous behaviour recognition in less controlled environments [5]. Robust audio analysis [6] will be used to study the dynamics of speech. Short-term behaviour specific cues (e.g. body gestures) shall further refine the classification task by modelling interaction aspects that could influence the users' emotional state. We will also tackle the challenging problem of detecting human affective behaviour cues in less constrained settings.

3 Data fusion and interpretation

Data fusion puts together a coherent 'snapshot' of the user's status from the different sensor data analysis results. Such snapshots aggregate data over periods of time short enough to be considered as a unit by the DSS and long enough to not swamp the DSS with unnecessary detail. Furthermore, fusion cross-validates sensor input from different modalities to detect hardware failures or other abnormal system conditions; similarly to the way that multimodal document understanding uses multi-modal information to fill gaps and detect errors [7].

Data fusion and interpretation will be based on symbolic event recognition techniques. More precisely, the Event Calculus [8] will be used in order to recognise composite events of interest given the sensor data, and make them available to the DSS for further, longer-term reasoning. The Event Calculus is a logic programming language for representing and reasoning about events and their effects. It allows for complex temporal representation and has recently been extended to support real-time reasoning in large-scale distributed systems [9]. Furthermore, it has direct routes to reasoning under uncertainty [10] and, therefore, it is suitable for noisy environments such as sensor networks.

This approach provides a clean interface between non-medical knowledge pertaining to the interpretation of physical measurements and medical knowledge pertaining to decisions based on these measurements. Furthermore, this fusion and interpretation layer is a natural position for implementing *personalization* and *adaptivity*, so that deviations from normal values are understood in the context of the different individual users and circumstances.

4 Decision support

DSS combines information extracted from sensors with medical history information, to produce indicators pertaining to mental and physical status and to

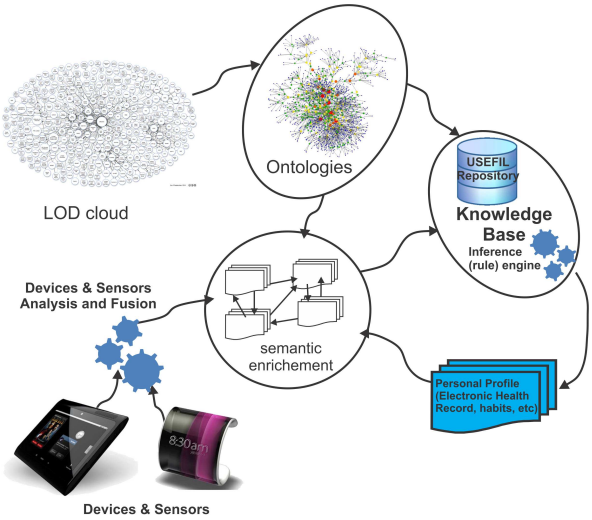


Figure 2. USEFIL DSS architecture.

suggest interventions or actions (Figure 2). As such, DSS will be the main communication layer between the USEFIL system and the health professionals and informal caregivers; including methods to sustain the knowledge acquisition process from elders and health professionals.

DSS will be based on *semantic rules* and a rule-based inference mechanism: expert knowledge and clinical guidelines can be naturally expressed in a rules language such as SWRL, and are some of the most important source of information for likely diagnosis and therapy prescriptions/recommendations.

A commonly recurring issue in clinical decision support systems is that of *uncertainty*, introduced by limited sensor accuracy or even sensor unavailability. In the presence of uncertainty, probabilistic techniques enhance system flexibility and robustness. For instance, the Naïve Bayesian approach combined with Fuzzy Cognitive Maps [11] may offer a more elaborated way of highlighting interactions among clinical features that may in turn facilitate the expert's decision.

Besides reaching real-time decisions regarding current health status, long-term monitoring is also a core element of the proposed architecture. *Trend analysis* encompassing the notion of *temporal logic* can provide baseline alterations of the participant's long-term health status Adopting this approach, the feasibility of preventing future risky health situations using the knowledge derived from past alerting instances will be investigated. By expressing and responding to user needs that are particular to each individual user and also evolve and change over time, DSS is another natural position (besides personalized data interpretation discussed in the previous section) for *personalization* and *adaptivity* to be implemented.

5 Conclusions

This paper presents our position on clinical evidence-based DSS, that adoption and uptake of machine learning and statistical analysis technologies will be enhanced by clearly separating data processing functionality from medical knowledge and restricting the application of such methods to the former.

Besides motivating our position, we propose an architecture that supports it and present state-of-the-art technologies that can implement it. Our research objectives are (a) to explore implementing specialized sensors by fusing data from low-cost general-purpose sensing hardware; and (b) to define the data fusion-DSS interface and the position of personalization; that is to say, to explore personalized interpretation vs. personalized decision making.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 288532. For more details, please see `http://www.usefil.eu`

References

1. Broens, T.H.F., Huis in 't Veld, R.M., *et al.*: Determinants of successful telemedicine implementations: A literature study. J Telemed. and Telecare **13** (2007)
2. Demiris, G., Oliver, D.P., *et al.*: Findings from a participatory evaluation of a smart home application for older adults. Technology and Health Care **16** (2008)
3. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans. Biomed. Engineering **58**(1) (2011) 7–11
4. Calvo, R., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Trans. Affective Comp. **1**(1) (2010) 18–37
5. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: Proc. Workshop for Human Communicative Behaviour Analysis, held at Computer Vision and Pattern Recognition (CVPR 2010), San Francisco, USA (2010) 9–14
6. Giannakopoulos, T., Petridis, S.: Fisher linear semi-discriminant analysis for speaker diarization. IEEE Trans. Audio, Speech, Language Processing **99** (2011)
7. Paliouras, G., Spyropoulos, C.D., *et al.*, eds.: Knowledge-Driven Multimedia Information Extraction and Ontology Evolution. LNCS 6050. Springer (2011)
8. Kowalski, R.A., Sergot, M.J.: A logic-based calculus of events. New Generation Comput. **4**(1) (1986) 67–95
9. Artikis, A., Sergot, M.J., Paliouras, G.: Run-time composite event recognition. In: CM Proc. ACM Intl Conf. on Distributed Event-Based Systems (DEBS 2012),
10. Filippou, J., Artikis, A., Skarlatidis, A., Paliouras, G.: A probabilistic logic programming event calculus. CoRR **abs/1204.1851** (2012)
11. Papageorgiou, E.I.: A fuzzy inference map approach to cope with uncertainty in modeling medical knowledge and making decisions. Intelligent Decision Technologies **5**(3) (2011) 219–235

A Framework for AI-Based Clinical Decision Support that is Patient-Centric and Evidence-Based

John A. Doucette, Atif Khan, Robin Cohen, Dan Lizotte, and Hooman M. Moghaddam

David R. Cheriton School of Computer Science
University of Waterloo, Ontario, Canada

Abstract. In this paper, we present a framework which enables medical decision making in the presence of partial information. At the core is ontology-based automated reasoning; this is augmented with machine learning techniques to enhance existing patient datasets. Our approach supports interoperability between different health information systems. This is clarified in a sample implementation that combines three separate datasets (patient data, drug drug interactions and drug prescription rules) to demonstrate the effectiveness of our algorithms in producing effective medical decisions. In short, we demonstrate the potential for artificial intelligence to support a task where there is a critical need from medical professionals, coping with missing or noisy patient data and enabling the usage of multiple medical datasets.

1 Introduction

Medical decision support systems (MDSS) map patient information to promising diagnostic and treatment paths. The value of such systems has been shown in various healthcare settings [1–3]. The properties of data, including representation, heterogeneity, availability and interoperability play a critical role in ensuring the success of MDSS. A decision making process should use all relevant data from many distributed systems instead of a single data source to maximize its effectiveness [4], but real-world medical decisions are often based on incomplete information due to the challenges posed by these properties when engaging in data synthesis.

Many artificial intelligence (AI) techniques (including knowledge-based and learning-based techniques) have been employed to deal with this information challenge, and to create a robust, practical MDSS - most notably MYCIN [5], Internist & Cadence [6], DXplain [7] and HIROFILOS-II [8]. Although prior approaches have enjoyed partial success, neither alone has been completely successful in real-world medical settings. Knowledge-based systems can suffer a significant loss of performance when patient data is incomplete (e.g. patients omit details, or access restrictions prevent viewing of remote medical records). In contrast the decisions of learning-based systems cannot be easily explained,

and may have difficulty differentiating correlation from causation when making recommendations [9].

Although both approaches have enjoyed partial success, neither alone has been completely successful in real-world medical settings. Knowledge-based systems can suffer a significant loss of performance when patient data is incomplete (e.g. patients omit details, or access restrictions prevent viewing of remote medical records), while the output of learning-based systems cannot be easily explained, may have difficulty differentiating correlation from causation when making recommendations, and can produce models which are opaque to laypeople [9].

Our system leverages the benefits of machine learning, structured knowledge representation, and logic-based inference in a novel fashion. We demonstrate on real world data that it is capable of providing robust, intelligent decision support, despite the complexity of medical relationships and the inter-dependencies inherent in medical decisions. Where previously machine learning (ML) in isolation has been demonstrated to fall short[9], our hybrid architecture produces decisions that are easy to verify and explain and, more critically, is also robust to missing data.

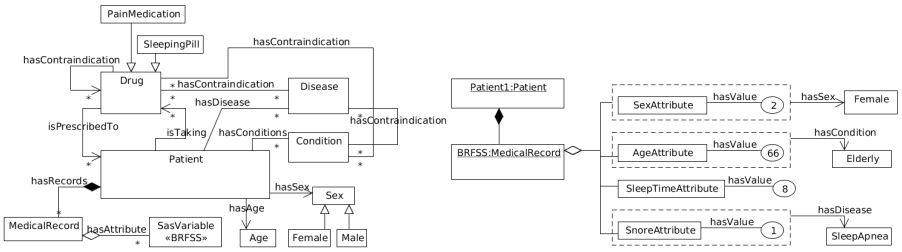
To realize our system, we represnted raw patient information using ontological concepts and placed it in structured triple-stores. Inference rules were designed by a domain expert, and applied using a semantic reasoner to generate decisions. This made decisions produced by the system easy to validate and explain, but the resulting knowledge-based system required complete information, which limited its usefulness in the real world. We overcame this limitation by augmenting a semantic reasoner with machine learning techniques to impute values for missing data. Imputation models are generated in a pre-processing stage and then integrated with the ontological system, allowing the system to perform in real time. This results in a patient-centric, evidence-based, decision support system.

Our proof-of-concept implementation employs three sources of information: a large, real-world dataset of patient medical information, a drug interaction registry, and a collection of medication prescription protocols. Preliminary results confirm that for practical medical scenarios, where patient data may be missing or incomplete, our hybrid design outperforms both solutions which rely exclusively on knowledge-based techniques and those which rely exclusively on machine learning.

2 Implementation & Evaluation

In order to validate our proposed framework, we created a proof of concept implementation focused around the knowledge management component and the query execution component from an existing ontological decision support system design [9]. We chose insomnia treatment as our line of inquiry, and used the following real-world datasets:

1. Patient records drawn from the Center for Disease Control (CDC) Behavioral Risk Factor Surveillance System (BRFSS) telephone survey for 2010 [10]. The BRFSS dataset contains a wide array of respondent information including age, race, sex, and geographic location, along with information about a wide range of common medical conditions like cancer, asthma, mental illness, and diabetes. Many behavioural risk factors including alcohol consumption, drug use, and sleep deprivation are also tracked. The dataset contains information on 450,000 individuals defining over 450 attributes for each individual. All of the data is numerically coded and stored in a structured format similar to a relational database.
2. A prescription protocol drawn from the Mayo Clinic [11] for use as expert decision making rules corresponding to the prescription protocol for various sleep aids.
3. A drug interaction registry [12] to identify drug-to-drug interactions.



(a) Ontology graph (b) BRFSS based patient record

Fig. 1: The figure depicts an ontological model representing the core concepts used in the and their relationships. It also depicts the use of inference rules to map raw data onto ontological concepts.

2.1 Knowledge Management Component

To instantiate the knowledge management component of the system design [9], we created a simplified ontology to define the relevant key concepts and their various relationships, shown in Figure 1a. We created inference rules in accordance with the BRFSS codebook which defined the semantics of different values for the data attributes, to transform the numerically coded BRFSS data records into corresponding instances of the 'Patient' concept. These rules were then applied to all records to create a semantic knowledge-store of the BRFSS dataset.

Figure 1b describes a particular patient instance and the corresponding medical information. The rules *hasValue* capture the raw BRFSS data, and the BRFSS codebook based inference rules enrich the knowledge base by linking the raw values to ontological concepts For example, a patient might have an

attribute *SEX* defined with a value of *2*. The *hasSex* rule maps all patients with a value of *2* for the *SEX* field as instances of the *Female* concept. Other patient information (such as medical conditions, diseases etc.) was mapped in similar fashion.

2.2 Query Execution Component

To instantiate our query execution component, we combined a semantic reasoner called 'Euler Proof Mechanism: EulerSharp' [13] with the WEKA machine learning toolkit [14]. The semantic reasoner provided the main mechanism for logic-based decision making in the system, while WEKA acted in a supporting role to impute missing data.

We identified a subset of sleep aids and applied the Mayo clinic sleep aid prescription protocol[15] to identify the conditions under which each drug should be prescribed. Using the ontological concepts, this information was then translated into inference rules for the decision making process. A local family physician assisted in selecting the various drugs and validating our translation of the Mayo clinic sleep aid prescription protocol[16]. Although the inference rules have been kept simple, they do reflect real medical considerations for sleep aid prescription. The generic forms of the resulting rules are given below, but specific interactions were also verified with the physician.

1. **drug-to-drug interaction rule:** *If a patient is currently taking an existing drug D_1, and D_1 cannot be given with drug D_2, then the patient cannot be given drug D_2.*
2. **drug-to-condition interaction rule:** *If a patient has some existing medical condition C, and a drug D has contraindication to the condition C, then the patient cannot be given drug D.*
3. **drug-to-disease interaction rule:** *If a patient has a disease E, and a drug D has contraindication to the disease E, then the patient cannot be given drug D.*

In order to deal with the missing values in patient records, we created classifiers using machine learning to predict values for the missing data fields. We trained a classifier to predict each attribute using all complete data from the BRSS dataset for the attributes of interest as a training set. For example, suppose the sleep aid estazolam cannot be prescribed to elderly patients, making it important to know a patient's age. We first take all patient records in the BRFSS set where the patient's age is known, and partition this subset into training and validation data. A classifier is then built using the training data. The performance of this classifier on validation data can provide a point estimate of the classifier confidence when making a decision, though we note that more elaborate estimations of confidence are a possible avenue for future work. In future cases where the patient's age is missing, we apply our classifier to label the patient's age as either elderly or not. The predicted value is substituted into the patient's record, and the semantic reasoner is run again. The confidence of the decision made by the semantic reasoner is based on the point estimate mentioned above.

2.3 Experimental Comparison

We conducted several experiments to evaluate the effectiveness of the proposed hybrid decision making system. Patients who should be given sleep aids, according to the Mayo clinic prescription protocol, were labelled as 'positive' exemplars, and those who should not as 'negative' exemplars. When a system labeled a patient correctly in response to a query, a 'true positive' (tp) or 'true negative' (tn) was produced. Otherwise, a 'false positive' (fp) or 'false negative' (fn) was produced. The results were evaluated in terms of:

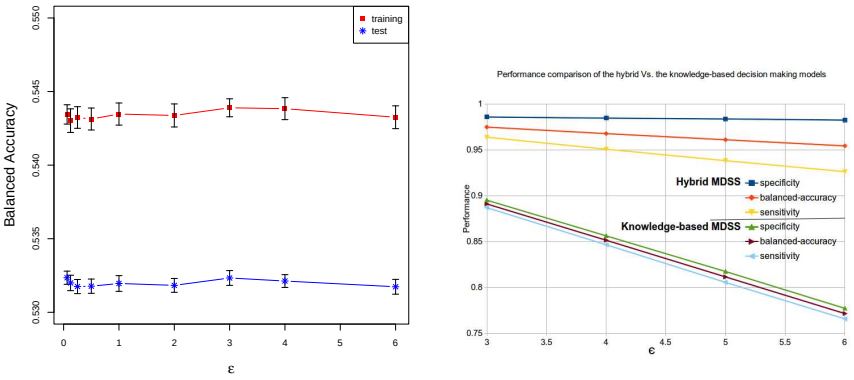
1. **Sensitivity**: *rate of positive exemplars labeled as positive.*
2. **Specificity**: *rate of negative exemplars labeled as negative.*
3. **Balanced accuracy**: simple average of *specificity* and *sensitivity* [17].

$$Spec = \frac{tn}{tn + fp} \qquad Sens = \frac{tp}{tp + fn} \qquad balAcc = \frac{Spec + Sens}{2}$$

2.4 Learning-based system

In order to assess the usefulness of our hybrid system over a purely learning-based system, we began by evaluating the performance of four different machine learning algorithms (*decision stump*, *C4.5-R8*, *Bagging* and *AdaBoost*) using the BRFSS dataset as follows. We generated 50 different randomly selected training sets (of two sizes: 2,500 exemplars and 5,000 exemplars), and used an information gain based *feature selection* algorithm [18] to reduce each set to its 30 most informative attributes. For each algorithm, we trained a predictive model (classifier) for every sleep aid in question, to predict whether a patient can be given that sleep aid or not – essentially they were trained to produce the output of the knowledge-based system based on patient data. We established the ground truth for each data record using the semantic reasoner (since when a knowledge-based decision can be made, we can assume 100% accuracy [19, 9]), and then compared the predictive accuracy of each machine learning algorithm. AdaBoost had the best performance across all four algorithms. Therefore, we only compared our system to AdaBoost-derived classifiers. The overall accuracy of AdaBoost when predicting the correct medical decision is low (roughly 0.5 on a scale of 0 to 1, nearly equivalent to the performance of a degenerate classifier but still statistically significantly better.)

Despite the poor performance of the learning-based system, we suspected it would be tolerant to missing data. We evaluated the impact of missing information on the performance of our learning-based system by removing known values from the patient records. We defined ϵ as the average number of attribute values removed from a patient's record, and varied ϵ to from an average of $\frac{1}{16}$ removed values per record to an average of 6 removed values per record. For each value of ϵ, we trained an AdaBoost-based classifier using 50 sets of 5000 exemplars from the partially-missing data. We then analyzed the impact of ϵ (missingness) on the performance of the AdaBoost-based learning-based system.



(a) Machine learning technique (b) Hybrid construction

Fig. 2: a. Impact of data missingness on balanced accuracy of AdaBoost-based classifiers for training & test data.
b. Performance comparison of hybrid & knowledge-based models for noisy data

The results in Figure 2a show the mean values for *balAcc* across the range of ϵ values for the AdaBoost-based classifier. We found the distributions to be approximately normal and no statistically significant differences in performance across the different values of ϵ. Furthermore, the performance of the classifier was very similar across the training and the test data, suggesting that AdaBoost is not over-fitting and is quite resilient to data omissions in the BRFSS-based patient records.

2.5 Knowledge-based and Hybrid Systems

Finally, we compared the performance of our hybrid system with that of the purely learning-based system described above and a purely knowledge-based system that had no imputation capability. We used EulerSharp for the knowledge-based reasoning and an AdaBoost-based classifier for the machine learning recommendation component. We selected the four data-sets corresponding to the four highest values of ϵ. For each ϵ value we measured the degradation of the knowledge-based decision making process. We trained an AdaBoost-based classifier to predict each patient attribute impacted by ϵ. For each patient record, the missing data values were replaced by the predicted values generated by the machine learning models. The semantic reasoner then reevaluated the system decision.

Figure 2b provides a performance comparison between the hybrid model and the knowledge-based model for the four highest levels of missingness (ϵ). We note that the hybrid decision making model experiences slight performance degradation in *balanced accuracy* as ϵ increases (an increase of 0.5 in ϵ causes

a decrease in performance of less than 1 percentage point). However, the performance of the knowledge-based decision support model degrades substantially for the same range of ϵ (an increase of 0.5 in ϵ causes a decrease in performance of roughly 4 percentage points). Overall the hybrid model achieves excellent *balanced accuracy*, meaning that its recommendations for medical decision making are effective.

2.6 Standard Imputation Methods

Analysis of datasets with missing values is a very well-studied problem in statistics, where *multiple imputation* (MI) techniques are often used [20]. When performing multiple imputation, each missing value is imputed several times by drawing feature values from a predictive distribution. This results in a collection of imputed datasets. Each imputed dataset is the same shape as the original dataset, and all of its non-missing values are identical to those in the original dataset, but its missing values are "filled in" differently for each imputed version. The collection of imputed datasets can then be used to produce unbiased estimates of summary statistics like means and regression coefficients, as well as statistically valid confidence measures for these statistics. Note that this goal of producing an accurate *summary* of the dataset is different from our goal of accurately predicting the missing values of *individuals* in the dataset. Nonetheless, for completeness we investigated the use of multiple imputation in our system.

We examined one popular MI technique, *Bayesian multiple imputation* [21], which assumes a particular joint probability model over the feature values and draws imputed datasets from the posterior distribution of the missing data given the observed data. This approach has been used in health survey analysis [22] in the past. We used the `mix` open source package [23] for the the R [24] software environment in order to test the "off-the-shelf" capabilities of the method. However we found that `mix` has several limitations that impede its performance: it cannot use more than approximately 30 features (recall BRFSS has over 400 features), it works very slowly, and it is not capable of making use of features that have a high degree of missingness. In order to test `mix`, we had to hand-select 17 features create a single imputed dataset. Note that because of the modelling assumptions inherent in Bayesian multiple imputation, it is not possible to do separate feature selection when predicting different features.

Although we were eventually able to run `mix` on a portion of our dataset, the process was not straightforward, and required many error-prone translations between different dataset formats. As a consequence, we are unable to confidently describe the outcome of using `mix` for imputations of this nature, except to note that it is clearly not intended for this purpose. Future work may examine the construction of a more appropriate problem-specific version of `mix` suitable for use in a decision support environment.

In conclusion, our proposed hybrid system offers a substantial performance advantage over alternatives both in the absence of missingness (compared with machine learning systems), and the presence of missingness (compared with

purely knowledge-based systems or alternative methods of imputation). Consequently, our system represents a robust solution to the problem of partially missing data for decision support systems, especially in the medical domain.

3 Discussion & Conclusion

We have examined a real world problem of high importance: assisting medical professionals in making decisions based on current patient data and best practices encoded in a rule base, in scenarios where there may be missing data. Medical professionals with whom we have consulted[16] consider this to be a critical challenge for which solutions are needed, one that is in fact commonplace (with patients routinely omitting or misrepresenting their current profile). We demonstrate here that artificial intelligence techniques can be introduced to great advantage in order to address this problem, yielding accurate medical advice appropriate for patients and that in particular, simply relying on more traditional probabilistic reasoning in isolation would not deliver what is needed for this task.

Hybrid construction: We presented and validated a specific hybrid construction of a medical decision support engine. Our proposed system processes user queries mainly using logic based reasoning, and uses machine learning inference models to cope with missing information. This approach has the distinct advantages that all query results can be explained to the end user, and can be independently verified for correctness by a third party (since the answers are based on logic).

Although our validation strategy used a very specific sleep aid prescription scenario, the framework is generic enough to be used in other medical applications. In order to construct a solution for a different problem domain, a problem-specific ontological model for data representation is defined, along with the expert inference rules for decision making. Then a machine learning algorithm that works well with the given dataset can be used to predict missing values directly from the raw data. Once the basic primitives have been defined, the system construction is identical to the one proposed in this paper.

Related work: There is a great deal of interest in applying machine learning techniques for clinical decision support systems [25, 26]. For this purpose there are other approaches orthogonal to our work.

Zhu et al. [27] explored the use of machine learning algorithms in a geriatric patient rehabilitation setting. They provided a comparative evaluation of two machine learning techniques against the existing decision making process (using only a clinical assessment protocol-CAP). Their results demonstrated a definite advantage of using machine learning algorithms. However, they noted that the machine learning techniques (alone) produced more false positives and false negatives. Furthermore, the machine learning results were less readily interpretable.

Frize et al. [28] presented a different approach where a knowledge-based expert system was created to provide case-based reasoning capabilities. They transformed raw patient data into patient cases, and then provided inference rules

to perform "near-match" search queries. Their particular construction is different from Holmes as they only utilized statistical analysis (of the raw data) to determine weights for ranking the results

Conclusion: Our approach of integrating machine learning with ontological reasoning makes use of the inherent advantages of both approaches in order to offer the required accuracy for the medical domain. While we have sketched our framework in operation with specific real world data sets and rule bases, we have outlined how it can be employed in any medical decision context. Future work will focus on designing an effective user interface to the decision support system, with a view to progressing from an emerging application to one that is in fact deployed.

References

1. Garg, A., Adhikari, N., McDonald, H., Rosas-Arellano, M., Devereaux, P., Beyene, J., Sam, J., Haynes, R.: Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. JAMA: the journal of the American Medical Association **293**(10) (2005) 1223
2. Hunt, D., Haynes, R., Hanna, S., Smith, K.: Effects of computer-based clinical decision support systems on physician performance and patient outcomes. JAMA: the journal of the American Medical Association **280**(15) (1998) 1339
3. Kawamoto, K., Houlihan, C., Balas, E., Lobach, D.: Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Bmj **330**(7494) (2005) 765
4. Abidi, S., Hussain, S.: Medical knowledge morphing via a semantic web framework. In: Computer-Based Medical Systems, 2007. CBMS'07. Twentieth IEEE International Symposium on, IEEE (2007) 554–562
5. Shortliffe, E.: Mycin: Computer-based medical consultations (1976)
6. Pople, H.: Evolution of an expert system: from internist to caduceus. Artificial Intelligence in Medicine (1985) 179–208
7. Elhanan, G., Socratous, S., Cimino, J.: Integrating dxplain into a clinical information system using the world wide web. In: Proceedings of the AMIA Annual Fall Symposium, American Medical Informatics Association (1996) 348
8. Koutsojannis, C., Nabil, E., Tsimara, M., Hatzilygeroudis, I.: Using machine learning techniques to improve the behaviour of a medical decision support system for prostate diseases. In: Intelligent Systems Design and Applications, 2009. ISDA '09. Ninth International Conference on. (30 2009-dec. 2 2009) 341 –346
9. Doucette, J., Khan, A., Cohen, R.: A comparative evaluation of an ontological medical decision support system (omed) for critical environments. In: IHI 2012 - 2nd ACM SIGHIT Internatioanl Health Informatics Symposium. (January 2012)
10. Centers for Disease Control Prevention, .: Behavioral risk factor surveillance system survey data (2010)
11. Richardson, J.: (expert opinion). Mayo Clinic, Rochester, Minn. Nov. 11 (2009)
12. Drugs.com: Drugs.com — prescription drug information, interactions & side effects. `http://www.drugs.com/`
13. Roo, J.D.: Euler proof mechanism. `http://eulersharp.sourceforge.net/`
14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations **11**(1) (2009)

15. Mayo Clinic, S.: Prescription sleeping pills: What's right for you? http://www.mayoclinic.com/health/sleeping-pills/SL00010 (2011)
16. Gupta, H.: Personal communication with the family physician (2011)
17. Buettcher, S., Clarke, C.L.A., Cormack, G.V.: Information Retrieval: Implementing and Evaluating Search Engines. The MIT Press (2010)
18. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of ICML-97, 14th International Conference on Machine Learning. (1997) 412–420
19. Khan, A., Doucette, J., Jin, C., Fu, L., Cohen, R.: An ontological approach to data mining for emergency medicine. In: 2011 Northeast Decision Sciences Institute Conference Proceedings 40th Annual Meeting, Montreal, Quebec, Canada (April 2011) 578–594
20. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, Second Edition. Wiley-Interscience (2002)
21. Schafer, J.L.: Analysis of Incomplete Multivariate Data. CRC Press (1997)
22. Ezzati-Rice, T., Johnson, W., Khare, M., Little, R., Rubin, D., Schafer, J.: A simulation study to evaluate the performance of model-based multiple imputations in nchs health examination surveys. In: Proceedings of the Annual research Conference. (1995) 257–266
23. Brian Ripley, J.L.S.: mix: Estimation/multiple imputation for mixed categorical and continuous data. http://cran.r-project.org/web/packages/mix/index.html
24. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2011) ISBN 3-900051-07-0.
25. Harrison, R., Kennedy, R.: Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. Annals of emergency medicine **46**(5) (2005) 431–439
26. Pearce, C., Gunn, S., Ahmed, A., Johnson, C., et al.: Machine learning can improve prediction of severity in acute pancreatitis using admission values of apache ii score and c-reactive protein. Pancreatology **6**(1-2) (2006) 123–131
27. Zhu, M., Zhang, Z., Hirdes, J., Stolee, P.: Using machine learning algorithms to guide rehabilitation planning for home care clients. BMC medical informatics and decision making **7**(1) (2007) 41
28. Frize, M., Walker, R.: Clinical decision-support systems for intensive care units using case-based reasoning. Medical engineering & physics **22**(9) (2000) 671–677

A Multi-Agent Approach for Health Information Systems Domain

Luca Palazzo, Aldo Franco Dragoni, Andrea Claudi, Gianluca Dolcini

Dirpartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche
via Brecce Bianche, Monte Dago, 60131, Ancona, IT
{l.palazzo, a.f.dragoni, a.claudi, g.dolcini}@univpm.it

Abstract. Health Information Systems are characterized by an extremely wide and articulate domain: this is due to the distributed nature of data, processes workflow complexity, transactions and communications handling. One of the main problems which affects this context is the high system fragmentation caused by the different adopted solutions, the lack of communication standards and, thus, weak interoperability. Because of this, most efforts are going towards a message exchange standardization, integration profiles definition, homogeneous data handling, coordination among the various system entities, real time information management to support decision making and better patients involvement in their treatment process.
The aim of this paper is to describe why and how the HIS domain could take advantage from Software Agents Technology to deal with this kind of issues. After outlining the theoretical motivations, the paper focus on the development of an IHE profile conceived as a Multi-Agent System, trying to meet the needs previously argued.

Keywords: HIS, IHE Integration Profiles, HL7 v2, HL7 CDA, Multi-Agent Systems, Jade Framework

1 Introduction

The use of *Information and Communication Technology* (ICT) in health-care is necessary to achieve an effective quality of service, a better coordination among medical professionals and facilities. Hospital information systems are nowadays evolving towards *Health Information Systems* (HIS) [1], which, due to their inbuilt complexity, have become one of the most challenging and promising fields of research [2], with significant benefits to medicine and health-care in general. The complexity comes from an excessive fragmentation of solutions adopted by the health facilities in the various contexts, in terms of processes definitions and ICT choices. This means that most part of existing sub-systems hardly makes use of communication standards, process definition protocols and homogenous data representations, and thus completely lack of interoperability: this is valid not only for different health-care (public or private) systems, but also within the same medical institution. In addition, ICT is often totally absent from crucial

aspects in the treatment processes workflow [3], thus patient medical records and history in the HIS are hardly available. ICT allows, lastly, a much better citizens involvement and awareness towards their care process, making all useful information accessible and respecting, at the same time, their privacy.

The *TSE*, following the European Union guidelines, promulgated a document [3] which expresses HIS goals and requirements:

- patient clinical information must be available from every point in the territory
- it must respect, at the same time, the federated architecture of the existing HIS
- it must have a high security level and comply with privacy laws
- it must ensure a high reliability/availability level (24x7)
- it must be developed as a modular structure resistant to obsolescence
- it must be less invasive as possible on the already existing legacy systems, in order to safeguard investments
- it must adopt open standards

Based on recent feasibility studies [5], our work purpose is to integrate the so-called Multi-Agent paradigm to support the evolution of existing Health Information Systems: we thought that Software Agents Technology is particularly suited to reach these goals, especially because of its modular and distributed nature [11]. The key point is the adoption of standardization protocols: with the aid of *wrapper agents*, we propose a new software layer on top of the existing ones, compliant with established standards. That's why our project is based on *Integrating the Healthcare Enterprise* [6] (IHE) profiles, which define technical guidelines of typical health-care ICT scenarios combining a series of open standards, with the aim to improve systems interoperability.

The aim of this paper is to present how *Multi-Agent Systems* (MAS) can satisfy some of the previously claimed needs in term of standardization and interoperability, and how this technology could evolve to support a more Patient Oriented System. After a brief view on current MAS applications in the e-Health field and on what distinguishes our contribution, we will focus on modeling an IHE Integration Profile as a Multi-Agent System. The requirements analysis will be followed by a design and implementation phase: we chose *Tropos Modeling Language* [18] for system modeling and *Jade Platform* [9] for software development, even if this one is started recently and, thus, still in an embryonic stage.

1.1 Related Works

Nowadays Multi-Agent Systems are considered of the most interesting technologies for the development of healthcare applications and services [4]. In fact, many of the previously mentioned requirements in this field recall the MAS characterizing features: a distributed context marked by the needs of integration, communication, availability, reliability, fault tolerance, and so on. That's why it doesn't seem absurd to think of ubiquitous and pervasive e-Health systems built with Multi-Agent Technology.

Most of MAS healthcare projects belongs to Ambient Assisted Living context [19-20], mainly due to recent MAS applications in the home automation field. The goal of these systems is to create a smart environment equipped with sensors to gather information about patient status and with multimedia devices to interact with him. Other MAS applications are connected to diagnosis area: diagnosis support systems use data mining techniques to extract useful information from different data sources [21].

With this paper, instead, we want to focus primarily on interoperability issues: we consider our work as a first step to make HIS an ubiquitous entity where all legacy systems can communicate each other, handle the same information, use the same process workflow. To reach this aim, we can consider MAS as an additional abstraction level, on top of the existing systems, compliant to communication and data management standards promoted by the research community.

2 Scenario Description

As we stated in the introduction, our main goal is to demonstrate how multi-agent technology can address the lack of interoperability of existing e-Health systems. To reach this aim, we have to start from analyzing a typical healthcare process model to examine which entities are involved in the process, what kind of information they need to execute their activities, how data are managed or shared by each other, and so on.

For this reason we decided to use the documentation provided by IHE, a consortium of healthcare and ICT experts created to improve the way computer systems should share information using established standards. In particular, IHE profiles define all the actors involved in several healthcare scenarios, the process workflow of these scenarios and the messages exchange among the various actors using the *HL7* [7] *v2* standard.

Regarding the patient documentation management, instead, we adopted the *HL7 Clinical Document Architecture* (CDA) structure, which is part of the *HL7 v3* standard and lately stands out [8] thanks also to its persistent and human-readable XML-based markup language. This choice is driven by the fact that some processes could end with a document generation (e.g. a report), thus it's equally important that patients data can be available and readable from every node of our ubiquitous system in a widely accepted standard.

2.1 IHE Integration Profiles - Radiology Scheduled Workflow

The process flow considered in our experimental scenario is the IHE *Radiology SWF* [12] (Fig.1).

All IHE technical frameworks are organized in the same way. In particular they are based on three main concepts:

- *actors*, which perform specific informatics micro-systems
- *transactions*, that is actors interactions

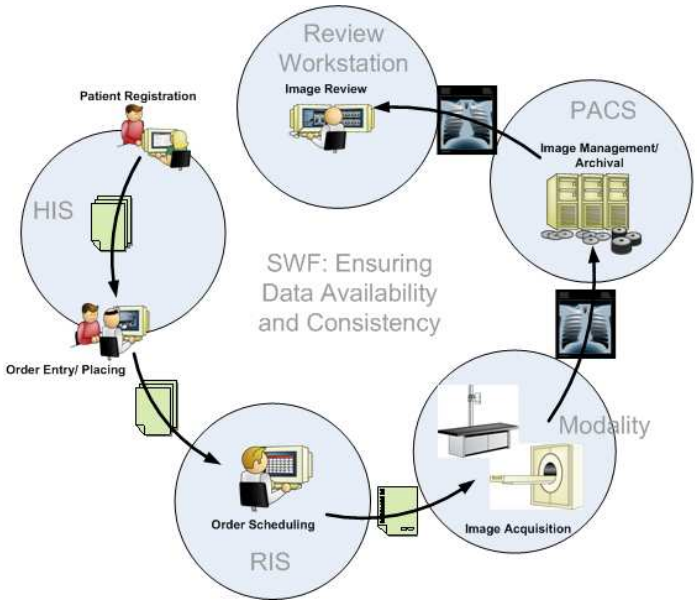


Fig. 1. This figure summarizes the information flow among the various actors of "Radiology Exams" domain

– *integration profiles*, that, as we said, represent typical scenarios and solutions defined by IHE for the various contexts

Let's analyze more closely the actors of our case of study [13], each one of them will be modeled by a software agent in the architecture design:

1. *ADT*: it has the role to register new patients, storing his personal details, contact details, etc. Subsequently, registration of the same patient will be no longer be necessary, even if it will be useful to record any additional information
2. *Order Placer*: it has the role to ask the reference department (RIS in this case) to schedule an examination
3. *DSS/RIS*: it must first determine the availability for a specific examination. After the scheduled performance, it must notify the Image Manager
4. *Acquisition Modality*: it is part of the diagnostic, and it must communicate with the Image Manger/Archive to request the image processing/saving
5. *Image Manager/Archive*: the images generated from the diagnostic machine must be store in reliable and durable archive, making them also available for reporting
6. *Clinical Reporting*: once the examination is ended and all images are available in the PACS (Picture Archiving and Communication System), it is possible to draw the clinical report

From the transactions point of view, what occurs is much more articulated, and it is expressed through the aid of the subsequent sequence diagrams (Fig.2).

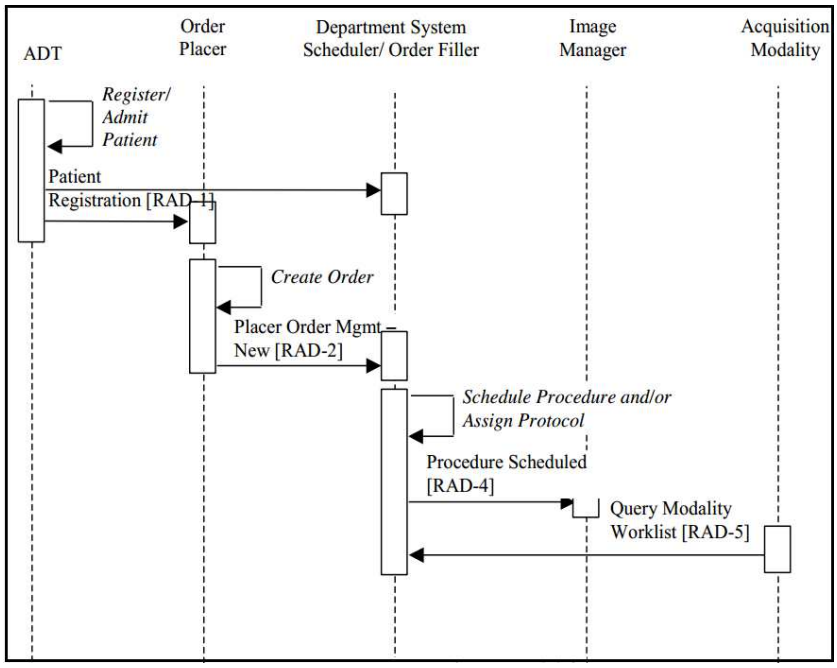


Fig. 2. Sequence diagram of administrative process flow transactions in the IHE radiology profile

Each transaction consists of a series of message exchanges, modeled by HL7 v2 standard, with, obviously, a specific syntax and semantic (Fig.3).

HL7 v2 Messaging *Health Level Seven* (HL7) is a no-profit corporation that deals with developing standards in health-care. These ones can be conceptual (RIM), documentation (CDA), application (CCOW) or messaging (HL7 v2 - v3) standards. Even if it is gradually migrating from HL7 v2 to HL7 v3 messaging protocols due to a more semantic oriented message definition and a better human-readable structure [14], we chose to adopt version 2 for our project. This is because it's still a very common format and, in contrast with v3, it's used on large scale. In addition, IHE profiles are entirely defined by transactions of HL7 v2 messages, and this is a crucial aspect to reach our interoperability goal.

2.2 HL7 Clinical Document Architecture (CDA)

We expressed our motivations on the choice of the HL7 v2 messaging standard, but this doesn't mean we rejected v3 at all. For the persistent handling of patient health information (Patient Summary, Clinical Reports, etc.) we decided to follow the TSE guidelines [15] with the adoption of the HL7 CDA, an XML-based clinical document developed on the *Reference Information Model* (RIM), and

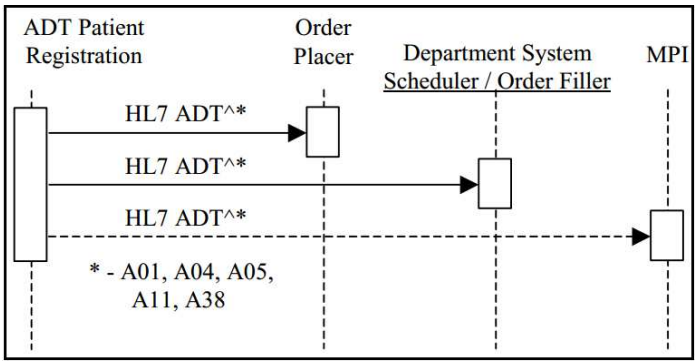


Fig. 3. HL7 v2 messages exchange in a RAD-1 transaction

thus belonging to HL7 v3 set of standards. The document is composed of two sections: the former, named "header", consists in a set of document identification data and patient personal details, the latter, named "body", contains clinical information in a both structured and unstructured way (Fig.4).

The CDA specifies that the content of the document consists of a mandatory textual part (which ensures human interpretation of the document contents) and optional structured parts (for software processing). The structured part relies on coding systems (such as from SNOMED and LOINC) to represent concepts [16]. As we said, TSE reports the meaning of all XML tags in the CDA *Schema Definition* (XSD) and how they should be filled.

3 Architecture Design

In this section we illustrate how the previously presented IHE process workflow can be modeled using MAS technology. A software agent is an entity located in an environment usually with other agents, which is able to perceive and manipulate the environment to reach its goals and to operate in an independent manner, thanks to its distributed nature [17]. It is, obviously, able to exchange messages with other agents in the same environment: depending on their aims, they can cooperate, negotiate or compete each others. A *Multi-Agent System* is, therefore, a distributed social entity where each member has its own tasks and an environment partial vision, but they can share their knowledge and skills to reach complex goals.

Our goal requires a high standardization level: in this field it comes thanks to *FIPA* (Federation for Intelligent Physical Agents) [10]. FIPA is an international non-profit organization, founded in 1996, which developed a set of specifications relating to MAS technology, in terms of agents architecture and communication. We chose to develop our project through the use of *JADE Framework* [9], a java platform developed by Telecom Italia published through LGPL license, which allows and simplify the implementation of FIPA compliant software agents. Jade Framework is composed of three fundamental elements:

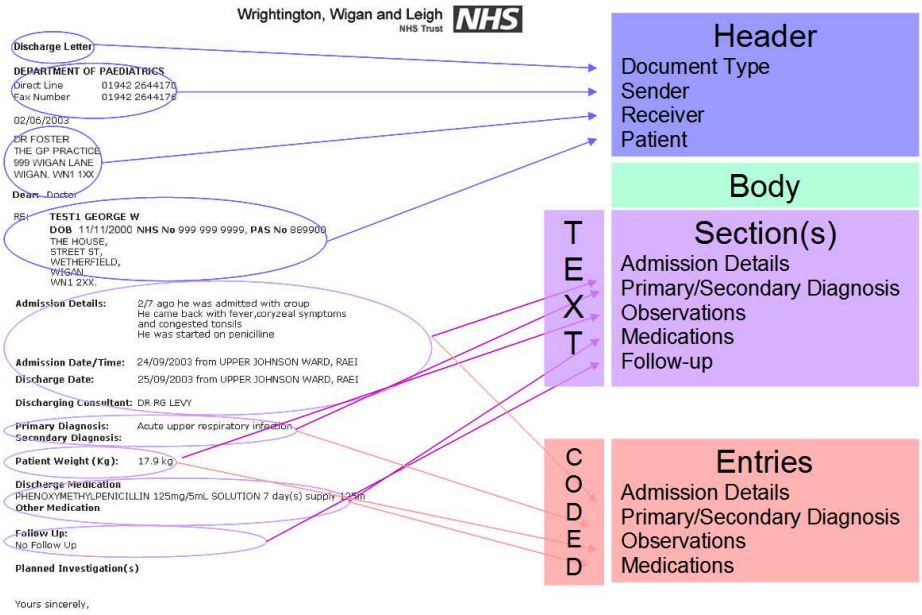


Fig. 4. HL7 v3 CDA structure example

- a runtime environment where agents can live
- a set of APIs to simplify programmers multi-agent systems development
- a graphical tools collection, to handle and monitor active agents

FIPA, besides, developed a specific language named FIPA SL (Semantic Language) to compose the content of a message through the use of previously built ontologies. As stated on IHE guidelines, the content of messages exchanged among the agents (integration profile actors) is expressed according to HL7 v2 standard, but we use FIPA SL for the communication between an agent and its own wrapper.

3.1 Modeling with Tropos

Tropos is a recent design methodology based on the multi-agent paradigm [18], which supports the project from requirements to system implementation. It offers a structured approach to software development, based on conceptual models construction defined in a visual modeling language, whose elements are concepts such as actors, roles, beliefs, goals, and so on.

In this way we modeled the IHE Radiology Scheduled Workflow profile and defined each entities involved in the process. In Tropos design there are two level of depth, known respectively "Early" and "Late" Requirements Analysis: the former is the domain analysis to identify individual actors and their goals

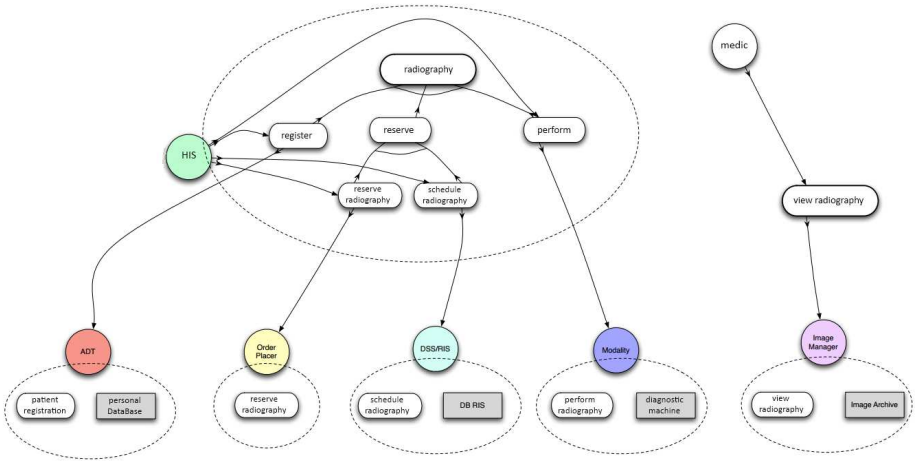


Fig. 5. Tropos Early-Requirements Analysis of IHE Radiology Scheduled Workflow

(Fig. 5); the latter has much more complex architecture and reports all the tasks and communications needed to reach these goals.

Each actor is performed by a distributed software agent which must interact with the other ones, following IHE guidelines presented in the previous section of this paper. Furthermore, some actors interacts directly with human users (for instance to register patient personal details or to draw a clinical report), so we need to provide certain agents of graphic user interfaces (GUIs).

However, as we already mentioned, agents provide a higher software layer on top on the existing informatics architecture: we need an interface to communicate with legacy systems to ensure interoperability with MAS. Most efforts are in *wrapper agents* designing, since they perform the crucial task to insert and obtain the appropriate information from legacy DBMS, to convert standard messages in understandable information, to make data consistent. This is achieved only through a start-up phase, where, for each software system, an "ad hoc" own wrapper agent will be designed. In addition to these, we have two more kind of agents, introduced by JADE Platform: the AMS and the DF. The former is unique for each platform and contains the address of agents bound to its container, the latter keeps track of the services provided by each agent in its container. Despite MAS distributed nature, these kind of agents could represent single points of failure, that's why we thought to use recovery techniques to keep the system operational. Finally, we must pay close attention to transmitted data with the adoption of cryptographic protocols to provide communication security.

3.2 Implementation

Implementation is currently in an early stage: we started to develop the first agents behaviours through Jade Platform and to define ontologies for their communication, but we still need much work in order to deploy a system compliant

with all IHE profile specifications, so we deal with this aspect in our future works. For the generation and parsing of HL7 v2 messages we use *HAPI* [22], an open source object oriented HL7 2.x parser and library for Java (Fig.6).

```
ADT_A01 adt = new ADT_A01();

//MSH Segment
MSH mshSegment = adt.getMSH();
mshSegment.getFieldSeparator().setValue("|");
mshSegment.getEncodingCharacters().setValue("^~\\&");
mshSegment.getSendingApplication().getNamespaceID().setValue("Databasedell'ADT");
mshSegment.getReceivingApplication().getNamespaceID().setValue("OP");
mshSegment.getDateTimeOfMessage().setValue(convertedDate);
mshSegment.getMessageType().getMessageCode().setValue("ADT");
mshSegment.getVersionID().getVersionID().setValue("2.6");
mshSegment.getSequenceNumber().setValue("123");
mshSegment.getMessageType().getTriggerEvent().setValue("A01");
.....
```

Fig. 6. Example of a MSH segment filling of an ADT message type using HAPI

4 Conclusions

The purpose of this paper is to show the effective usefulness of Multi-Agent Technology in a highly complex environment as HIS are. We demonstrate how it can be used to address certain issues that afflict this domain:

- Multi-Agent Systems can provide a higher interoperability level, thanks to the most established standard adoption for messages exchange among the actors of the system
- they can, with wrapper agents aid, represent a higher software layer on top of existing legacy systems, so they form a minimally invasive architecture respectful to the federated HIS one
- they have a distributed nature and are characterized by a high modularity level, so they can be modified without an excessive impact on the system and can better resist to obsolescence
- they can ensure a high data reliability and availability level from any point in the territory

To do this we chose an Integration Profile defined by IHE and reconsider it in a Multi-Agent perspective. This means that we considered each actor of the Scheduled Workflow as an agent with his goals and roles. At the same time, following IHE guidelines, we adopted HL7 v2 standard to message exchange among the entities, and HL7 CDA for clinical documents handling, to overcome the lack of interoperability. We used Tropos Modeling Language to design the software architecture and Jade Framework for agents implementation, even if this one is still in an early stage.

To conclude, this paper just represents a first step in the adoption of this kind of technology inside HIS articulate domain, but we believe that it will support

evolution in this field, not only, as we argued, in terms of interoperability support, but even to achieve a more Patient Oriented System: this path will be subject of future development and implementation.

References

1. K. A. Kuhn; D. A. Giuse. From hospital information systems to health information systems: Problems, challenges, perspectives. Methods of Information in Medicine, 40, no4:275–287, 2001.
2. Peter L. Reichertz. Health information systems past, present, future. International Journal of Medical Informatics, 75, Issue 3(3):282–299, May 2006.
3. Ministro per l'Innovazione e le Tecnologie. Strategia architetturale per la Sanità Elettronica delle Regioni e delle Province Autonome, 31/03/2006.
4. Federico Bergenti; Agostino Poggi. Multi-Agent Systems for e-Health: Recent Projects and Initiatives. Università degli Studi di Parma, 2011.
5. Minh Tuan Nguyen; Patrik Fuhrer; Jacques Pasquier-Rocha. Enhancing E-Health Information Systems with Agent Technology, International Journal of Telemedicine and Applications, vol. 2009 doi:10.1155/2009/279091.
6. Integrating the healthcare enterprise. http://www.ihe.net/
7. HL7 International - http://www.hl7.org/
8. Michel Treins, Olivier Curé, Gabriella Salzano. On the interest of using HL7 CDA release 2 for the exchange of annotated medical documents. Université de Marne-la-Vallée, 2006.
9. Java agent development framework. http://jade.tilab.com/, 2005.
10. Foundation for intelligent physical agents. http://www.fipa.org, 2005.
11. F. Bellifemine, G. Caire, D. Greenwood. "Developing Multi-Agent Systems with JADE", 2007.
12. Integrating the Healthcare Enterprise, IHE Technical Framework, Integration Profiles. Revision 10.0 – February 18, 2011.
13. Andrea Spada. "L'informatizzazione del Workflow in Sanità – Il modello di integrazione IHE".
14. "The HL7 Evolution", 2012 - http://www.corepointhealth.com
15. TSE – Specifiche tecniche per la creazione del "Profilo Sanitario Sintetico" secondo lo standard HL7-CDA Rel.2 – 24 Novembre 2010.
16. H. Yun, K. Kim. Processing HL7-CDA Entry for Semantic interoperability, 2007
17. Franco Zambonelli; Nicholas R. Jennings; Michael Wooldridge. Organizational Abstractions for the Analysis and Design of Multi-agent Systems, In Proceedings of AOSE'2000, pp.235 251
18. Tropos Modeling Language, http://www.troposproject.org/
19. Jih, W., Hsu, J.Y., & Tsai, T. "Context-awareservice integration for elderly care in a smart environment." In D.B. Leake, T.R. Roth-Berghofer, & S. Schulz, (Eds.), 2006 AAAI Workshop on Modeling and Retrieval of Context Retrieval of Context, (pp. 44-48). Menlo Park, CA: AAAI Press.
20. K4CARE (2007) K4CARE project Web site, http://www.k4care.net.
21. Hadzic, M., Chang, E., & Ulieru, M.. Soft computing agents for e-health applied to the research and control of unknown diseases. Information Sciences, 176:1190-1214, 2006.
22. HAPI Project - http://hl7api.sourceforge.net/

A comparative analysis of SNOMED CT and the reference ontology ROME

Marta Gentile, Aldo Franco Dragoni,
Università Politecnica Delle Marche, Italy

Abstract. Nowadays the healthcare domain shows interoperability problems, which can be solved by ontologies. This paper proposes a comparative analysis between two ontologies in biomedical field: SNOMED CT (the most comprehensive clinical healthcare terminology in the world) and ROME (an alternative one developed by the italian National Research Council - CNR). Key parameters of the analysis regard the international diffusion and acceptance, the ontological correctness, the identification of their expressivity and the evaluation of their computational complexity. The goal is to get a clear picture of these two important alternatives for Italy, to understand whether and why SNOMED CT would be the best choice for our country.

Keywords. ontology, SNOMED CT, OWL, Description Logic.

Introduction

The healthcare domain currently shows interoperability problems: clinical and organizations often use different clinical terms that represent the same concepts, and whenever they have to enter data in different databases, there are inevitably integration problems that must be settled case by case. An accepted and verified system of international codes solves only part of this problem. In fact, there is also the need to provide this information in such a way that they can be understood and processed from computers: so we need ontologies.

The term ontology is borrowed from philosophy: in computer science an ontology is the attempt to formulate an exhaustive and rigorous conceptual schema within a given domain. Particularly, in the healthcare domain we talk about biomedical ontology: it is focused on defining the main biological classes and relations among them. The principal advantages of ontologies are enhanced advanced software and better exchange of information among different systems: in the healthcare domain ontologies improve quality and safety.

1. Material and Methods

1.1. SNOMED CT

SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) is a clinical terminology increasingly guided by ontological principles. It is considered to be the most comprehensive and multilingual clinical healthcare terminology in the world: it is

a resource with comprehensive and scientifically-validated content, it is scalable and flexible, and it is already used in more than 50 countries around the world, but not yet in Italy. The relevance that SNOMED is increasingly taking in recent years in the international scientific community is demonstrated by his presence in many scientific papers: for instance, SNOMED CT is a recurrent issue in the proceedings of MIE 2009 [1].

The IHTSDO is the international organization that owns and administers the rights of SNOMED CT, including the rights to issue SNOMED CT licence. Members have rights and responsibilities, including the right to help improving this ontology. IHTSDO invoices the member for the annual fees: according to the World Bank GNI Atlas Based Fees for 2007, published on HITSDO official website, Italy would pay about 680,000 USD annually [3].

The basic components of SNOMED CT are:

- Concepts: SNOMED CT includes more than 311,000 unique concepts, but this number is growing. The concepts represent clear clinical significances, and they are organized in hierarchies, from the general to the specific.
- Concept descriptions: they are the terms or names assigned to a SNOMED CT concept. There are almost 800,000 descriptions in SNOMED CT, including synonyms that can be used to refer to a concept.
- Relationships: there are approximately 1,360,000 links or semantic relationships among the SNOMED CT concepts. These relationships provide formal definitions and other characteristics of the concept.

When a country decides to use SNOMED CT, it has to translate his resources into the target language. This is a complex process and it shows many language and terminology issues[2]. Once translated and nationalized, SNOMED is used in different experimental situations, as suggested by some of the papers presented at MIE 2009 [1]: such as the automatic mapping of clinical documentation to SNOMED CT terms [3], or the identification of Reference Sets for the structured representation of well defined clinical domains [4]. The work described in this paper belongs to the largest panorama of analysis or comparison among SNOMED and similar regional nomenclatures/ontologies pertaining to the health sector, in order to determine and justify the specific uses [1].

1.2. ROME

ROME (Reference Ontology in Medicine) is an italian ontology, recently developed by the CNR. A reference ontology is an intermediate layer between the top-level ontologies (formal ontology) and domain ontologies (which relates to specific domains). ROME consists of about 200 general entities, therefore it has a much smaller size than SNOMED CT. ROME is based upon the DOLCE top-level ontology. From DOLCE, it inherits the basic distinction between endurants and perdurants [5].

The main application of ROME is the design of several domain-specific ontologies that are mapped to it, and may be regarded as a specialized plug-ins covering different domains.

1.3. Analysis: ontological and knowledge-engineering correctness, expressiveness and computational complexity

The first aspect considered in this paper is the ontological correctness of both SNOMED CT and ROME. In general, a high-quality ontology should be modularized, with a clear separation between the formal top-level ontology (e.g. DOLCE), a shared reference ontology of medical knowledge and a set of (sub)domain ontologies of different medical specialties. The methodology followed was different in these two ontologies, because ROME was available for us, while SNOMED was not, since Italy has not the license yet. To analyze ROME we used Protégé, a software tool to develope and edit ontologies: we used it to navigate the ontology, and his reasoner to check consistency of classes. We followed another method to study SNOMED CT: starting from the most common errors that are reported in the literature [6], we have made a punctual examination using free browsers on the web, that allow to browse this ontology [7]. In this way we found concrete examples of SNOMED CT ontological errors, and we also found that many errors were corrected in the latest releases.

The second aspect of our analysis concerns an evaluation of expressivity and computational complexity of the two ontologies. They are both expressed using different sublanguages of OWL: it is based on description logics (DLs), which are a family of logics that are decidable fragments of first-order logic [8]. There are many varieties of DLs, and any kind of DL has a different level of expressiveness, thus complexity [9]. Therefore the aim of this paper is to determine which of the two ontologies reaches the best trade-off about expressiveness an computational complexity. About SNOMED CT, we trusted publications about it [10, 11]. On the contrary, about ROME, we did a comprehensive analysis (using Protégé) of OWL constructs actually used in its entities definition. ROME also uses such entities defined in DOLCE LITE 397, which in turn are defined in DOLCE LITE: for these concepts we have analyzed the original definitions. Then, for each construct found, we looked for correspondence with DL operators, and at the end we found the specific DL on which ROME is based.

Finally, we have identified the computational complexity relates to expressivity of both the two ontologies. About ROME, it was sufficient to observe that its DL is an extension of another DL whose complexity is known in the literature [9], and then we verified this result by a free application available on the web [12]. For SNOMED, we still trusted about results published in the literature [10, 11]

2. Results: ontological and knowledge-engineering correctness, expressiveness and computational complexity

About the correctness of ontologies, after we have performed the analysis described in the previous section, the following most important types of ontological errors can be abstracted:

- Hierarchy violating the rules of sound ontology engineering (inconsistent classification to DOLCE). For example “smoker” (a kind of agent) is subsumed by “tobacco smoking behaviour-finding” (a role). DOLCE clearly distinguishes between a role and the agent which plays that role [6].
- Common use of multiple inheritance, with frequent subsumption errors. For example, alcoholic beverage (through its parent ingestible alcohol) is subsumed by central depressant, ethyl alcohol and psychoactive substance of

abuse non-pharmaceutical. From a philosophical point of view none of these subsumptions is true. Alcoholic drinks contain ethyl alcohol which plays a role of depressant and substance of abuse (with respect to human beings) [6].

- Sometimes lack of exact mereology anatomy, omission of apparently obvious relations, violation of medical thought and biomedical knowledge, etc [6].

The frequent occurrence of these errors can create problems to the rationality of automated reasoning. Luckily SNOMED CT is continuously updated to meet the needs of users around the world. Therefore we can rely on work of users of the terminology to discover and revise any mistakes and ambiguities.

On the contrary, ROME is a reference ontology, it is based upon the DOLCE top-level ontology and it is separated from other domain ontologies. Therefore, it is a correct ontology according to definition described in the previous section. After a more careful analysis, we found that it have inaccuracies and it is incomplete (often it provides only a partial instantiation of certain levels of detail) and unclear (e.g. about property names).

As far as the study of expressiveness and computational complexity, we have found that SNOMED CT is expressed in OWL2 EL (which corrsponds to the DL EL++), that is a subset of OWL2, with very few constructs available (therefore it is not very expressive), particularly useful in applications employing ontologies that contain very large numbers of properties and/or classes [8]. Instead Rome, in its namespace declaration, says that it refers to things drawn from the namespace of OWL2. Protégé indicates that ROME is expressed by the DL ALE. On the contrary, the DL of DOLCE LITE 397 is SHIF, while the DL of DOLCE LITE is SHI. The DL SHIF extends SHI, that extends ALE. ROME uses seveal ontologies entities based on DOLCE: analyzing only endurant entities, we have found such constructs that are a subset of DL SHIF operators. Therefore the analysis focused on finding examples of transitive roles and functional properties, which would ensure the achievement of DL SHIF. Transitive roles are used by defining endurant and non - physical-endurant (entities that are used also in ROME), while the only functional property of DOLCE (life) is not used in Rome. So we can assert that ROME is based on the DL SHI.

A comparison among the DL EL++ operators (SNOMED CT) and the SHI operators (ROME) presents a subset of constructs that are identical, and constructs in excess in both sides.

About complexity of ROME, since its DL SHI is an extension of DL ALE (which have a well known complexity in literature), subsumption and satisfability problems (therefore also other reasoning problems due to the available operators) are at least NP-complete, already for simple expressions of concepts [9]. In detail, reasoning in Rome is EXP-TIME complete [12]. This complexity is due to the disjunction operators and the universal quantifier used in conjunction with the existential quantifier construct. However, an ontology with this complexity is usable only for small knowledge bases, thanks also to sophisticated optimization techniques: but increasing the knowledge base size, reasoning grow up and it becomes intractable. Rather SNOMED has good computational properties for large-scale ontologies: it provides reasoning in polynomial time (the basic reasoning problems can be performed in time that is polynomial with respect to the size of the ontology), thanks to the absence of universal quantifier and the use only of a restricted form of role-inclusion [10, 11].

3. Conclusions

In this paper we compared two biomedical ontologies: the international ontology SNOMED CT and the italian ontology ROME.

Firstly, the most obvious differences are their diffusion and acceptance: SNOMED CT is the most comprehensive healthcare terminology in the World, while ROME is used in sporadic applications only in Italy! The second point is the ontological correctness: SNOMED CT has ontological errors that are absent in ROME, which, on the other hand, has some inaccuracies. About the expressiveness, we can not assert which one is more expressive; but about their computational complexity we can assert that it is *exponential* for ROME (so increasing the knowledge base size, reasoning grows up and becomes intractable) while it is *polynomial* for SNOMED, so that the latter remains tractable for large amounts of data.

In order to evaluate which alternative would be better for Italy, we must remember that the primary purpose of ontologies is to ensure interoperability or, better, international interoperability, that is of course assured only by SNOMED CT. Even renouncing to reach interoperability at an international level, ROME should be expanded by adding other domain ontologies to achieve a greater coverage of topics in health-care domain: but due to its complexity, reasoning would become intractable, which means that the ontology would become unusable. According to our analysis, Italy should become a member of HIT SDO and collaborate to fix contents of SNOMED CT and "nationalize" it to achieve a complete interoperability and benefit from its excellent computational properties.

References

[1] K.P.Adlassnig, B.Blobel, J.Mantas, I.Masic: *Medical Informatics in a United and Healthy Europe, Proceedings of MIE 2009,* IOS Press, Netherlands, 2009

[2] HITSDO, *http://www.hitsdo.org*

[3] H.Stenzhorn, E.J.Pacheco, P.Nohama, S.Schulz, Automatic mapping of clinical documentation to SNOMED CT*, Medical Informatics in a United and Healthy Europe, Proceedings of MIE 2009* (2009), 228-232

[4] A.G.James, EugeneNG, P.S. Shah, A Reference Set of SNOMED terms for the Structured Representation of Respiratory disorders of the Newborn Infant*, Medical Informatics in a United and Healthy Europe, Proceedings of MIE 2009* (2009), 243-247

[5] D.M.Pisanelli, M.Battaglia, C.De Lazzari: *ROME: a Reference Ontology in Medicine,* New Trends in Software Methodologies, Tools and Techniques-Proceedings of the sixth SoMeT 07, (2007).

[6] G.Hja,G.Surjn,P.Varga: *Ontological analysis of SNOMED CT,* First European Conference on SNOMED CT,Copenhagen-Denmark(2006).

[7]Snoflakebrowser,*http://snomed.dataline.co.uk*,SnomedCtCoreBrowser, h*ttp://terminology.vetmed.vt.edu/default.htm*

[8] W3C, *http://www.w3.org/*

[9] F. Baader, D. L.McGuinness, D.Nardi, P.F.Patel-Schneider: *The description logic handbook: theory, implementation, and applications,* Cambridge University Press, (2003).

[10] F.Baader,S.Brandt,C.Lutz: *Pushing the EL Envelope*, Institute for Theoretical Computer Science TU Dresden,Germany

[11] F.Baader, S. Brandt, C.Lutz: *Pushing the EL Envelope Further*, Institute for Theoretical Computer Science TU Dresden, Germany

[12] Complexity of reasoning in Description Logics, *http://www.cs.man.ac.uk/ ~ezolin/dl/#note8*

Searching for patterns in clinical data - Choosing the right data mining approach

A. Fazel Famili[1], Ziying Liu[1], Andrea Bravi[2], and Andrew Seely[2]

[1] Knowledge Discovery Group, National Research Council Canada, 1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada
{Fazel.Famili, Ziying.Liu}@nrc-cnrc.gc.ca

[2] University of Ottawa, 501 Smyth Road, Ottawa, K1H 8L6, Canada
{andrea.bravi, ajeseely}@gmail.com

Abstract. Clinical parameters are normally collected for all classes of patients, including the ones in critical care. Focusing on the objective of identifying interesting patterns related to physiological variability, we briefly report on our investigation in searching for useful patterns from clinical parameters that include heart rate and respiratory rate measurements. We report on our investigation in understanding the domain, the breadth and depth of data, what should be the right data mining paradigm and how it is related to a novel application of Artificial Intelligence (AI) in life sciences.

Key words: Knowledge Discovery, Clinical Data Analysis.

1 Introduction

Advanced data analysis methods for data storage, analysis, pattern discovery and visualization have been developed in response to the substantial challenges posed by the quantity, diversity and complexity of clinical data. This data consists of various forms among which are: (i) meta data, (ii) genomics, (iii) proteomics, (iv) metabolomics, and (v) clinical reports. The latter normally consists of textual information (clinical reports containing structured fields and free text) and granular variables such as heart rate and respiratory rate. However, whether the objective is early detection or prognosis, a number of these clinical parameters need to be measured and monitored over time for which many forms of variability techniques have been considered [1]. Similar to the idea of constructive induction in machine learning where the argument is: data representation contributes to 99% of a good learning performance and normally one transforms the original data into (or derives from) new attributes, it has been understood that effective variability analysis requires deriving secondary attributes from the original parameters [1, 7]. However, regardless of the amount of data, learning meaningful patterns, explanatory or predictive, from variability analysis of clinical parameters is not a trivial task. This short paper summarizes our attempts to discover meaningful patterns from transformed clinical data obtained from a group of patients in critical care. The objective is to investigate use of unsupervised and supervised machine learning methods and identify the right approach for this investigation. We briefly explain the problem, the data used in this study, our data preprocessing and understanding and preliminary data analysis methods applied.

2 Problem statement

Identifying abnormal physiological variability has been attributed to a more accurate diagnosis and successful treatments of patients [4]. It is therefore important to accurately analyze critical care patient data and discover patterns that assist in choosing the best treatment strategies, such as risk-free extubation. Here the main problem under study is to predict whether a patient under mechanically assisted ventilation can be safely extubated, a problem of high interest in the clinical practice [5, 6].

The standard procedure to do so is called a spontaneous breath trial (SBT), where the machine providing the assistance is turned off, and the conditions of the patient are monitored for a variable number of minutes (usually at least 30 min). At the end of the SBT the physician decides whether it is appropriate to extubate the patient, or keep him intubated and then perform another SBT at a successive time. If the extubated patient needs to be reintubated (meaning his cardio-respiratory system does not allow him to breath autonomously), it is said that the patient failed the extubation; if no reintubation is needed, then the patient is passed the extubation. In our case, we plan to predict the extubation outcome by monitoring heart rate and respiratory rate variabilities, for those clinical parameters are easily available in the intensive care units. The problem is therefore considered highly suitable for an active and incremental learning application in which one would first start with an appropriate batch of data to identify the most suitable data mining approach. The study can then be continued into an active and incremental learning application. Our objective is to generate sufficiently validated results so that we can eventually build a decision support system that may be considered partially autonomous.

3 The Data and its scope

The data used in this study consisted of 102 measures of Heart Rate Variability (HRV) and Respiratory Rate Variability (RRV) recorded for one hour simultaneously. The total number of patients was 42; of these, 36 were labeled as passed extubation, and 6 failed extubation. These patients belonged to similar group with ethical permission. Each measure of variability consisted of 12 values obtained through a windowed analysis (window size of 5 minutes, with no overlap), representing two sets of time intervals; 6 before and 6 during SBTs – i.e. 30 minutes each (Fig 1). Since patients related to this class of problems (critical care) were acquired at different times of the day, this data may represent a small batch of what could be available on a regular basis. A review of some of the variability attributes is given in [1].

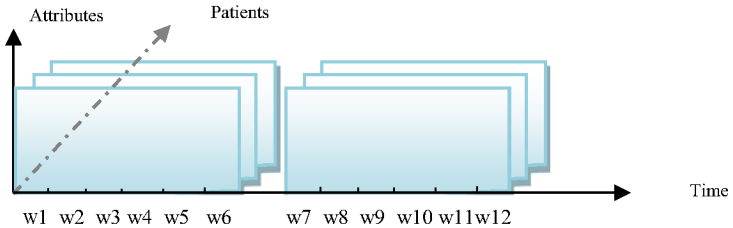


Figure 1. Example of a three dimensional clinic data

4 Data Preprocessing and Data Understanding

This is a unique case study for which although the data is collected over time (a total of 12 windows for each measure of variability), it cannot be considered as a normal time-series. A time series is defined as a sequence of data points, measured typically at successive time instants spaced at equal time intervals. Time series data have a natural temporal ordering.

Following is a list of our data preprocessing and data analysis efforts for HRV dataset

a. We removed attributes for which ≥80% of values were missing. We also removed 3 of 42 patients because of missing values.

b. Since derived attributes have various distribution properties (e.g. min/max), we decided to normalize the data using linear transformation in the range [-1, 1]. Through the analyses described in the following, we learned it was appropriate to highlight the clinically relevant information.

c. Assuming that we can treat the data as a normal time-series, we tried to analyze it using our cluster mapping (CM) algorithm [2]. This algorithm can identify a set of variability measures that at any time point along the time axis can influence the behaviour of another group of attributes at other time point. Using CM algorithm we performed a s et of experiments to discover rules/ associations between different classes of patients. With respect to the functionality of the CM algorithm, the results were not interesting and seemed that the data cannot be treated as a time-series data.

d. We then took mean ratio of the two windows (before and during extubation) for all attributes across all the patients. We removed outliers using BioMiner software [3] and domain knowledge. Then we tried to visualize the difference between passed and failed patients using a heat-map graph. We also used hierarchical clustering to cluster the mean ratio data, displayed with a dendogram. The idea was to identify and remove some redundant attributes through visual inspection. Although the results seemed interesting (as we observed noticeable intensity spectrum distribution for all 102 derived attributes) and could be used for attribute selection, we decided that the initial attribute filtering/selection should be done by the domain expert, who performed this task and selected 43 out of 102 attributes for subsequent analysis.

e. Looking at the distribution of all 43 parameters, we noticed anomalies existed in certain attributes. Therefore, we decided to treat all values that were far by 3-4 folds in total from their nearest neighbours as anomalous and replaced them with "NA". This would prevent creating biased results in the next steps. We also removed the data for two attributes that had too many missing values for all patients. Therefore we continued with 41 attributes.

f. At this stage, we used the newly selected attributes (41) and calculated the covariance ratios of all attributes from the two windows (before and during extubation) for all patients. This was done using the covariance equation {1}:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{(n-1)} \qquad \{1\}$$

where X_i, Y_i indicate the i^{th} value in the sets of before and during SBTs, respectively and $\bar{X}, \bar{Y}$ indicate the mean of the sets before and during SBTs, respectively.

It was our understanding that given the statement of the problem (two state situation, before and during extubation) and based on our preliminary data preprocessing, we should pursue any form of data analysis using transformed data in the form of covariance ratios.

5 Data analysis and prospects for learning patterns

Based on our understanding of the problem under study and the type of data (batches and their availability), we can consider this problem as being highly suitable for the applications of artificial intelligence in life sciences. However, there is a limit on what can be discovered so that it is considered as new (not previously known) and useful knowledge.

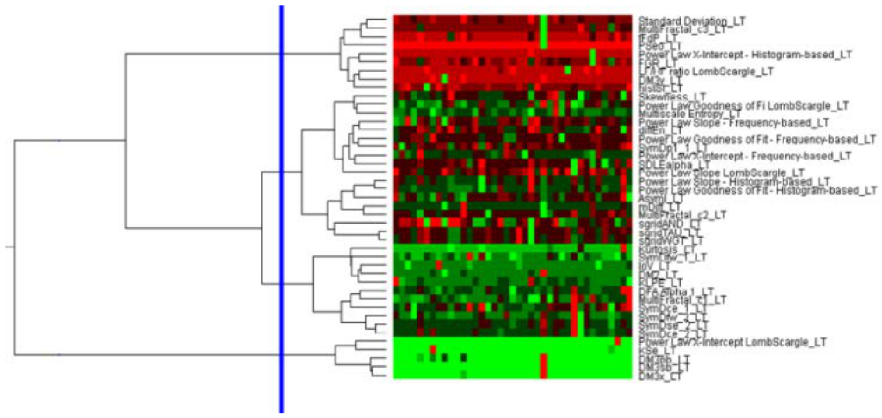


Figure 2. Hierarchical clustering of all 41 attributes for all 39 patients

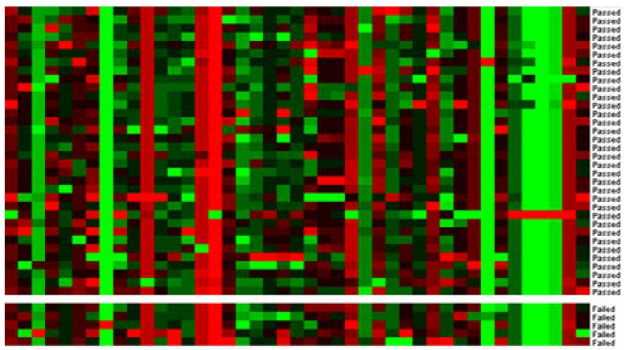


Figure 3. Heat map of all 41 attributes for all 39 patients

Figure 2 shows the hierarchical clustering of linearly transformed (-10, 10) covariance ratios for all 41 attributes derived from HRV for all 39 patients. In this case columns are patients. Here we notice that all 41 attributes are in four groups based on their intensity spectrums. This would raise the possibility of some form of a learning based on subsets identified from these four groups which would help us for better feature selection. Heat maps were generated using the BioMiner software [3].

Figure 3 shows the heat map of linearly transformed (-10, 10) covariance ratios for all 41 attributes derived from HRV, for all 39 patients. Here columns are attribute values. The white line separates the two classes (passed and failed). As we observe in

this heat map, there is no noticeable difference between the significance of attribute intensity spectrums between passed and failed patients.

Similarly, figure 4 shows the hierarchical clustering of linearly transformed (-10, 10) covariance ratios for all 43 attributes derived from RRV for all 41 patients. In this case columns are patients. Here we also notice that all 43 attributes can be observed in four groups based on their intensity spectrums. This would raise the possibility of some form of a learning based on subsets of patients identified from these four groups which would help us for better feature selection.

In addition, figure 5 shows the heat map of linearly transformed (-10, 10) covariance ratios for all 43 attributes derived from RRV, for all 41 patients. Here columns are attribute values. The white line separates the two classes (passed and failed). As we observe in this heat map, there is no noticeable difference on the significance of attribute intensity spectrums between passed and failed patients.

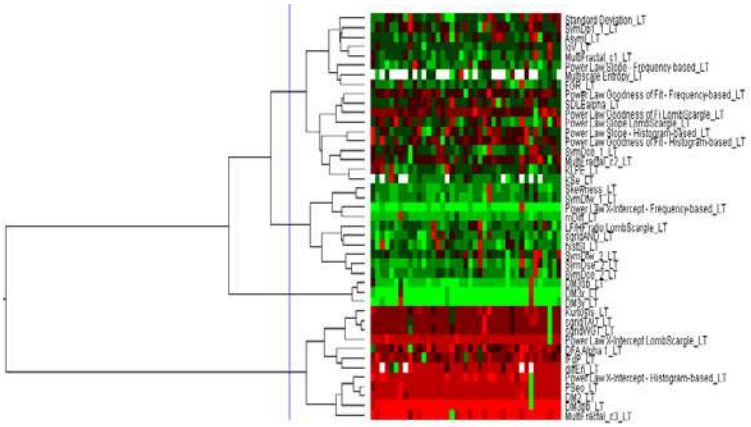


Figure 4. Hierarchical clustering of all 43 attributes for all 41 patients

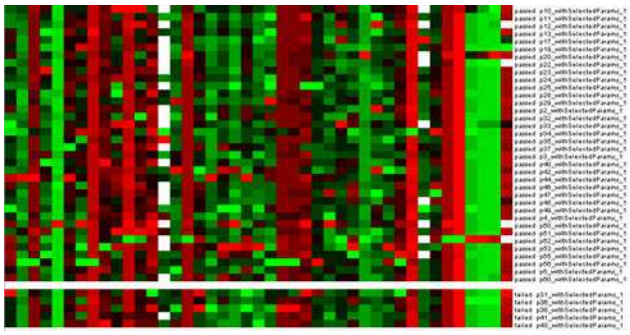


Figure 5. Heat map of all 43 attributes for all 41 patients

We are currently investigating the possibility of applying SVMs and other ML methods for additional analysis and identifying the most suitable methods to accurately separate the two classes of patients. We realize that more data would be needed.

6 Discussion and future directions

Despite its limited amount of data, this application is of high clinical interest. We are currently investigating how supervised or unsupervised learning methods could be applied to discover interesting patterns from this class of data, especially when more data becomes available. Some of the options are:

a. Using a classifier to generate models (hyperplanes) that would contain four pieces of information and could be used to classify new patients using their collected data. The four pieces of information would be the most informative derived attributes, their thresholds, the particular relations and particular confidence measures.

b. A second option would be to apply SVMs (Support Vector Machines) to this data and preferably larger amounts of it, when available. A support vector machine is a computer algorithm that learns by example to assign labels to objects (cases). A common biomedical application of support vector machines is the automatic classification of microarray gene expression profiles for a two-class application such as Leukemia (ALL vs. AML). Here in this context we can apply SVMs to a reasonable amount of data to generate a hyperplane that separates passed patients from failed ones. The hyperplane and its associated graph can be useful to identify the risk associated with future patients.

c. A third option is to consider an instance based learning approach, the most popular form of it is *k*-nearest neighbor. Unlike classification algorithms, instance based learning methods are non-parametric and memory-based. The key idea is to store all available examples (e.g. passed and failed patients) in memory and when the information for a new patient arrives, the algorithm computes the value of the new instance based on the values of the closest ones (the most similar). This is done using some standard distance measures, such as Euclidean distance. The result would be *x* number of closest neighbours along with some certainty measure that would provide some confidence.

Our contributions until now are more along the line of understanding the problem and narrowing down the path to identify the most suitable strategy for an AI application in life sciences. We are currently pursuing this research and we think that our biggest challenge would be validation of our discoveries. Another challenge is dealing with imbalanced data since the number of cases in the two classes is substantially different. Several approaches exist for handling imbalanced data for learning, once large data sets are available [8, 9, 10]. We have investigated an approach in which samples from the majority class are selected based on their inherent data characteristics and combined with the small class. And finally, being of high clinical interest, we believe in the importance of developing a decision support tool that can help physicians to better understand the risk associated with extubating patients who are in critical care.

References

1. A. Bravi, A. Longtin, and A.J.E. Seely, Review and classification of variability analysis techniques with clinical applications, *BioMedical Engineering onLine*, **10**, 90, (2011).
2. A. Famili, Z. Liu, J., Ouyang, P.R. Walker, B. Smith, M. O'Connor and A. Lenferink, 'A Novel Data Mining Technique for Gene Identification in Time-Series Gene Expression Data', *ECAI conf*, Valencia, Spain (2004).

3. A. Famili and J. Ouyang, 'Data mining: understanding data and disease modeling' *Applied Informatics*, 32-37, (2003).
4. A. Seely, and P. MacKlem, 'Complex systems and the technology of variability analysis'. Critical Care 8(6):R367-84 (2004).
5. L.N. Segal, E. Oei, B.W. Oppenheimer, R.M. Goldring, R.T. Bustami, S. Ruggiero, K.I. Berger and S.B. Fiel, 'Evolution of pattern of breathing during a spontaneous breathing trial predicts successful extubation' *Intensive Care Med*. 36(3):487-95 (2010).
6. A. Savi, C. Teixeira, J.M. Silva, L.G. Borges, P. A. Pereira, K. B. Pinto, F. Gehm, F. C. Moreira, R. Wickert, C. B. Trevisan, J. G. Maccari, R. P. Oliveira and S. R. Vieira, 'Weaning predictors do not predict extubation failure in simple-to-wean patients'. J Critical Care 27(2):221 (2012).
7. R. Sutton, Invited talk at the Constructive Induction Workshop, ICML-94 (1994).
8. X. Guo, Y. Yin, C. Dong, G. Yang. and G. Zhou: 'On the Class Imbalance Problem', In *Proc. of 4th International Conference on Natural Computation,* Jinan, October 18-20, pp. 192–201. IEEE, Los Alamitos (2008).
9. N. Japkowicz, and S. Stephen, 'The class imbalance problem: A systematic study' *Journal of Intelligent Data Analysis* 6(5), IOS Press Amsterdam, The Netherlands, (2002).
10. A. Famili, Z. Liu, and S. Phan, 'Identifying informative genes in highly imbalanced gene expression data, ECML-PKDD 11 (2011).