IDAMAP 2005

Intelligent Data Analysis in Medicine and Pharmacology

John H. Holmes and Niels Peek (Program Chairs)

A one-day workshop during the 10th European Conference on Artificial Intelligence in Medicine 2005 (AIME 05) in Aberdeen, Scotland, UK

Sunday, July 24, 2005

Co-Sponsored by

American Medical Informatics Association
Knowledge Discovery and Data Mining
Working Group

International Medical Informatics Association
Intelligent Data Analysis and Data Mining
Workgroup (WG 03)

IDAMAP 2005
Intelligent Data Analysis in Medicine and Pharmacology

John H. Holmes and Niels Peek (chairs)

A one-day workshop during the 10th European Conference on Artificial Intelligence in Medicine 2005 (AIME 05) in Aberdeen, Scotland, UK
Sunday, July 24, 2005

1. Introduction

Welcome to IDAMAP 2005! This is the tenth anniversary of IDAMAP, and as such is a special occasion for us. This year, IDAMAP is organized in collaboration with and sponsored by the Intelligent Data Analysis and Data Mining Workgroup of the International Medical Informatics Association, and the Knowledge Discovery & Data Mining Working Group of the American Medical Informatics Association.

The IDAMAP workshop series is devoted to computational methods for data analysis in medicine, biology and pharmacology that present results of analysis in a form communicable to domain experts and exploit expert knowledge of the problem domain. Methods include data mining, temporal abstraction, machine learning, and data visualization. We gather today in an informal setting, with ample opportunity to meet one another and discuss selected technical topics in an atmosphere which fosters the active exchange of ideas among researchers and practitioners. The workshop is intended to be a genuinely interactive event and not a mini-conference. Please take advantage of this unique workshop: ask questions, participate in discussions, introduce yourself to new colleagues, and enjoy the day!

2. Program

Today's program is quite full, including 13 long and four short papers, covering a wide spectrum of topics including Microarray analysis, Temporal reasoning, Prognosis, Subgroup mining, Case-based reasoning, and Visualization methods. In addition, we are most fortunate to have two exciting guest speakers. Arno Siebes from the University of Utrecht will speak on "Understanding Classifiers on Discrete Data," and Lucila Ohno-Machado from Harvard University will present "Reverse the Curse (of Dimensionality)." The full program schedule appears on the next page of the proceedings.

2. Program Committee

As program committee co-chairs, we are very grateful for the assistance of the members of the program committee. These individuals took time from their busy schedules to review a large number of submissions and assist with the final paper selection process. We thank them all for their hard work:

- Ameen Abu-Hanna, Academic Medical Center, Amsterdam, The Netherlands
- Lars Asker, Stockholm University, Sweden
- Riccardo Belazzi, University of Pavia, Italy
- Carlo Combi, University of Verona, Italy
- Janez Demsar, University of Ljubljana, Slovenia
- Michel Dojat, Universite Joseph Fourier, Grenoble, France
- Dragan Gamberger, Rudjer Boskovic Institute, Croatia
- Werner Horn, Austrian Research Institute for AI, Austria
- Jim Hunter, University of Aberdeen, UK
- Nicolette de Keizer, Academic Medical Center, Amsterdam, The Netherlands
- Elpida Keravnou-Papaeliou, University of Cyprus, Cyprus
- Matjaz Kukar, University of Ljubljana, Slovenia
- Pedro Larranaga, University of the Basque Country, San Sebastian, Spain
- Nada Lavrac, J. Stefan Institute, Slovenia
- Xiaohui Liu, Brunel University, UK
- Peter Lucas, Radboud University Nijmegen, The Netherlands
- Silvia Miksch, Vienna University of Technology, Austria
- Lucila Ohno-Machado, Harvard Medical School and M.I.T., Boston, USA
- Marco Ramoni, Harvard Medical School, Boston, USA
- Steve Rees, Aalborg University, Denmark
- Paola Sebastiani, Boston University School of Public Health
- Yuval Shahar, Ben-Gurion University of the Negev, Israel
- Stephen Swift, Brunel University, UK
- Allan Tucker, Brunel University, UK
- Frans Voorbraak Academic Medical Center, Amsterdam, The Netherlands
- Adam B Wilcox, University of Utah, USA
- Blaz Zupan, University of Ljubljana, Slovenia

3. Special acknowledgements

We would like to thank our guest speakers and authors for their contribution to the success of IDAMAP 2005. We are indebted to Silvia Miksch and Jim Hunter, Programme and Local Chairs, respectively, of AIME 05 for their support of this workshop. Finally, we are especially grateful to AMIA for their financial support and sponsorship of IDAMAP 2005.

IDAMAP 2005 Schedule

Morning Session

0900-0905	**Opening of IDAMAP2005**	
0905-0915	**10 years of IDAMAP** *Riccardo Bellazzi*	
0915-1000	**Invited Presentation** **Understanding Classifiers on Discrete Data** *Arno Siebes*	
1000-1100	**PAPER SESSION 1** **MICROARRAY ANALYSIS**	
	Signature Mining: a Heuristic Approach to Biochemical Pathway Analysis *E Panteris, S Swift, A Payne, X Liu*	Page 5
	Learning Gene Regulatory Networks with an Intelligent Data Analysis Approach: An Application to the Yeast Cell Cycle *R Bellazzi, R Amici, F Ferrazzi, P Magni, L Sacchi, S Sotgiu*	Page 11
	Mutant vs. Gene Expression Profiles for Function Prediction *T Curk, U Petrovic, G Shaulsky, B Zupan*	Page 15
1100-1130	**Break**	
1130-1255	**PAPER SESSION 2** **TEMPORAL REASONING AND PROGNOSIS**	
	Instance-based Prognosis in Intensive Care Using Severity-of-illness Scores *C Tan, L Peelen, N Peek*	Page 21
	The Predictive Value of Consecutive Event Episodes in the Intensive Care *T Toma, A Abu-Hanna, R-J Bosman*	Page 27
	A Dynamic Bayesian Network for Diagnosing Ventilator-Associated Pneumonia in ICU Patients *T Charitos, LC van der Gaag, S Visscher, K Schurink, P Lucas*	Page 32
	A Probabilistic Method for Multiple-Patient Temporal Abstraction *M Ramati, Y Shahar*	Page 38
	Short Term Blood Glucose Measurements may be severely Biased *J Randløv, J Kildegård*	Page 44
1255-1425	**Lunch**	

IDAMAP 2005 Schedule

Afternoon Session

1425-1510	**Invited Presentation** **Reverse the Curse (of Dimensionality)** *Lucila Ohno-Machado*	
1510-1600	**PAPER SESSION 3** **SUBGROUP MINING, RECORD LINKAGE, AND CASE-BASED REASONING**	
	Profiling Examiners using Intelligent Subgroup Mining *M Atzmueller, F Puppe, H-P Buscher*	Page 46
	Data Analysis Based on Subgroup Discovery: Experiments in Brain Ischaemia Domain *D Gamberger, A Krstacic, G Krstacic, N Lavrac, M Sebag*	Page 52
	Probabilistic Medical Record Linkage Incorporating Close Agreement *M Tromp, N Méray, ACJ Ravelli, JB Reitsma, GJ Bonsel*	Page 57
	Diagnosis of Dysmorphic Syndromes Using Prototypes and Adaptation Rules *T Waligora, R Schmidt*	Page 59
1600-1620	**Break**	
1620-1745	**PAPER SESSION 4** **VISUALIZATION METHODS, INFORMATION RETRIEVAL, AND DECISION MAKING**	
	FreeViz - An Intelligent Visualization Approach for Class-Labeled Multidimensional Data Sets *J Demsar, G Leban, B Zupan*	Page 61
	Gravi++: Interactive Information Visualization of Highly Structured Temporal Data *K Hinum, S Miksch, W Aigner, S Ohmann, Ch Popow, M Pohl, M Rester*	Page 67
	Intelligent Interface for Adjuvant Treatment Planning in Breast Cancer *IH Jarman, TA Etchells, PJG Lisboa*	Page 73
	Multi-Classification of Clinical Guidelines in Concept Hierarchies *D Sona, P Avesani, R Moskovitch*	Page 78
	Influence Diagrams for Medical Decision Problems: Some Limitations and Proposed Solutions *M Luque, FJ Diez, C Disdier*	Page 85
1745	**Closing**	

Signature Mining: a Heuristic Approach to Biochemical Pathway Analysis

Eleftherios Panteris*, Stephen Swift, Annette Payne, Xiaohui Liu

School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, Middlesex UB 8 3PH, UK
*Contact e-mail Eleftherios.Panteris@brunel.ac.uk
http://www.ida-research.net

Abstract

Microarrays have revolutionised biology, and bioinformatics is now a powerful tool in the hands of biologists. Gene expression analysis has attracted a large amount of attention over the last few years mostly in the form of algorithmic explorations of cluster relationships, and software that try to display the multidimensionality of microarray data in biological relevant formats. In this paper we propose a simple yet effective approach to biochemical pathway analysis based on biological knowledge, to select a subset of genes for each pathway that fully describes the behaviour of the pathway at a given experimental condition in a bid to reduce the dimensionality of microarray data and make the analysis more biologically relevant.

1 Introduction

A new field attempting to describe biology called systems biology is currently emerging trying to depict biology at an organisation level by multidisciplinary research [Aggawal *et al.*, 2003]. Microarrays are an essential member of this multidisciplinary approach and a lot of interest has focused on gene expression analysis. Informatics and computer science are important members of this field with a heavy interest in microarray data analysis and data storage, as well as in distribution and display of data in terms of clustering programs and large databases. Network modelling is also very active trying to describe biochemical pathways and biological processes in general [Huang, 2004].

These multidisciplinary approaches aspire to combine and produce practical descriptive models of biological systems that can be used among others to predict drug response and aid in cancer prevention and treatment.

Analysis of microarray gene expression [Eisen *et al.*, 1998] has attracted a lot of attention over the years mostly the form of algorithmic explorations of cluster relationships, and software that try to display the multidimensionality of microarray data in biological relevant formats [Slonim, 2002]. The multi-dimensionality of the microarray experimental data has made this into a daunting task and there still a lot to be desired from the current work [Claverie, 1999]. Meanwhile, the modelling community has a growing interest in the complexity of biochemical pathways and various modelling methods exist that try to predict how such pathways behave [Papin *et al.*, 2003].

We have taken a simple yet effective approach to pathway analysis using the idea of *signatures* for each pathway. An algorithm based on hill climbing was used to mine for the *signatures* in all the 108 pathways from *E.coli*. The algorithm is effective in finding biologically relevant *signatures* and the results are promising that this is a valid way forward in the field.

The background behind this study is in Section 2, giving the reasons why we used a novel interpretation of biochemical pathways for our problem. Section 3 gives information about the data and their sources, and section 4 describes the *signature* mining process and its algorithm. Section 5 deals with the biological verification of the results. Section 6 summarises the findings and proposes future directions of work.

2 Background

The identification and validation of drug targets depends critically on knowledge of the biochemical pathways in which potential target molecules operate within cells. For this reason, the study of biochemical pathways is the focus of numerous drug discovery researchers and is central to the strategy of many biopharmaceutical and genomic companies.

There is intense research going on in systems biology, with scientists using different methods to solve similar problems [Aggawal *et al.*, 2003]. From the biological point of view, most scientists use methods that offer some but not all of the functionality a biologist would like to have, often with rather complex and time consuming implementations. If pathway analysis and visualisation is going to be performed by biologists alone, it should be done in a straightforward and with few intermediate steps way so scientists can focus on the biological significance of the findings and not the programming implementation of the methods. So far this is not currently available in the research community, and the software available, both public and commercial, do not provide all the functionality they should. [Goesmann *et al.*, 2002; Toyoda *et al.*, 2003; Dahlquist *et al.*, 2002; Kolpakov *et al.*, 1998]

These constraints infiltrate the relationship biologists have with other sciences and computer science in particular. Ma and Zeng have shown in their paper [2003] that modelling biochemical pathways is not straightforward, and mistakes can be made if all parameters are not taken into account.

Utilising biological knowledge about biochemical pathways and their components, this study produces a practical picture of the behaviour of the whole genome of an organism based on microarray data and pathway data from major databases like KEGG [Kanenisha and Goto, 2000]. By collecting numerous experiments from a given organism, *E.coli* was used, for distinct environmental conditions and treatments and then combining it with well-established pathway information about genes and their biological contribution, we choose a sub set of genes from each pathway, a '*signature*', which is used to describe the behaviour of that pathway under the given condition.

A pathway's *signature* is a unique set of genes that can be monitored in any given microarray experiment to illustrate that pathway's behaviour. The *signature* is the collection of the 'true' *expression indicators* from the pathway. They are the most 'expressively active' genes, in the sense that they are the more sensitive part of the pathway, the ones most responsive to external stimuli, i.e., the change in the environmental conditions affects them in such a way as to alter their expression in the cell. The rest of the genes in that pathway are transcriptionally dormant in the sense that they do not respond readily to change, since they form the infrastructure of the pathway in the cell, and as building block they are not sensitive to external stimuli as much.

Pathway analysis methods of expression data currently in use, which include all the recent clustering techniques, require *all* the genes of a pathway to be taken into account, and may lead to the erroneous conclusion that the activity of a pathway has remain unchanged. For example, if more genes in a pathway are transcriptionally dormant than transcriptionally active, the more numerous dormant ones mask the true picture of a change in the activity of that pathway.

By monitoring the *signature* of a pathway in all subsequent microarray experimental data we would have an immediate description of the behaviour of the pathway and subsequently of the whole organism in a *global pathway /signature network*. In essence, we aim to reduce the dimensionality of cDNA microarray data to provide a biologically relevant picture of the whole organism immediately, before resorting to clustering methods.

Our emphasis lies on using pathway knowledge to group all the scattered genes in a microarray dataset as pathways and observe the pathway's behaviour as a whole, rather than genes individually. It is a different concept that aims to help biologists in pathway analysis, by representing microarray data in a pathway-orientated view, with genes grouped not only by expression similarity but also biologically.

Furthermore, it offers a simplified view of these pathways by using a specific subset of genes to depict the behaviour in each experiment. This offers new options to biologists who could group or 'cluster' the pathways according to behaviour in an experiment thus, finding interesting connections, not easily observed in gene clustering techniques and visualisations.

3 Data

Gene **E**xpression **O**mnibus data repository at NCBI was the source of the microarray datasets. They come from *E.coli* and represent three different experimental conditions in 51 experiments in total. We exploited the variety of conditions to find the most sensitive genes under these conditions, since the larger the number of experimental conditions and number of experiments, the more fine-tuned the dataset is. There are global cDNA microarray experiments containing the majority of the *E.coli* genes. The experimental data, representing 51 microarray experiments, were normalised to Standard Deviation of 1 and Mean of 0 so that they can be compared together. No further normalisation was necessary since the data were already normalised to log ratios when they were released in **GEO**.

The genes are chosen according to their variability in expression and have to be above a certain empirically defined global threshold, as used in microarray analysis to be considered as statistically significant. The threshold is empirically selected depending on the dataset used and is considered for each time point independently and the selection process is repeated for every experiment. The threshold is the statistically significant fold difference between the two copies of the gene in the control and test conditions. Its range is usually between 1 and 2 fold and the researcher chooses an appropriate value depending on the general levels of expression of all the genes in the experiment. [Schena *et al.*, 1996; Dunggan *et al.*, 1999]
The KEGG *E.coli* files were taken from the KEGG portal [Kanenisha and Goto, 2000]. By combing the two, a list of important genes was assembled and these were used as the base of the algorithm.

4 Algorithm

Choosing the best selection of genes in each pathway that represent that pathway's behaviour is challenging because each gene can be a member of several pathways and we needed to find a way to choose genes that represent each pathway out of the 108 of *E.coli*. Basically we tried to find a way to move genes from one pathway to another based on their similarity of expression for the whole of the 51 experiments not just one experiment. We opted for an algorithm with a hill climbing [Michalewicz *et al.*, 1998] step described below.

Let G be the set of n genes, $G = \{1,\ldots,n\}$, let $X \in \Re^{n \times T}$ be the n by T gene expression matrix for the n genes where the ith row of X, x_i, is the gene expression

profile for gene *i*. x_{ij} is defined as the *j*th element of the vector x_i. Let the pathway list *P* be a list of $m>0$ lists where $p_i \subseteq G$ is the *i*th element of *P*, where $|p_i| > 0$. A *signature* s_i of a pathway p_i is defined as $s_i \subseteq p_i$ where $|s_i| > 0$. The list of *signature*s is denoted as *S*, where $|S| = m$. s_{ij} is defined as the *j*th element of the list s_i, such that s_i is a subset of the corresponding p_i. How close two expression profiles *a* and *b* are, is given by the Euclidean Distance formally defined in formula 1.

$$d(a,b) = \sqrt{\sum_{i=1}^{T} (x_{ai} - x_{bi})^2} \quad (1)$$

$$D \in \Re^{n \times n}, \text{ where } D_{ij} = d(i,j) \quad (2)$$

The *n* by *n* symmetric matrix *D* contains all of the pair-wise similarities between genes. Note that the larger *d* (*a*, *b*) is, the more dissimilar the genes *a* and *b* are. How close together the genes within a *signature* are is defined as follows:

$$FS(s_i) = \sum_{a=1}^{|s_i|-1} \sum_{b=a+1}^{|s_i|} d(s_{ia}, s_{ib}) \quad (3)$$

This is the sum of all pair-wise differences between the elements of a *signature*. Equation 4 represents how well fitted the *signatures* are, and equation 5 represents how many genes have been allocated from each pathway. To 'mine' the *signatures* for each pathway we need to find a set *S* where F_1 is minimised and F_2 is maximised:

$$F_1 = \sum_{i=1}^{m} FS(s_i) \quad (4)$$

$$F_2 = \sum_{i=1}^{m} |s_i| \quad (5)$$

$$F_3 = \frac{F_1}{F_2} \quad (6)$$

The algorithm fitness F3 is represented in equation 6 and needs to be minimised for the optimum solution, i.e. the smallest *signature* possible with the genes best describing the pathway.

The signature mining algorithm takes as input a Euclidean distance comparison matrix of all the genes from all the pathways, and a pathway list of lists from KEGG of all the pathways and their genes. To mine the appropriate genes for each signature, we decided to randomly remove or replace a gene from a pathway and use a hill climbing technique to evaluate the solution. The evaluation is based on a similarity and a size function, requiring minimisation of their fraction to progress. The algorithm is described below (algorithm1).

Algorithm 1 *Signature* Mining:

```
(1) INPUT:
    Euclidean Distance Matrix (D)(eq.
    2),Filtered Pathway list of lists
    (PA)(108 pathways long).
(2) ITERATIONS
    For w=1:ITER do
(3) RANDOM SELECTION with REMOVAL or
    REPLACEMENT
    Randomly chose a pathway and randomly
    chose a gene position from the
    pathway.
    If gene is present:
       REMOVE
    Else
       REPLACE gene back to position.
    End
(4)     GET EUCLIDIAN DISTANCES FROM ALL
        PATHWAYS
        For i=1:length of PA do
            OBTAIN all unique distances
            between the pathway genes
            from D for comparison.
        End
(5) GET F1 (SIMILARITY FUNCTION)(eq. 4)
    For i=1:length of PA do
        F1(i)= SUM(distances of all
        genes from P(i))
    End
       Store F1(i)
(6)     GET F2 (SIZE FUNCTION)(eq. 5)
        For i=1:length of PA do
            F2(i)= length of P(i)
        End
            Store F2(i)
(7) GET F3(EVALUATION FUNCTION)(eq. 6)
    For i=1:length of PA do
        F3(i)=SUM(F1(i)/F2(i))
        F3new(w)=F3new(w)+F3(i)
    End
(8)      EVALUATION
         If F3new(w) < F3old
            SET as F3old
         Elseif F3new(w) > F3old
               RESET to previous value
         End
    End
(9) OUTPUT:Signature list for all (108)
    pathways
```

The convergence of the algorithm is shown below (Fig.1), the number of iterations being 20000. The convergence graph shows that the algorithm performs well, by sharply dropping for the first 4000 iterations and then slowly stabilising to the minimum evaluation value possi-

ble, from the 6000 iteration onwards the slope levels up to almost a straight line.

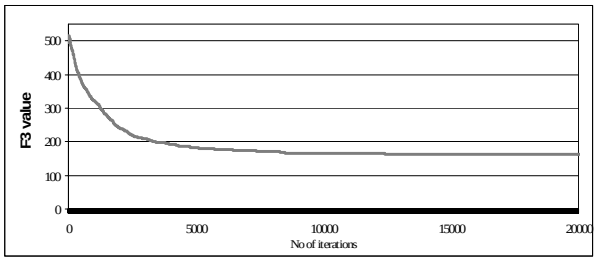


Fig. 1. Convergence plot of the Algorithm. The y axis F3 value refers to the evaluation function F3 (equation 6).

The performance of the algorithm accordingly can be seen from the histograms that represent the distribution of genes per pathway (Fig.2), both before (a) and after (b) the application of the algorithm. By keeping only the *signature* specific genes after using the algorithm, each pathway is reduced to at least 1/3 of its previous size, making it more specific, allowing for easier biological interpretation of the pathway behaviour.

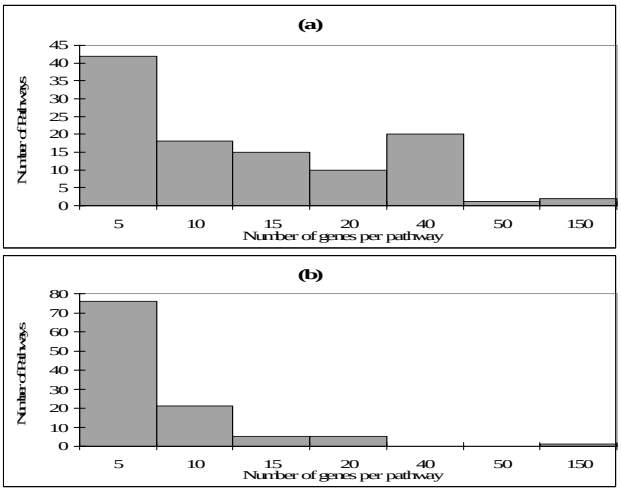


Fig. 2. The distribution of genes per pathway before (a) and after (b) the application of the signature algorithm, with number of pathways and number of genes per pathway.

5 Application

In order to biologically validate our method we chose a pathway from *E.coli*. The Phenylalanine, Tyrosine and Tryptophan biosynthesis pathway, as defined in the KEGG pathway database, was chosen with focus on Tryptophan. The Tryptophan production is regulated from a specific operon that contains five genes, B1260, B1261, B1262, B1263, and B1264.

A microarray experiment of *E.coli* under Tryptophan starvation from Khodursky et al [2000] was used. They observed a very specific response from the Tryptophan operon genes. These genes are activated in the absence of Tryptophan and induce its production. So by starving the organism in their experiments, they monitored the activation of the pathway.

Using the signature mining algorithm we 'mined' a signature for the specific pathway that portrays the behaviour of the pathway according to Khodursky et al [2000]. The importance of the signature lays in the fact that we used the GEO dataset, to find the signature that describes the pathway in the Khodursky et al [2000] dataset.

The Phenylalanine, Tyrosine and Tryptophan biosynthesis pathway includes genes from the biosynthesis of these three amino acids. They are grouped together in the KEGG database due to the chemical similarity these amino acids have.

The pathway contains 26 genes and the signature mining algorithm produced a pathway signature of 6 genes. The 6 genes are from all the three branches of the specific pathway as shown in table 1. Genes B0928 and B2021 are present in both the phenylalanine and tyrosine processes.

Table 1. Distribution of signature genes per branch of the Phenylalanine, tyrosine and tryptophan pathway.

PHENYLALANINE	TYROSINE	TRYPTOPHAN
B1713	B4054	B1260
B0928	*B0928*	B1262
B2021	*B2021*	

Khodursky et al [2000] are interested only in Tryptophan starvation so their dataset contains only the genes B1260 and B1262. The starvation response of *E.coli* is to activate the genes that produce Tryptophan [Khodursky et al., 2000]. The response can be observed in Fig. 3 where the gene expression of the genes that constitute the tryptophan operon is plotted in a six part starvation time course. The organism is placed in an environment without tryptophan at the start of the experiments (see Fig.3) and gene expression measurements are taken at 20 minutes intervals. It is obvious from the graph (Fig 3) that the genes are highly up-regulated moments after the starvation initiation.

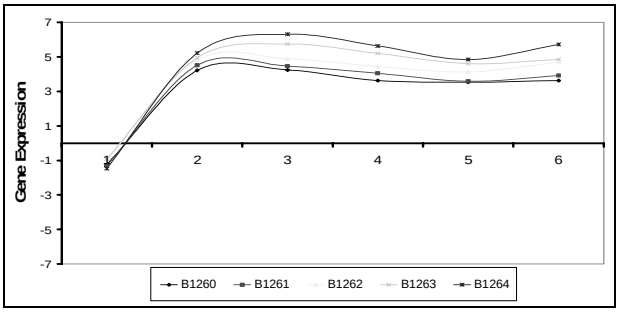


Fig. 3. The Tryptophan operon activation from the Khodursky et al [2000] dataset. See text for further details.

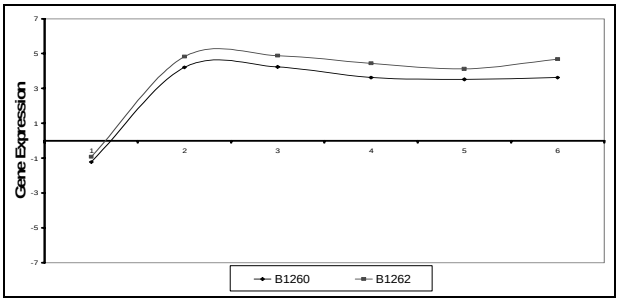


Fig. 4. The Signature genes present in the dataset. The genes describe the activation of the operon during Tryptophan starvation.

Our signature has two out of five genes from the tryptophan operon. As it can be seen from Fig. 4 they are sufficient to portray the behaviour of the pathway during the experiments. As mentioned above briefly, the signature was 'mined' from the **GEO** dataset and applied to the Khodursky *et al* dataset [2000].

This has provided early evidence that signature mining can be an effective way of analysing biochemical pathways because once they are chosen they can be applied across experimental conditions and datasets with ease.

The biological relevance of the *signature* mining algorithm is extensive, especially in the biochemical and pharmaceutical community, since it allows the biologist to observe the behaviour of a specific pathway in a clear and definitive way that does not involve genes that do not affect the pathway's regulation. Its relevance is evident in drug related research. *Signature* mining could help by monitoring the effect of the drug on the whole of the organism by reducing the complexity of the pathways and showing a holistic view of the organism. An up or down regulated behaviour could be easily identified in the signature context and that pathway chosen for further investigation to find the specific genes affected. In essence, it allows the researcher to have an overview of the entire organism processes in an simple and obvious way, easy to understand and use.

6 Conclusions

To conclude, we have shown that a specially selected sub group of genes from a biochemical pathway, a *signature*, is able to depict its behaviour under a given experimental condition. A simple yet effective algorithm based on hill climbing was created to mine for the appropriate genes for each pathway based on the pathway's behaviour across a large set of experiments of varying conditions. The algorithm was able to select *signatures* for all 108 pathways, of which one was used as an example here.

Using biological knowledge in the design of the algorithm was very important as it was the biological verification of the results. The preliminary results have clearly shown this interpretation of biochemical pathways to be an interesting way of describing microarray data used for pathway analysis.

Future work will improve the algorithm run time and functionality. The algorithm needs to be more selective and not prone to be trapped in local optima. Ideally the algorithm should have picked all five genes that form the operon, this maybe due to the common problem of hill climbing algorithms of being easily trapped in local optima and it is currently being addressed using a global search method such as simulated annealing [Kirkpatrick *et al.*, 1983].

Additionally, the algorithm will be a part of a framework for microarray datasets for full exploitation of microarray data in relation to pathway analysis and pharmaceutical research. Application of the algorithm will not be restricted only to *E.coli* but to other organisms with specific pharmaceutical concerns and in due course to human data, with a continuation of the framework steps to include gene networks and interactions with protein-protein networks, offering a solid solution in that area of systems biology.

Acknowledgements

The authors would like to thank the reviewers for their useful insights and comments about the manuscript.

References

[Aggawal *et al.*, 2003] Aggawal, K., Lee, K.H.,; Functional genomics and proteomics as a foundation for systems biology.' Briefings in functional genomics and proteomics Vol 2, No 3, 175-184 2003

[Claverie, 1999] Claverie, J.; Computational methods for the identification of differential and coordinated gene expression. Human Molecular Genetics. Vol 8, No 10 1821-1832 1999

[Dahlquist *et al.*, 2002] Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C., Conklin, B.R.,; GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways, Nature Genetics 31(1):19-20 2002

[Duggan *et al.*, 1999] Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J.; Expression profiling using cDNA microarrays. Nature Genetics. 21, 10-14 1999

[Eisen *et al.*, 1998] Eisen, M., Spellman, P.T., Botstein, D., Brown, P.O,; Clustering Analysis and display of genome-wide expression patterns. PNAS 95. 14863-14868 1998

[Goesmann *et al.*, 2002] Goesmann, A., Haubrock, M., Meyer, F., Kalinowski, J., Giegerich, R.; PathFinder: reconstruction and dynamic visualization of metabolic pathways. Bioinformatics. 18: 124-129 2002

[Huang, 2004] Huang, S.; Back to the biology in systems biology: What can we learn from biomolecular net-

works? Briefings in functional genomics and proteomics. Vol 2, No 4, 279-297 2004

[Kanehisa and Goto, 2000] Kanehisa, M., Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 28, 27-30 2000

[Kirkpatrick *et al.*, 1983] Kirkpatrick, S., Gelatt, Jr., C.D., Vecchi, M.P.; Optimization by Simulated Annealing, Science. 220, No. 4598, 671-680 1983

[Khodursky A., *et al* 2000] Khodursky AB, Peter BJ, Cozzarelli NR, Botstein D, Brown PO, Yanofsky C. DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *E.coli***.** PNAS. USA. Oct 24;97(22):12170-5 2000

[Kolpakov *et al.*, 1998] Kolpakov, F.A., Ananko, E.A., Kolesov, G.B., Kolchanov, N.A.; GeneNet: a gene network database and its automated visualization. Bioinformatics 14: 529-537 1998

[Ma and Zeng, 2003] Ma, H., Zeng, A.,; Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics 19: 270-277 2003

[Michalewicz *et al.*, 1998] Michalewicz, Z., Fogel, D. B.; How To Solve It: Modern Heuristics, Springer, 1998.

[Papin *et al.*, 2003] Papin, J., Price, N., Wiback, S., Fell, D., Palsson, B.; "Metabolic pathways in the post-genome era," Trends in Biochemical Sciences., Vol. 28, 250-258 2003

[Schena *et al.*, 1996] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., Davies, R.W.; Parallel Human Genome Analysis: Microarray-Based Expression Monitoring of 1,000 Genes. PNAS 93. 10614-10619 1996

[Slonim, 2002] Slonim, D. K.; From patterns to pathways: gene expression data analysis comes of age. Nature Genetics. 32, 502 - 508 2002

[Toyoda *et al.*, 2003] Toyoda, T., Mochizuki, Y., Konagaya, A.,; "GSCope: a clipped fisheye viewer effective for highly complicated biomolecular network graphs," Bioinformatics, Vol. 19, 437-438 2003

Learning gene regulatory networks with an Intelligent data analysis approach: an application to the yeast cell cycle

Riccardo Bellazzi, Roberta Amici, Fulvia Ferrazzi, Paolo Magni, Lucia Sacchi, Stefano Sotgiu
Dipartimento di Informatica e Sistemistica, Università di Pavia
Via Ferrata 1, 27100 Pavia, Italy
riccardo.bellazzi@unipv.it

Abstract

This paper presents a novel approach for the extraction of gene regulatory networks from DNA microarray data. The method is applied to the reconstruction of a network of interactions of genes involved into the cell cycle of Saccharomyces Cerevisiae. The approach is characterized by the integration of data coming from different experiments together with the knowledge available on the biological process under analysis and on the dynamics of the process itself. The method is capable to reconstruct known relationships among genes and to provide meaningful biological results.

1 Introduction

A noteworthy research effort in Biomedical informatics has been recently devoted to the development of methods for the automated extraction of gene regulatory networks from DNA microarrays data. Such interest is motivated by the capability of DNA microarrays to describe cell molecular processes at the whole genome level. The availability of experiments in which a certain cell condition is followed over time gives the chance to learn dynamic models of gene to gene interactions. Several algorithms have been implemented so far: a pioneering work is represented by the REVEAL approach, which extracts networks expressing Boolean relationships between genes through a heuristic search strategy based on mutual information [Liang *et al.*, 1998]. More recently, other methods have been presented to derive regulatory networks from microarray data, including methods based on differential equations [De Jong, 2002] and dynamic probabilistic networks [Perrin *et al.*, 2003]. All those methods have pros and cons; however, given the very nature of the data, none of the approaches may lead to reveal all the biochemical pathways underlying the observed processes. As a matter of fact, a certain mRNA stream does not always correspond to the same protein, due to potential modifications after transcription and after translation; even more importantly, the dynamics of biochemical reactions cannot be captured by the (low) sampling time available in DNA microarray experiments. For these reasons, it is of interest to integrate data coming from different sources, multiple experiments and the available background knowledge to derive models which should be able to describe as close as possible regulatory interactions occurring between genes. In this paper we present a novel approach to derive a network of potential interactions of genes involved in the yeast cell cycle. The approach integrates data coming from two different experiments and the knowledge available on the biological process and on the dynamics of cell cycle.

2 Modeling gene networks

Following the approach proposed by Schlitt and Brazma [Schlitt, 2005], it is possible to model gene networks at different levels of detail. As a consequence, four basic classes of models can be distinguished: a) *Parts lists*, referring to the collection and systematization of the network components; b) *Topology models*, describing the interactions between the parts; c) *Control logic models*, describing the effect of regulatory signals; d) *Dynamic models*, modeling the dynamics of gene interactions.

The so-called *part list* is often directly extracted from knowledge available in Gene Ontology (Gene Ontology™ Consortium, http://www.geneontology.org). Such information allows to select only the genes which are known to be involved in the process which is under study. However, other secondary bioinformatics databases can be conveniently exploited, such as the *Gene* database, maintained at NCBI (http://www.ncbi.nlm.nih.gov).

The gene-gene interaction *network topology* is learned from data. In this case, it is crucial to assign a meaning to the network connections. In the literature, a first interpretation is that, given two genes G1 and G2 represented in the network as nodes, G1 is directly linked to G2 only if G1 is a transcription factor for G2. In this case the link describes a physical interaction between the two genes. A second interpretation is that an edge between G1 and G2 means a generic "cause-effect" relationship, such that a change in the expression of G1 causes a change in the expression of G2. In this case we are describing a phenomenological event, regardless of the physical interactions between the two genes. Rather interestingly, in some model organisms, such as Saccharomyces Cerevisiae (baker's yeast), it is now possible to learn from data both kind of networks.

An important data set on the interactions between the genes and their transcription factors has been collected by Lee et al [Lee et al 2002] in the so-called ChIP-on-chip experiments. Such data have been used to derive the topology of a network of physical interactions.
On the other side, Hughes et al. [Hughes *et al.*, 2000] performed a complex experiment to detect the effects of a single gene mutation. Given a DNA microarray experiment on a mutant, corresponding to a single knocked-out gene, a significant change of the expression level in any of the non-mutated genes with respect to the wild-type case is supposed to highlight a relationship with the knocked out gene.
As mentioned in the introduction, a large number of control models have been studied in the literature, starting from Boolean relationships and moving towards probabilistic ones [Liang *et al.*, 1998; De Jong, 2002, Perrin *et al.*, 2003]. All those models can be considered also dynamic models, although the emphasis is not given to the description of the biochemical reactions, but rather to the phenomenological relationships between the problem variables, i.e. the genes. Such models are often derived from "dynamic data", i.e. time series of gene expression profiles usually collected with experiments carried on in cell cultures [Spellman *et al.*, 1998].
A consistent literature is also available on the quantitative modeling of the biochemical networks. For what concerns yeast, for example, several papers appeared on the cell cycle dynamics [Sveiczer *et al.*, 2004]. It is important to notice that such models are designed for simulation purposes, and aim at describing at a "physical" level the gene product interactions. Since they must model also fast reactions, they are typically not identifiable from data, but they require knowledge on the stoichiometric coefficients of each single biochemical reaction.
In our case, we are interested in providing a description of the interactions of the genes involved in the cell cycle of Saccharomyces Cerevisiae, taking into account all the four levels mentioned above: we will propose a network model based on different data sources and on the knowledge available in the knowledge repositories (parts lists), which relies on a network topology derived from data (topology modeling), and which models the dynamics of control interactions between genes (control logic and dynamic models).

3 The proposed approach

In this paper we propose a method to infer gene to gene interaction networks in Saccaromyces Cerevisiae cell cycle. The basic steps of the method are described in Figure 1; they can be summarized as follows: 1) learning of an initial network topology from mutant data; 2) selection of the genes involved in the cell cycle; 3) filtering of the selected genes on the basis of the available data on the cell cycle dynamics; 4) learning the final interaction network and a dynamic model of control with a genetic algorithm search.

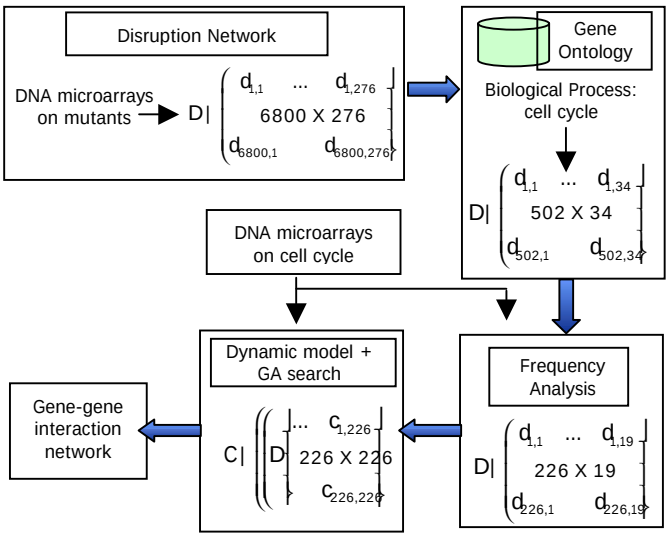


Figure 1. The proposed method

3.1 Learning the initial network topology from mutant data

This step is based on the analysis of the data made available by Hughes et al. [Hughes *et al.*, 2000], already introduced in Section 2. They collected the data of about 300 experiments in which a single gene has been knocked-out and the RNA abundance of all the other genes (about 6800) has been measured through c-DNA microarrays. The goal of this study was the detection of the functional modules of each mutated gene. Starting from the mutants experiments, it is possible to derive a first network of gene interactions: this network can be easily represented with a connection matrix D with elements D_{ij} which express the relationships between gene i and gene j; if $D_{ij}=1$ the connection is present, if $D_{ij}=0$ the connection is absent.
After the analysis of the Huges data, we obtained a matrix of 6800 x 276 elements, where each column corresponds to an experiment with a single mutated gene, while each row corresponds to a certain gene. The semantic of the network can be augmented with the sign of the relationship (enhancement or inhibition).

3.2 Gene ontology and dynamic networks filtering

The dimension of the matrix D can be conveniently reduced by resorting to the knowledge available in Gene Ontology. In our case we selected only the genes involved in the cell cycle biological process, thus reducing the matrix D to 502 x 34.
Since our main goal was to learn a dynamical model of the control of genes involved in the cell cycle, we then resorted to the "dynamic" data sets available in the literature. In the case of yeast cell cycle, the reference data are the ones coming from a well-known experiment from Spellman [Spellman *et al.*, 1998]. In this case the mRNA data have been collected in 18 different time points (one each 7 minutes). Since the cell cycle for the yeast under the experimental conditions lasts 66 minutes, it is possible to observe almost two complete mitotic cycles.

The knowledge on the dynamics of the cell cycle period, together with the information on the sampling time, limits the scope of the investigation to search for relationships which can be reasonably detected in the available data.
In particular, given the sampling time, we cannot detect signal with frequency components higher than (1/(2*7) min^{-1}). For this reason, we have filtered out the gene profiles with energy content located in high frequencies, with a cut-off frequency of 0.05 (1/20) min^{-1}. Such a choice is able to preserve the cell cycle frequency and its first harmonic component. In this way, the matrix D dimension has been then further reduced to 226 x 19.

3.3 Learning dynamic models

Starting from the connection matrix obtained after filtering, we implemented a novel algorithm to select the final model of the gene network interactions. Such step needs two ingredients: a) the choice of a dynamic mathematical model able to describe the available data; b) a strategy to search for potential relationships in the unexplored portion of the connection links (a matrix D' 226 x 207). In order to accomplish with this goal, we have exploited discrete time dynamic linear models and a Genetic Algorithm (GA) search.
Dynamic linear models have been selected, since they are the simplest class of models which allows periodic or damped oscillation behaviors.
The dynamics of the mRNA ratio[1] (x) of the i-th gene is therefore described as:

$$x_i(k+1) = a_{ii}x_i(k) + \sum_{j=1, i\neq j} a_{ij}c_{ij}x_j(k)$$

where the a_{ij}s are connection weights and the matrix $C=|D\ D'|^{226x226}$ is the connection matrix obtained by concatenating the known matrix D^{206x19} and the unknown matrix $D^{206x207}$ that has to be learned from the data. Given a certain network topology, i.e. a matrix C, we can easily learn the parameters a_{ij} from the available data through least square fitting. Different models, i.e. different C matrixes, can be compared, and hence selected by applying a model selection score. In our case we exploited the Akaike Information Criterion score (AIC).
The space of all possible models (i.e. possible connections) is super exponential; therefore it has been searched through a Genetic Algorithm strategy, with a fitness function given by the AIC score. In particular, the Genetic algorithm has been implemented by selecting 20 "individuals" (i.e. initial samples for the matrix C) which have evolved for 400 generations with the following parameters: cross-over probability = 0.9, mutation probability = 0.1, and probability of selecting the i-th individual (i.e. a certain matrix C) which is proportional to the fitness. Convergence of the solution has been visually inspected.

[1] The available measurements are coming from cDNA experiments. Therefore the problem variables are expressed as ratios of mRNA with respect to the basal condition (time =0)

3 Results

Interesting results have been obtained in all phases of the learning process. To evaluate such results, we considered 22 genes whose role in the cell cycle is well characterized and we investigated the capability of our method of reconstructing the known relationships on the basis of the available data.

We first exploited the data coming from the Huges disruption experiment, in which only 6 of those 22 genes have been mutated. We thus inferred a network (shown in Figure 2a) in which some connections appear to be supported by the information available in the literature (e.g. some links involve a gene and its transcription factor). This network was extended following the strategy proposed in this paper: in the final graph obtained (shown in Figure 2b) a significant number of the inferred connections between the 22 cell cycle genes reflects the knowledge available in the literature about the gene to gene interactions. In particular, the network shows the following interesting relationships:
a) *Mcm1* interacts with *Clb1*: the genes that normally exhibit a G2-to-M-phase-specific expression pattern, such as *Clb1*, are not induced in the absence of functional *Mcm1;* moreover, it was demonstrated that *Clb1* transcript levels are substantially reduced when functional *Mcm1* is absent. b) the *Clb5*-*Clb1* and *Clb2*-*Clb1* links express complex (indirect) interactions between cyclins, the proteins which regulate the overall cell cycle (see http://mips.gsf.de/genre/proj/yeast/). c) *Far1* is a cyclin-dependent kinase inhibitor, and it is therefore activated by the cyclin levels, such as *Clb1*.

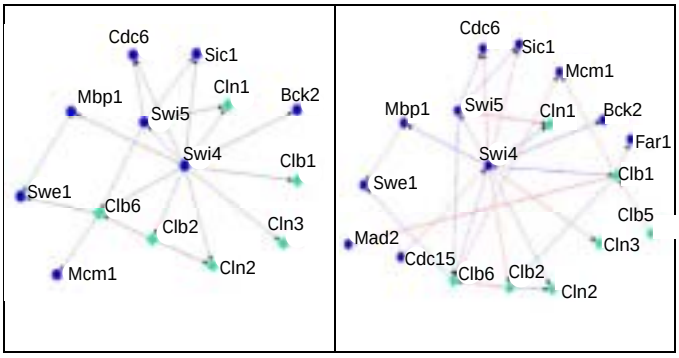


Figure 2. Graph connectivity of some of the 22 well-characterized cell cycle genes: a) initial disruption network, b) final network obtained exploiting background knowledge and dynamic data

Examining the overall derived network, we observed a scale-free connectivity: about 170 genes out of 226 are linked with no more than 5 genes, while only 10 genes are connected with more than 40 other genes. Such latter genes are the hubs of the final gene interaction network. Some of the hubs are: *Swi4*, the DNA binding component of SBF transcription factor; the two B-type cyclins *Clb1* and *Clb2*, activators of Cdc28 at G2/M phase of the cell cycle; *Cdc46*, that encodes a member of the Mcm2-7 family of proteins involved in the initiation of DNA replication; *Cdc27*, subunit of the Anaphase-Promoting

Complex/Cyclosome (APC/C); *Orc1* which directs DNA replication; *Bim1* which is the microtubule-binding protein that together with Kar9p delays the exit from mitosis when the spindle is oriented abnormally; *Rnr1* (Ribonucleotide-diphosphate reductase), which is regulated by DNA replication and DNA damage checkpoint pathways; *Dsk2*, a nuclear-enriched ubiquitin-like polyubiquitin-binding protein, required for spindle pole body (SPB) duplication and for transit through the G2/M phase of the cell cycle; *Tub2* the beta-tubulin, which associates with alpha-tubulin to form tubulin dimmer; the dimers polymerize to form microtubules, required for mitosis.

We also carried out several tests by repeatedly running the Genetic Algorithm for 400 evolutionary steps with different initializations. We compared the final and initial generations observing that: 1) in the final population some of the hubs are unchanged (*Bim1*, *Clb2*, *Dsk2*, *Rnr1* and *Swi4*), while some are added. 2) The number of connections varies approximately from 1100 to 1300. This means that the majority of the links comes from the experimental data of Hughes and that the method used adds approximately the 19% of the initial connections. 3) The improvement of the fitness of the best model with respect to the initial conditions ranges between 3% (worst case) to the 4.5% (best case). We are now performing other tests with different fitness functions, such as AIC or BIC, to evaluate the robustness of the results herein described.

For what concerns the analysis of the best model obtained, we evaluated also the variability of the network topology across the members of the final population. Again, some of the hubs are unchanged (*Bim1*, *Clb2*, *Dsk2*, *Rnr1* and *Swi4*), while there is one gene which is suggested to be a Hub in 15 over 20 members of the population (*Tub2*), and a set of other genes has variable frequency (*Cdc46*, *Ctf18*, *Scm4*, *Sth1* and *Taf6*).

The goodness of fit of the learned model is satisfactory, with an overall RMSE of 0.047. An example of the one step ahead prediction for one of the analyzed genes is shown in Figure 3.

4 Conclusions

The approach described in this paper is an example of how different knowledge and data sources can be conveniently integrated in gene network learning. The method was able to reconstruct known relationships among genes and to provide meaningful biological results. It seems therefore suitable of further investigations and refinements. In particular, we plan to include in the strategy also data available from protein-protein interactions.

Acknowledgments

This work is part of the PRIN Project “Modelli dinamici dell'espressione genica da dati di microarray: tecniche di clustering e reti di regolazione” funded by the Italian Ministry of Education. We gratefully thank Sonia Rinaldi for her help in running the Genetic algorithm tests.

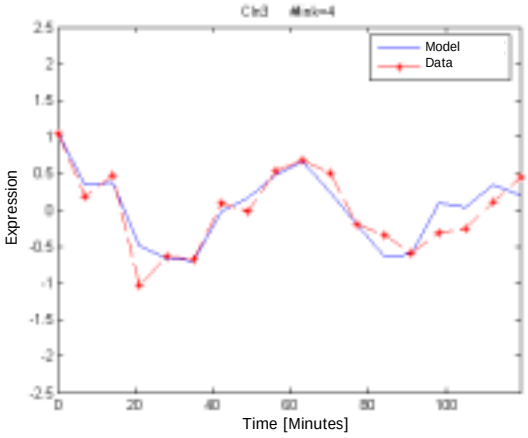


Figure 3. Raw data and one step ahead prediction for gene Cln3.

References

[De Jong, 2002] H. De Jong. Modeling and Simulation of Genetic Regulatory Systems: A literature Review. *Journal of Computational Biology*, 9(1): 67-103, 2002.

[Hughes *et al.*, 2000] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109-26, 2000.

[Lee *et al.*, 2002] T.I. Lee et al Transcriptional regulatory networks in Saccharomyces cerevisiae. Science. 2002 Oct 25;298(5594):799-804

[Liang *et al.*, 1998] S. Liang, R. Somogyi, S. Fuhrman. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In *Proceedings of Pacific Symposium on Biocomputing*, 3: 18-29,1998.

[Perrin *et al.*, 2003] B.E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, F. D'Alche-Buc. Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, 19 Suppl 2:138-148, 2003.

[Schlitt *et al.*, 2005] T. Schlitt, A. Brazma Modelling gene networks at different organisational levels. FEBS Lett. 2005 Mar 21;579(8):1859-66. Review.

[Sveiczer *et al.*, 2004] A. Sveiczer, J.J. Tyson, B. Novak. Modelling the fission yeast cell cycle. Brief Funct Genomic Proteomic. 2004 Feb;2(4):298-307.

[Spellman *et al.*, 1998] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces Cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12): 3273-97, 1998.

Mutant vs. Gene Expression Profiles for Function Prediction

Tomaz Curk,[a] Uros Petrovic,[b] Gad Shaulsky,[c] Blaz Zupan[a,c]

a) University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia

b) J. Stefan Institute, Department of Biochemistry and Molecular Biology, Ljubljana, Slovenia

c) Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX

tomaz.curk@fri.uni-lj.si, blaz.zupan@fri.uni-lj.si, gadi@bcm.tmc.edu, uros.petrovic@ijs.si

Abstract

A popular utility of microarray data is to make inference on gene function based on similarity of its expression to expressions of other, functionally already annotated genes. This approach may use available collections of gene expression measurements that study organisms under different conditions. An alternative way, enabled by recent advances in biotechnology, is to associate gene function to a phenotype of its corresponding mutant defined by expressions of all other, not mutated genes. In the paper, we use a technique called gene-coexpression networks to compare the two approaches, and apply it to data on budding yeast *S. cerevisiae*. In terms of gene function prediction and contrary to our expectations, we found that mutant phenotypes are on the overall not more informative than gene expression profiles, and we provide biological explanation why some gene functions can be better predicted using one type of data and not the other.

1 Introduction

The development of DNA microarrays allows whole-genome expression profiling, measuring the expression of each of the genes in a single assay. Changes in gene expression are often related to specific cellular needs and most expression profiling studies try to identify genes that respond to specific conditions or treatments. Recently the idea that expression of all genes could be used as an indication of cellular state has received great attention [Alizadeh *et al.*, 2000; Bittner *et al.*, 2000; Hughes *et al.*, 2000]. Whole-genome expression profiles of mutants thus hold great promise for rapid genome function analysis. It is plausible that the mutant expression profile could serve as a universal phenotype [Hughes *et al.*, 2000; Hughes, 2005; Van Driessche *et al.*, 2005] and as such is believed to be very informative for assigning gene function.

Instead of associating gene function to its expression pattern under different conditions, we can consider the entire microarray profile of a strain that is mutated in one gene as an indicator of that gene's function. This method has been demonstrated successfully in yeast [Hughes *et al.*, 2000] and in cancer cell characterization [Alizadeh *et al.*, 2000; Bittner *et al.*, 2000]. The reason to use this alternative also follows the finding that gene expression and gene function show very little correlation on a global scale (less than 10% of the cases) [Winzeler *et al.*, 1999]. When reasoning on gene function classical genetics has much depended on observational mutant phenotypes (*e.g.* "mutant grows", "does not grow", "sporelates", etc.), leading us to believe that their modern transcriptional variants will provide a robust funding for function prediction.

The study reported here was inspired by investigation of Stuart *et al.*, who showed that, in contrary to above, predicting gene function using evolutionary conservation in the wild is more sensitive than scoring the phenotype resulting from strong loss-of-function mutants in the laboratory [Stuart *et al.*, 2003]. They base their work on the assumption that genes commonly found in diverse organisms and with by-organism correlated expression patterns under a large number of diverse conditions imply functional relation. In order to distinguish accidentally co-regulated genes from those that are physiologically important they observe the evolutionary conservation between multiple species (yeast, worm, fly and human). They believe that evolutionary conservation is a powerful criterion to identify genes that are functionally important from a set of co-regulated genes and that co-regulation of a pair of genes over large evolutionary distances implies that the co-regulation confers a selective advantage, most likely because the genes are functionally related [Stuart *et al.*, 2003]. Using the method of gene-coexpression networks they were able to identify several examples of evolutionary conserved functional groups with high gene coexpression. Importantly, they also show that predictive accuracy is much poorer when using only the data on a single-organism.

In this paper we compare the level of information in mutant expression data with the information in gene expression data. Similarly to Stuart *et al.* we base our work on the assumption that similar expression profiles imply similar function and use the method of gene-coexpression networks to make the comparison.

The method of gene-coexpression networks requires a measure of distance in gene expression. The most common way, the one also used in [Stuart *et al.*, 2003], to measure distance (or similarity) in gene expression between pairs of genes is to compare their expression profiles, where expression of a gene is measured under different experimental conditions. We will call this type of

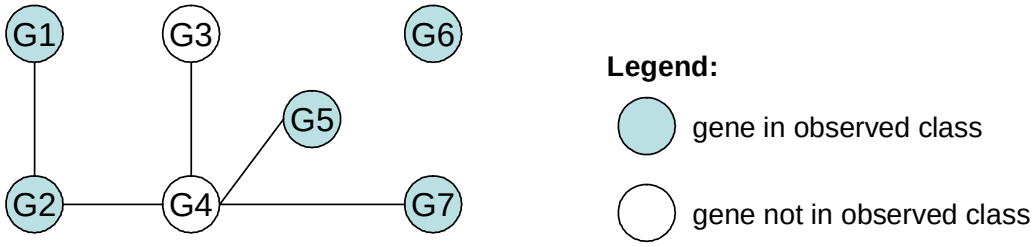


Figure 1. Gene-coexpression network for seven genes (G1-G7). We connect those pairs of genes that are correlated more than an arbitrarily selected threshold (some genes might not be connected, gene G6 in this example). There are five genes annotated to the observed class (nodes G1, G2, G5, G6 and G7), but only two of them are connected to each other (G1 and G2). The class coverage of this network is then: 2/5=0.4. To calculate accuracy we need to count edges. There is only one edge connecting two genes in observed class (edge G1-G2). There are four edges coming from class genes (edges G1-G2, G2-G4, G5-G4, G7-G4). The accuracy of this network is then 1/4=0.25.

data *gene profile*. As an alternative, we measure distance between genes using transcriptional profiles of mutant strains. For a mutated, usually deleted gene, expression of all genes in mutant's genome is measured. When comparing two genes, we compare transcriptional profiles of the respective two mutant strains. We will call this type of data *mutant profile*. In both cases and like in the study by [Stuart *et al.*, 2003], we use Pearson correlation as a distance function.

2 Data and methods

In this section we describe the microarray gene expression data sets and gene functional annotations used. We also describe the method of so called gene-coexpression networks that we applied to measure the ability to predict function from gene coexpression.

2.1 Gene expression and functional annotation

We have used two data sets of microarray gene expression measurements. For *gene profiles* we used data from a study of cell-cycle in *S. cerevisiae* where whole-genome expression under 73 conditions was measured by [Spellman *et al.*, 1998]. For *mutant profiles* the data was obtained from a compendium of whole-genome expression measurements of 300 diverse mutants and chemical treatments in *S. cerevisiae* as performed by [Hughes *et al.*, 2000].

To test the results of our predictions we have used existing functional annotations on 76 GO slim terms, which is a collection of high level Gene Ontology (GO) terms [Ashburner *et al.*, 2000] that best represent the major biological processes, functions, and cellular components that are found in *S. cerevisiae* (data available at http://www.yeastgenome.org). We also used KEGG annotation [Ogata *et al.*, 1999] for four functional classes described and used in [Stuart *et al.*, 2003]: Cell cycle, Proteasome, Oxidative phosphorylation and Ribosome.

2.2 Gene-coexpression networks

The method presented in [Stuart *et al.*, 2003] measures the correlation between gene coexpression and function. The method, called *gene-coexpression networks*, requires a measure of gene coexpression (or distance in gene expression) in order to build a network of coexpressed genes. In their paper, Stuart *et al.* used gene expression measured under different conditions. We have applied the same method to relate genes based on their mutant-based transcriptional phenotypes.

A gene-coexpression network is a graph where nodes represent genes. Edges in this network connect two nodes if coexpression of their corresponding genes is higher than an arbitrary threshold. By varying this threshold we can generate different networks: from relatively unconnected networks, where only the most coexpressed genes are linked, to highly connected networks with edges relating also genes with low correlation. Each time we can measure the connectivity properties of genes from a selected functional class. One such measure of connectedness is *coverage*: the percentage of class genes that are connected to at least one gene from same class. The other is *accuracy*: the number of edges connecting genes from same class divided by the number of all edges coming from class genes (see Figure 1 for example).

Gene-coexpression networks can be seen as a method to cluster genes. At the same time, by varying the threshold, they also give a general overview of the relation between function and gene coexpression. In their study, [Stuart *et al.*, 2003] verified the significance of the interactions in such networks by means of a variety of statistical and permutation tests. They compared the number of interactions (links) in random networks with real networks, the influence of selection of microarray experiments on the ability to identify interactions, and the influence of noise in microarray data on the constructed networks. They found the method to be robust and appropriate for the task. For details see [Stuart *et al.*, 2003]. Among other things, they showed that genes from some functional classes were highly inter-connected in the coexpression network, indicating a correlation between function and coexpression.

2.3 Performance of gene-coexpression networks

We can plot coverage and accuracy values of gene-coexpression networks obtained at different thresholds of gene coexpression for a selected functional class of genes (see Figure 2a). At high thresholds the class coverage is close to zero, because the networks include only a few edges. If genes from the same functional class are highly connected (and at the same time disconnected from genes

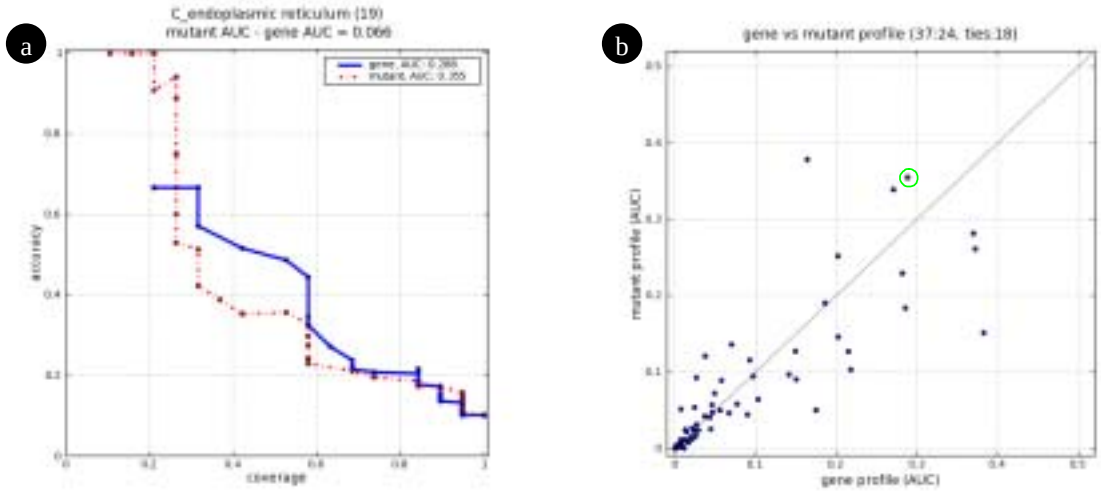


Figure 2. a) Comparison of the two performance curves of gene-coexpression networks built for class "C: endoplasmic reticulum" (C indicates that the term is from "cellular component" aspect of GO). There are 19 genes in the class (number indicated in parentheses). Solid line curve shows the performance of gene-coexpression network built using gene profiles (with AUC = 0.288), and dash-dot line curve the performance when using mutant profiles (with AUC = 0.355). A simple comparison of the two values tells us that mutant profiles are more correlated than gene profiles for the selected functional class. **b)** Graph showing AUCs obtained using gene profiles (X axis) and mutant profiles (Y axis) for ~80 functional classes. Each point represents a functional class. Its X coordinate is AUC of performance curve obtained using gene profiles, Y coordinate is AUC of performance curve obtained using mutant profiles. Encircled, at coordinates (0.288, 0.355), is class from example in Figure 2a. Gene profile wins 37 times, mutant profile wins 24 times. There are 18 ties – cases when both AUCs are equal.

in different functional class), we obtain high accuracy. Relaxing the threshold, coverage monotonically increases at increasing risk for lower accuracy (accuracy may increase, however, but would on average decrease monotonically; see mutant curve in Figure 2a). The closer is the curve to point (1, 1) – the highest coverage and the highest accuracy – the better. By calculating the *area under the curve* (AUC), we can summarize the two measures into a *single performance value* that describes the level of correlation between gene function and coexpression.

2.4 Quantitative comparison of gene function prediction from gene and mutant profiles

The "gene profile" curve in Figure 2a was obtained using coexpression of gene profiles from Spellman gene profile data. If we now derive coexpression of Hughes' mutant profiles and use it to build gene-coexpression networks, we can plot both curves and compare their AUCs for an observed functional class. Example in Figure 2a indicates that mutant profile might be more indicative for gene function "C: endoplasmic reticulum" because it has a higher AUC.

By observing performance plots of different functional classes we can count the number of times a profile type "*wins*," *i.e.* is more indicative to predicted class. Figure 2b is a summary of comparisons for all classes considered in this study. Each point represents a functional class, its X and Y coordinates are the AUCs of gene-coexpression networks obtained using gene and mutant profiles respectively. Points below the diagonal are functional classes that can be better predicted using gene profiles (AUC using gene profiles is higher than AUC using mutant profiles), whereas those above the diagonal are cases where mutant profiles are more indicative. By observing the number of points on each side of the diagonal, we can then easily see which profile type is generally more informative.

3 Results and Discussion

We measured the performance of gene-coexpression networks built from gene and mutant profiles for 80 functional classes. Results for selected functional classes are summarized in Table 3, where we give the difference in AUCs obtained using the two types of profiles. In Table 3 we subtracted the AUC obtained using gene profile from the AUC obtained using mutant profiles. A positive difference indicates that mutant profile is more informative, a negative difference that gene profile is more informative.

Looking at results in Table 3 and Figure 2b we find a slight indication that gene profiles might be more informative than mutant profiles. We base this on higher AUCs values for gene profiles (compare values in top and bottom rows in Table 3) and the prevailing number of times that gene profile wins in Figure 2b.

Analysis of functional classes, for which either gene expression profiles or mutant profiles are more informative than the other, resulted in a list of classes (Table 3) that are in accordance with current understanding of regulation of cellular functions in yeast. There are some functional classes of yeast genes that are transcriptionally regulated and thus have relatively uniform expression profiles. This is however not a general phenomenon and other functional classes include genes coding for proteins whose activities are not regulated on the level of transcription. For some of those genes, *e.g.* for those coding for

regulatory proteins that affect transcription of other genes, it is the coexpression of their target genes that can be used for functional classification, and these classes were primarily identified through mutant profiles in our study. Functional classes for which gene expression profiles are more informative than mutant profiles included genes involved in cell cycle-related processes and functions (classes "Cell cycle", "DNA metabolism") and genes for ribosomal proteins (classes "Protein biosynthesis", "Structural molecule"). These classes of genes have been previously shown as prime examples of agreement between function similarity and gene coexpression in several studies [DeRisi *et al.*, 1997; Spellman *et al.*, 1998]. Also classified as functional classes of genes for which gene coexpression correlates with functional similarity were genes involved in metabolism and transport, *i.e.* genes coding for enzymes and transporters, and genes involved in response to stress, in agreement with previous findings [Causton *et al.*, 2001; Gasch *et al.*, 2002].

On the other hand, in the group of genes for which gene coexpression is less informative than mutant profiles, based on previous knowledge about regulation of cellular functions in yeast we expected functional classes including regulatory genes. Their expression levels do not change significantly in response to perturbations, but rather they affect gene expression of their target genes. In agreement with our expectations, the list was composed of functional classes such as "Protein kinase activity", "Protein binding activity", "Transcription" and "Transcription regulator activity", "Protein modification", "Signal transducer" and "Enzyme regulator activity", that all consist of genes coding for regulatory proteins. Unexpectedly, however, there were three additional functional classes in this group that do not contain significant amount of genes known to have regulatory functions. These classes are "Lipid metabolism", "Cytoskeleton organization and biogenesis" and "Cell wall organization and biogenesis." Intriguingly, genes belonging to these classes could have roles in cellular physiology that are, from a global perspective, more important in regulatory activities than in their direct metabolic and structural functions.

3.1 Additional experimental studies

Since some of GO annotations are inferred from expression data (in particular, this includes all annotations with GO annotation evidence codes IEP and RCA), this could be the reason for higher performance of gene-coexpression networks in general. We therefore removed IEP and RCA annotations (~5% of all annotation for yeast), and thus removed about 120 genes with no annotation left. After this procedure the results changed only slightly (see Table 4), and gene profiles still appear to be more informative (compare the two graphs in first row in Table 4).

We then performed two more tests to see how the results change if we use other gene and mutant profile data (graphs in second and third row in Table 4). First, we tried to use a different set of gene profiles (second row in Table 4), and used gene profiles from the same data set as used for mutant profiling (from Hughes dataset). In this case gene profiles consisted of measurements from approx. 270 conditions (each mutant can be seen as a condition). The mutant profile data remained the same as in our first test (expression of 6316 genes in mutant's genome). In this test, mutant profiles are slightly more informative than their gene expression profiles alternative, but the later are still winning on the overall (see second row in Table 4 and compare it to first row).

In our last test we observed if there is any difference if we look at mutant data in two different ways: as a mutant profile or as a gene profile. To do so, we used only the expression of 270 mutated genes to build mutant profiles. In this case mutant profiles become more informative (see third row in Table 4 and compare it to first and second row).

mutant - gene profile performance AUC	Functional annotation
-0.1147	P: DNA metabolism
-0.1110	P: protein biosynthesis
-0.0880	Ribo
-0.0600	P: cell cycle
-0.0564	F: structural molecule activity
-0.0528	F: transporter activity
-0.0390	F: transferase activity
-0.0219	F: oxidoreductase activity
-0.0194	P: transport
-0.0191	F: hydrolase activity
…	…
0.0109	F: signal transducer activity
0.0229	P: protein modification
0.0230	F: transcription regulator activity
0.0293	P: cell wall organization and biogenesis
0.0313	P: transcription
0.0437	P: cytoskeleton organization and biogenesis
0.0498	P: lipid metabolism
0.0657	F: protein binding
0.0662	F: protein kinase activity
0.0681	Oxid

Table 3. Difference in gene-coexpression network performance when using mutant and gene profiles. Only top ten classes for each profile type are listed. Functional classes that can be better predicted using gene profiles are listed on the top, while those better predicted using mutant profiles are shown on the bottom of the list. Prefix "P" indicates the "biological process" aspect, "F" the "molecular function" aspect of GO. Classes Ribo and Oxid are taken from Stuart *et al.*

4 Conclusion

Overall, we found no clear difference between the information coming from gene and mutant profile data. On the contrary from what was our expectation (and perhaps an unstated belief of the community), there is a slight indication that gene profiles (*i.e.* observing gene expression un-

der different conditions) might, on the overall, be more correlated to gene function than mutant profiles (*i.e.* observing expression of mutants in same condition). But, when studying a particular function, there may be a clear difference between the two approaches that can be explained with existing biological knowledge. This is a clear indication that both sources of experimental data may be used in order to successfully predict gene function. We are currently investigating ways to automatically learn how to combine both profile types for better function prediction.

The principal novelty of reported work is in direct comparison of the utility of gene expression profiles and transcriptional phenotypes of mutants for gene function prediction. The two data sources were first studied together and qualitatively compared in [Hughes *et al.*, 2000], while other references on utility of mutant-based transcriptional phenotyping are at best rare. Gene expression networks and accuracy-coverage graphs, together with utility of function annotation data bases, allowed us to compare two sources quantitatively and to draw conclusions related to particular functional groups.

Acknowledgments

This work was supported in part by Program and Project grants from the Slovene Ministry of Education, Science and Sports and by a grant from the National Institute of Child Health and Human Development, P01 HD39691.

References

[Alizadeh *et al.*, 2000] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**(6769): 503-11.

[Ashburner *et al.*, 2000] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**(1): 25-9.

[Bittner *et al.*, 2000] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, *et al.* (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**(6795): 536-40.

[Causton *et al.*, 2001] H. C. Causton, B. Ren, S. S. Koh, C. T. Harbison, E. Kanin, E. G. Jennings, T. I. Lee, H. L. True, *et al.* (2001). Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell* **12**(2): 323-37.

[DeRisi *et al.*, 1997] J. L. Derisi, V. R. Iyer and P. O. Brown (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**(5338): 680-6.

[Gasch *et al.*, 2002] A. P. Gasch and M. Werner-Washburne (2002). The genomics of yeast responses to environmental stress and starvation. *Funct Integr Genomics* **2**(4-5): 181-92.

[Hughes, 2005] T. R. Hughes (2005). Universal epistasis analysis. *Nat Genet* **37**(5): 457-8.

[Hughes *et al.*, 2000] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, *et al.* (2000). Functional discovery via a compendium of expression profiles. *Cell* **102**(1): 109-26.

[Ogata *et al.*, 1999] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono and M. Kanehisa (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**(1): 29-34.

[Spellman *et al.*, 1998] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, *et al.* (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**(12): 3273-97.

[Stuart *et al.*, 2003] J. M. Stuart, E. Segal, D. Koller and S. K. Kim (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**(5643): 249-55.

[Van Driessche *et al.*, 2005] N. Van Driessche, J. Demsar, E. O. Booth, P. Hill, P. Juvan, B. Zupan, A. Kuspa and G. Shaulsky (2005). Epistasis analysis with global transcriptional phenotypes. *Nat Genet* **37**(5): 471-7.

[Winzeler *et al.*, 1999] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, *et al.* (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**(5429): 901-6.

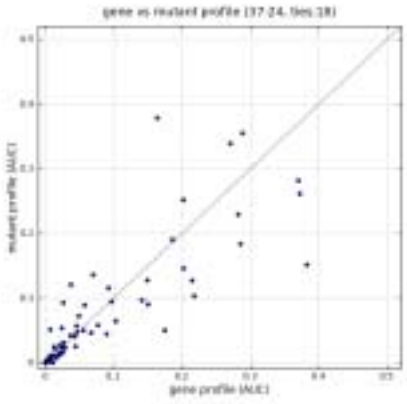
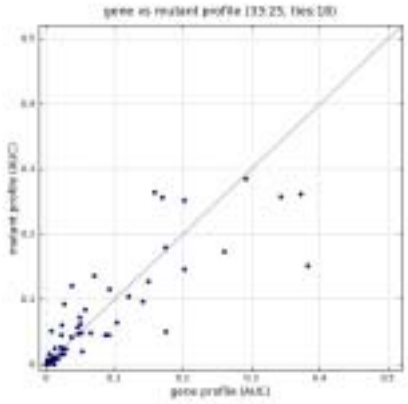
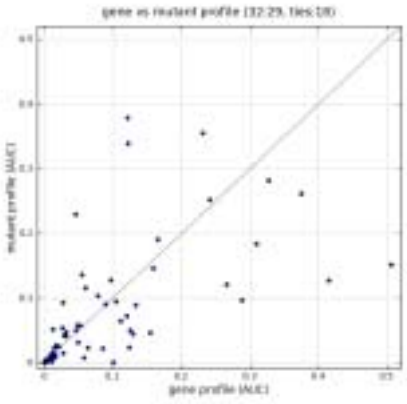
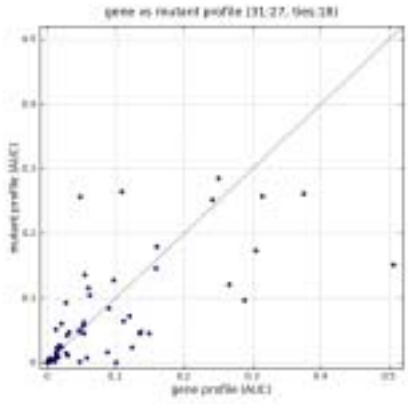
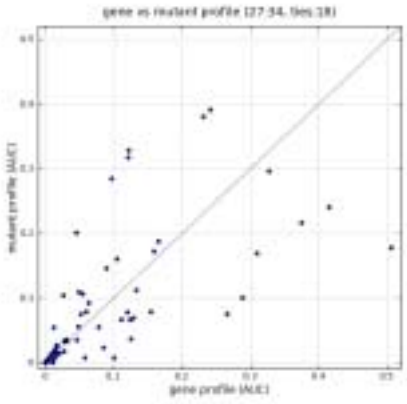
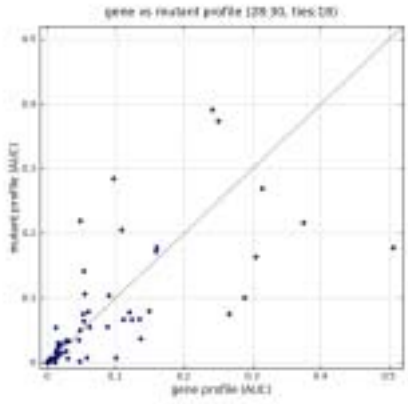
		Annotation	
		All annotation from GO slims	All annotation except IEP and RCA evidence codes (inferred from expression)
Profiles	**Gene profile:** expression in 73 conditions (Spellman data) **Mutant profile:** expression of all 6316 genes in mutant (Hughes data)		
	Gene profile: expression in 272 mutants (taken as conditions, Hughes data) **Mutant profile:** expression of all 6316 genes in mutant (Hughes data)		
	Gene profile: expression in 272 mutants (as conditions, Hughes data) **Mutant profile:** expression of only 270 mutated genes in mutant (Hughes data)		

Table 4. Comparison of performance of gene-coexpression networks built using different sources of gene and mutant profile data, for two types of annotation. Graphs in first column show the gene *vs.* mutant profile AUCs comparison when using all annotation; second column, after annotation with IEP and RCA evidence codes was removed.

Instance-based Prognosis in Intensive Care Using Severity-of-illness Scores

Clarence Tan[1], **Linda Peelen**[1,2,*], **and Niels Peek**[1]
[1] Department of Medical Informatics, Academic Medical Center
University of Amsterdam, The Netherlands
[2] Dutch National Intensive Care Evaluation (NICE) Foundation

Abstract

This paper explores the use of instance-based reasoning (IBR) to estimate the probability of hospital death in patients admitted to the Intensive Care Unit (ICU). The predictions are based on severity-of-illness scores that indicate the state of the patient. We have implemented an instance-based reasoning algorithm as an alternative to logistic regression (LR) models to predict hospital mortality. The performance was measured and prospectively validated. Results show that instance-based reasoning is competitive to logistic regression.

1 Introduction

Clinical scoring systems are tools for assessing the states of patients and quantifying the severity of their condition [Wyatt, 1990]. They are used in many medical disciplines, including cardiology, oncology, and critical care, and can be used for a variety of clinical and management tasks such as comparative audit among practitioners, measuring the effects of treatment, and risk assessment and prognosis. In this paper, we focus on the application of scoring systems in prognosis with binary outcome variables.

In most scoring systems, patient-specific data is used to arrive at an integer value that represents the severity of a patient's illness. Because points are assigned to deviations from normal values, low values (close to zero) generally represent mild conditions, whereas higher values are associated with more serious conditions. When clinical scores are used in prognosis, a model has to be developed that converts these scores into patient-specific predictions. With a binary outcome variable, the model needs to convert scores into either predicted outcome classes or into probabilities. The predominant methodology for doing this is logistic regression (LR) analysis [Hosmer and Lemeshow, 2000], where the score is used as a linear covariate.

Although LR analysis has proven to be a powerful modeling methodology in the biomedical field, it is based on assumptions that are questionable for most clinical scoring systems. In particular, logistic regression assumes that there exists a fixed (usually linear) relationship between score and log odds of the outcome probability over the entire score range. In practice, however, most scoring systems were not designed to have this property, and the relationship between score and (log odds of the) outcome may vary over the score range, and may be highly nonlinear.

In this paper, we study the use of instance-based reasoning (IBR) as an alternative for LR analysis in scoring-based prognosis. IBR is a nonparametric prediction method that is based on the assumption that the prognosis of a new patient resembles those of past patients with similar characteristics. The IBR method employed is the weighted k-NN regression algorithm with an adaptive neighborhood size. The main advantage of instance-based reasoning is that it makes few assumptions regarding the relationship between predictors and outcome. Furthermore, being a 'lazy' learning method, it is less sensitive to population drift than eager (model-based) learning methods such as LR. The main disadvantage is that it requires relatively large datasets (compared to parametric methods), and does not work well in high-dimensional domains. Finally, when it is used to estimate probabilities, as in our application, these may be biased (structurally too high or too low), a phenomenon that does not occur in model-based methods.

The method was applied to data from two popular scoring systems for intensive care patients, the APACHE II [Knaus *et al.*, 1985] and SAPS II [Le Gall *et al.*, 1993] scores. The resulting mortality estimators were validated and compared with LR models internally (with cross-validation on the training dataset) and externally (on a prospectively collected dataset).

The paper is organized as follows. Section 2 reviews the two scoring systems that were employed; Section 3 provides details on the datasets, IBR prediction method, and validation procedure. Section 4 describes the results from our study and Section 5 finishes the paper with a discussion and conclusions.

2 APACHE II and SAPS II scoring systems

Various scoring systems have been developed for the field of intensive care medicine [Gunning and Rowan, 1999]. In this study, we have used the Acute Physiological And Chronic Health Evaluation (APACHE) II [Knaus *et al.*, 1985] and the Simplified Acute Physiology Score (SAPS) II [Le Gall *et al.*, 1993] scores. Both scores are assessed during the first 24h of a patient's ICU stay, and can be converted into an estimated probability of death by means of

*Corresponding author. E-mail: l.m.peelen@amc.uva.nl

an associated LR model. The APACHE II score has a minimum of 0 and a maximum of 71 points; it summarizes mainly physiological information, and the associated LR model employs information on the patient's diagnosis at admission (54 categories) and type of admission (6 categories) besides the score. The SAPS II score ranges from 0 to 163 points; it summarizes physiological, diagnostic, and admission-type information; the associated LR model only employs the score itself.

An important difference between the APACHE II and SAPS II scoring systems is that the former is based on knowledge from practitioners, whereas the latter is based on data analysis. The APACHE II scoring system was designed during a consensus meeting with experienced intensive care clinicians; the associated prognostic model is based on LR analysis of a multicenter dataset of ICU admissions. The SAPS II scoring system, in contrast, was obtained by scaling the coefficients that were derived by multiple LR analysis on a large multicenter dataset.

Both scoring systems consider patients who have undergone cardiac surgery as special cases. These patients usually stay for observation at the ICU and leave for further recovery at the nursing ward once their condition is stable. We can compute scores for these patients, but the associated probability estimates from the LR models are believed to be unreliable [Knaus *et al.*, 1985; Le Gall *et al.*, 1993].

3 Data and methods

3.1 Data

The Dutch National Intensive Care Evaluation (NICE) register [NICE, 2005] provided two datasets containing information on ICU admissions. The first dataset describes 1559 ICU admissions from 7 Dutch hospitals between January 2003 and August 2003 and was used as a training set. In this dataset the hospital mortality is 14.8%.

During our study a second dataset was provided consisting of 1868 ICU admissions from August 2003 to June 2004. It was used to validate the IBR estimators that were developed on the first set. The hospital mortality in this dataset is 16.3%. The difference in mortality between the two datasets is not significant ($\chi^2 = 1.38$; $p = 0.24$). Both sets contain all variables required to compute the APACHE II and SAPS II scores, the scores themselves, and the associated probabilities of death estimated by the APACHE II and SAPS II LR models.

Using these data in total eight IBR estimators were developed using different (combinations of) predictive features. Two univariate IBR estimators were developed, one for the APACHE II score, and one for SAPS II score. Because the APACHE II score does not include information on the patient's diagnosis and type of ICU admission, also three multivariate estimators were developed for the APACHE II score in combination with diagnosis category, admission type, and both. Finally, a multivariate IBR estimator was developed on the basis of the two scores together.

As discussed in Section 2, predictions from the APACHE II and SAPS II LR models are believed to be unreliable for cardiac surgery patients and therefore should not be used. In the IBR estimators described above we have neglected this exclusion criterion and make predictions for all ICU patients in the same manner. Therefore we refer to these IBR estimators as *single method estimators*.

To take the exclusion criterion for cardiac surgery ICU admissions into account, we developed two more estimators, called the *dual method estimators*. Here we use the clinical scores (APACHE II and SAPS II respectively) to arrive at predictions for the patients who did not undergo cardiac surgery, and four alternative features for patients who arrive at the ICU after cardiac surgery. The four alternative features are minimum temperature, minimum systolic blood pressure, minimum bicarbonate, and maximum creatinine (all during the first 24h of ICU stay); they have been shown to be important predictors of mortality in cardiac surgery patients [Verduijn, 2002].

All IBR estimators were constructed with an extension of the *weighted k-NN regression algorithm*.

3.2 Prediction method

In weighted k-NN regression, predictions are obtained by computing a weighted average of the outcomes of the k training instances that are most similar to query instance $\mathbf{x}_q$. In the case of a binary outcome Y, we have

$$\hat{p}(Y = 1|\mathbf{x}_q) \quad = \quad \frac{\sum_{i=1}^{k} K_\lambda(\mathbf{x}_q, \mathbf{x}_{[i]}) \cdot y_{[i]}}{\sum_{i=1}^{k} K_\lambda(\mathbf{x}_q, \mathbf{x}_{[i]})}, \qquad (1)$$

where $\mathbf{x}_{[1]}, \ldots, \mathbf{x}_{[k]}$ are the k training instances most similar to $\mathbf{x}_q$, and $K_\lambda(\mathbf{x}_q, \mathbf{x}_{[j]})$ is the weight assigned to training instance $\mathbf{x}_{[j]}$. This is called the Nadaraya-Watson kernel-weighted average [Hastie *et al.*, 2001, Ch. 6]. In our application, $\hat{p}(Y = 1|\mathbf{x}_q)$ is the patient's estimated probability of hospital death, given the feature-value vector $\mathbf{x}_q$ (e.g. APACHE II score and diagnosis category).

Three important choices have to be made when weighted k-NN regression is applied: 1. How do we find similar training instances (*choice of distance metric*)?, 2. How are distances transformed into weights (*kernel function*)?, and 3. How many neighbors are used to make predictions (*neighborhood size*)? Each of these questions is addressed below.

Distance metric In the univariate IBR estimators we have used the *score difference* to quantify the distance between instances. In the multivariate estimators, local distance metrics were constructed for each of the predictive features. For non-numeric features (diagnosis category and ICU admission type), these local metrics were defined by distance matrices based on the hierarchical relations between feature values; we refer to [Tan, 2005] for details. In the prediction phase, local distances were normalized and then combined using the *Manhattan metric* (i.e. taking the unweighted sum of all normalized local distances).

Kernel function The kernel is a function that assigns a nonzero weight to all instances within the neighborhood of k nearest training instances, and zero weight to all other instances. We have used two kernel functions in our experiments, the *uniform kernel* and the *Epanechnikov kernel*

[Silverman, 1986]. The uniform kernel assigns unit weight to all k nearest neighbors, thus treating them as equally important. The Epanechnikov kernel, in contrast, is a non-linear function that approaches 1 at small distances to the query instance, and 0 at the boundaries of the neighborhood:

$$K_\lambda(\mathbf{x}_q, \mathbf{x}_{[i]}) \quad = \quad \begin{cases} \frac{3}{4}(1-t^2) & \text{if } |t| \leq 1, \\ 0, & \text{otherwise,} \end{cases} \qquad (2)$$

where $t = d(\mathbf{x}_q, \mathbf{x}_{[i]})/d(\mathbf{x}_q, \mathbf{x}_{[k]})$ is the normalized distance between neighbor $\mathbf{x}_{[i]}$ and the query instance $\mathbf{x}_q$.

Neighborhood size Most algorithms for k-NN classification and regression (e.g. those implemented in WEKA [Witten and Frank, 2001]) choose a fixed number of neighbors to make all predictions. However, usually the values of predictive features are not uniformly distributed over their theoretical range. As a result the width of the neighborhood that is necessary to obtain the k nearest neighbors varies with the sparsity of the data in the neighborhood of the query instance. However, when the neighbors are weighed according to their distance to the query instance, a single close neighbor yields the same amount of weight as multiple distant neighbors together. A better option is therefore to let the neighborhood width depend on the total weight of the neighbors rather than the number of neighbors [Hastie *et al.*, 2001]. This implies that the neighborhood width varies with the position of the query instance in the instance space and is locally adapted to the sparsity of the data.

In our application, a *target total weight* (`ttw`) of the instances in the neighborhood was established during the learning phase. The value of `ttw` is constant over the feature space, but needs to be optimized for the predictive feature(s) and the type of kernel function that are used to predict mortality. To find the optimal value for `ttw`, the following method was employed. For each IBR estimator, both kernel types and `ttw` values of 5, 10, 20, 50, 100, 200 and 500 were employed in a jackknife cross-validation procedure. In each run of the procedure, the estimator's accuracy was determined. Based on the results, the kernel type and `ttw` value were chosen.

Within this procedure, predictive accuracy was measured by the R^2 statistic [Ash and Shwartz, 1999]:

$$R^2 \quad = \quad 1 - \frac{\sum_{i=1}^{N}(\hat{p}(Y=1 \mid \mathbf{x}_i) - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2}, \qquad (3)$$

where N is the size of the training dataset and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ is the mean outcome value. The R^2 statistic is inversely proportional to the mean squared error and the Brier score.

Figure 1 shows an example for the APACHE II score. The best performance is obtained with the Epanechnikov kernel and `ttw` values of 50 and 100. Because larger `ttw` values correspond to simpler models, we choose the Epanechnikov kernel with a `ttw` value of 100.

This procedure of selecting the optimal settings has been applied for all IBR estimators.

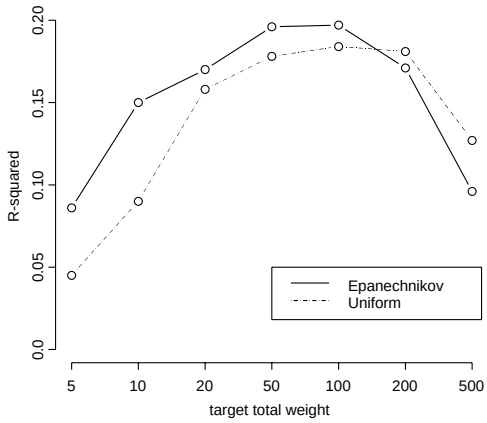


Figure 1: R^2 performance statistic for the APACHE II IBR estimator, plotted against the `ttw`, for both kernel types.

3.3 Validation

The IBR estimators were internally and prospectively validated. In the internal validation the performance of the estimator was measured by jackknife cross-validation on the training data. The estimator was also validated on prospectively collected data, using the second data set provided by the NICE register. Three different procedures were used in this prospective validation.

The first prospective validation procedure, the *settings validation*, aims to check whether the settings for kernel type and target total weight that were optimized on the training data, yield comparable performance on the prospective test dataset. To this end, we only use these settings, but not the training data for prediction; instead jackknife cross-validation is applied on the test set. Because the test set is larger than the training set, we expect the measured performance to be equally good or better if the chosen settings are valid.

The second prospective validation procedure is called the *plain prospective validation*. This procedure aims to investigate how well the algorithm generalizes to prospective data. To this end, predictions are made for all instances in the test set, while the training set serves as the instance base. We use the settings for kernel type and `ttw` value that were found on the training set.

One interesting property of IBR is the fact that it is a *lazy learning* method: generalization over examples in the instance base takes place no sooner than at the time of making predictions. The third prospective validation procedure, called *incremental prospective validation*, takes advantage of this property by incrementally adding instances from the test set and using them for future predictions. To this end, records in the test set were ordered by ICU admission date, and evaluated in that order. When evaluating a given record with admission date d, the instance base consists of all records from both the training set and test set with discharge date prior to d. For the first record from the test set this procedure yields the same prediction as in the second validation procedure. But for later records, the number of possibly similar instances is much larger, and there-

Estimator method	Predictive feature(s)	Kernel type	ttw	Relative bias	AUC $\pm$ S.D.
Single	APACHE II	Epan	100	-3.85	0.792 ± 0.033
Single	SAPS II	Unif	50	-1.65	0.860 ± 0.030
Single	APACHE II, SAPS II	Epan	20	4.50	0.854 ± 0.031
Single	APACHE II, admission type	Epan	20	0.04	0.828 ± 0.029
Single	APACHE II, diagnosis category	Unif	20	-4.22	0.831 ± 0.029
Single	APACHE II, adm. type, diag. category	Unif	20	-6.70	0.818 ± 0.029
Dual	APACHE II or alternative features	Epan	100	-11.90	0.818 ± 0.033
Dual	SAPS II or alternative features	Epan	50	-1.07	0.854 ± 0.030
LR model	APACHE II	-	-	-	0.796 ± 0.033
LR model	SAPS II	-	-	-	0.867 ± 0.027

Table 1: Results from the internal validation (jackknife cross-validation on the training set, 1559 ICU admissions). The predictive bias, averaged over all cases, is expressed as a percentage of the hospital mortality (14.8%). The alternative features for the dual method estimator are minimum temperature, minimum systolic blood pressure, minimum bicarbonate, and maximum creatinine values during the first 24h of ICU stay.

fore the predictions may be more accurate. Furthermore, in this way the prediction method accommodates to changes in the population characteristics (*drift*), a phenomenon that frequently occurs in medical applications.

In each validation procedure we computed the area under the ROC curve (AUC) for all IBR estimators. The AUC quantifies a estimator's ability to discriminate between patients who survive and those who die. An AUC value of 0.5 indicates that the estimator does not predict better than chance, while an AUC value of 1 indicates perfect discrimination. For the APACHE II and SAPS II scoring systems an AUC of > 0.80 is considered to be good.

4 Results

4.1 Internal validation

Table 1 shows the results from the internal validation. When regarding the AUCs, we see that the SAPS II IBR estimator is superior to the APACHE II IBR estimator (0.860 vs. 0.792). The LR model of SAPS II is better than that of APACHE II (0.867 vs. 0.796), and the SAPS II IBR estimator. The multivariate IBR estimator that uses both scores yields a slightly worse performance than SAPS II alone (0.860 vs 0.854) but these differences have not been tested for significance.

The APACHE II LR model employs information on the patient's diagnosis and type of admission besides the score, so employing this information with the IBR estimator should lead to better results as well. This is done by combining the APACHE II score and the admission type and/or diagnosis category in the IBR estimator. We see in Table 1 that adding either APACHE II admission type or diagnosis category leads to a increase in performance compared to that of the APACHE II alone in the IBR estimator. The performance is slightly worse when both the admission type and diagnosis category are used.

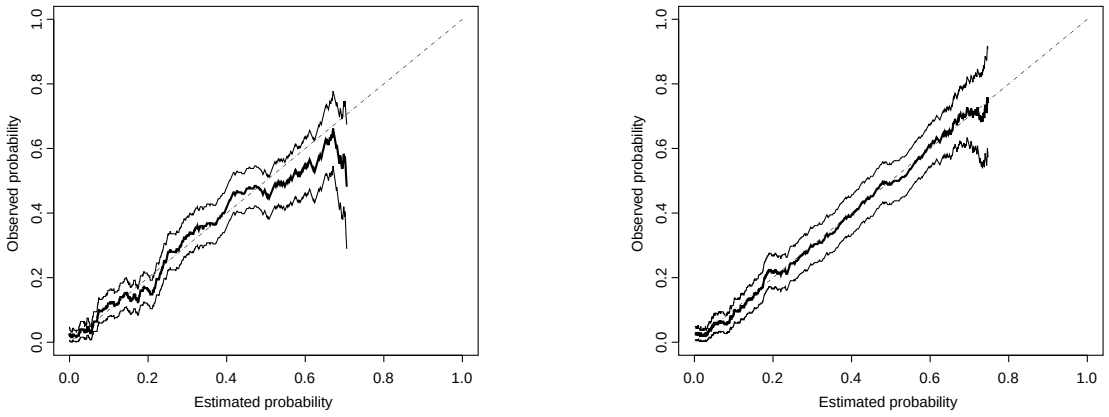
Since predictions for cardiac surgery patients by the APACHE II and SAPS II LR models are believed to be unreliable, the predictions by the single method IBR estimator may be unreliable as well. The dual method estimator attempts to improve performance by using alternative features for these patients. The desired effect is however only obtained for the APACHE II score and not for SAPS.

Table 1 also shows that the uniform kernel and Epanechnikov kernel were almost equally often selected by the optimization algorithm. So, the uniform kernel (i.e., equal weight for all instances in the neighborhood) may perform equally well or better than the Epanechnikov kernel in practical circumstances, even though the Epanechnikov kernel appears to be superior from a theoretical point of view.

Interestingly, the optimization algorithm has chosen `ttw` values (i.e., effective neighborhood sizes) that are relatively large compared to the values that are usually reported in the literature (less than 20 neighbors is common). Presumably, the explanation is that in our application, the neighboring outcomes are used to estimate the probability of death instead of the dominant class, and therefore a larger neighborhood size is required. Note that lower `ttw` values are selected for the multivariate estimators, due to sparsity in the multidimensional feature space of these estimators.

Because k-NN regression does not optimize a global likelihood formula, its predictions may show structural deviations from the observed outcome; we refer to this phenomenon as *predictive bias*. In Table 1 we have listed the predictive bias of each of the estimators, expressed as a percentage of the observed outcome. The APACHE II IBR estimator (first row), for instance, predicts a total of 221.1 deaths, whereas 230 out of 1559 patients actually died; the estimator thus underestimates mortality with 3.85%. The dual method estimator for APACHE II (seventh row) has a serious negative bias of -11.9% (202.6 deaths predicted).

Figure 2 shows smoothing plots of observed versus predicted probabilities for the APACHE II and SAPS II IBR estimators. The plots illustrate well the superior fit of the SAPS II estimator to the data: its plot is far more smooth and extends further into the upper region of the probability interval. The APACHE II plot, in contrast, is rather bumpy and the estimator appears to perform very poorly for patients with a high score. So, the APACHE II score appears to contain 'errors' that are difficult to repair, even for a highly adaptive method such as k-NN regression.



(a) (b)

Figure 2: Observed vs. predicted probabilities in the APACHE II (a) and SAPS II (b) IBR estimators. The observed probabilities (on the y-axis) are obtained by loess smoothing on the observed outcome values (0 and 1), and are surrounded by 95% confidence intervals.

4.2 Prospective validation

Table 2 shows the results of all prospective validations. For each prospective validation, the mean predictive bias and AUC are displayed.

In the settings validation, the IBR estimators are applied to new data with the kernel type and target total weight settings that were optimized on the training set. For all estimators, the performance is equally good or better on the test set (explained by the fact that this set is somewhat larger than the training set). We conclude that the settings that optimized on the training set generalize well to new data.

Also in the plain prospective validation, where instances from the training set are used to make predictions on the test set, the performance is similar to the internal performance on the training set. So, we can use the IBR estimators to make predictions for future, unseen cases. The predictive bias, however, increases.

In the incremental prospective validation, the performance of estimators based on the APACHE II score further increases. Apparently, these estimators take advantage of the increasing size of the instance base. This does not hold for the estimators based on SAPS II. Furthermore, the predictive bias now reduces. An explanation for the latter fact is that the feature space becomes more densely populated since instances are added. A denser population means that the neighborhood does not have to expand as much as with a sparse population. This is especially advantageous near the boundaries, where the predictive bias is usually larger.

5 Discussion and conclusion

We have used IBR to predict hospital mortality for patients admitted to the ICU. Comparing our study to other applications of IBR in medicine, we note that in most studies IBR is used for classification (e.g. [Schmidt and Gierl, 2005; Lopez and Plaza, 1997]) and only sparsely for prediction. Anand et al. [Anand *et al.*, 2001] use k-NN in a hybrid system to predict time to survival in cancer patients, Gottrup et al. [Gottrup *et al.*, 2005] predict infarcted regions of the brain after cerebral stroke, based on MRI scans. Often IBR is used as part of a larger system, e.g. as in [Montani *et al.*, 2000].

From our experiments we conclude that IBR can be used to make reliable prognoses from clinical scores, and is competitive to LR in this task. For the APACHE II score, IBR prediction even outperforms the LR model. The applied method has been shown to generalize well to future patients, especially when new patients are added to the instance base to compensate for drift in the population characteristics.

When comparing the performance of the APACHE II and SAPS II scores in the IBR algorithm, we see that the SAPS II score performs better than the APACHE II score. The SAPS II scoring system was developed by scaling the coefficients that were obtained with multiple LR analysis. In contrast, the APACHE II scoring system is based on expert knowledge and the associated prognostic model was obtained from a LR analysis. This may be the reason that the IBR estimator does not perform better than the SAPS II LR model. These different approaches (expert knowledge vs. multiple LR analysis) to the development of a scoring system appears to be an important factor in the performance of IBR compared to a LR model. We think that this difference may also be apparent in other medical domains.

In the multivariate IBR estimators, we have used the Manhattan distance metric. Euclidean distance or other more sophisticated metrics may lead to better results. Similarly, it may be beneficial to weigh the predictive features, instead of treating them as equally important. However, Kohavi et al. [Kohavi *et al.*, 1997] found that weighing features rapidly leads to overfitting. Furthermore, we note that these adjustments only affect the multivariate estimators, whereas very good results were obtained already with our univariate estimators.

In the multivariate experiments, the combination of APACHE II and SAPS II scores performed worse than the

Feature(s)	Settings validation		Plain prospective validation		Incr. prospective validation	
	Bias	AUC $\pm$ S.D.	Bias	AUC $\pm$ S.D.	Bias	AUC $\pm$ S.D.
APACHE II	-3.83	0.821 ± 0.026	-11.22	0.784 ± 0.029	-4.37	0.809 ± 0.027
SAPS II	-0.73	0.865 ± 0.024	-2.23	0.867 ± 0.024	0.26	0.867 ± 0.024
APACHE II, SAPS II	-3.30	0.869 ± 0.022	-3.02	0.861 ± 0.025	0.83	0.863 ± 0.024
APACHE II, admission type	-2.81	0.843 ± 0.024	-8.57	0.832 ± 0.025	-4.74	0.839 ± 0.024
APACHE II, diagnosis category	-3.01	0.840 ± 0.023	-7.60	0.829 ± 0.025	-4.88	0.835 ± 0.024
APACHE II, adm. type, diag. category	-13.78	0.826 ± 0.024	-7.88	0.828 ± 0.024	-5.64	0.831 ± 0.024
Dual method APACHE II	-12.12	0.818 ± 0.024	-16.95	0.812 ± 0.027	-11.82	0.834 ± 0.025
Dual method SAPS II	-1.52	0.863 ± 0.024	-1.15	0.870 ± 0.024	1.68	0.872 ± 0.023
APACHE II *LR model*	-	-	-	0.804 ± 0.027	-	0.804 ± 0.027
SAPS II *LR model*	-	-	-	0.877 ± 0.022	-	0.877 ± 0.022

Table 2: Results from the prospective validations (1868 ICU admissions). The hospital mortality in this dataset is 16.3% .

SAPS II score alone. A possible explanation is found in the fact that the distance metric regards these two scores on two independent axes, perpendicular to each other. This is not correct, because both scores indicate the severity of illness; they are collinear. In future experiments, we have planned to use local regression models [Cleveland, 1979], which is expected to adjust for this phenomenon.

Acknowledgement The authors wish to thank Ameen Abu-Hanna and Rob Bosman for valuable discussions on the work described. Niels Peek receives a grant from the Netherlands Organisation of Scientific Research (NWO) under project number 634.000.020.

References

[Anand *et al.*, 2001] S.S. Anand, P.W. Hamilton, and J.G. Hughes et al. On prognostic models, artificial intelligence and censored observations. *Methods Inf Med*, 40:18–24, 2001.

[Ash and Shwartz, 1999] A. Ash and M. Shwartz. R^2: a useful measure of model performance when predicting a dichotomous outcome. *Stat Med*, 18:375–84, 1999.

[Cleveland, 1979] W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc*, 74:829–836, 1979.

[Gottrup *et al.*, 2005] C. Gottrup, K. Thompson, and P. Locht et al. Applying instance-based techniques to prediction of final outcome in acute stroke. *Artif Intell Med*, 33:223–236, 2005.

[Gunning and Rowan, 1999] K. Gunning and K. Rowan. ABC of intensive care: Outcome data and scoring systems. *BMJ*, 319:241–4, 1999.

[Hastie *et al.*, 2001] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer, New York, 2001.

[Hosmer and Lemeshow, 2000] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, 2nd edition, 2000.

[Knaus *et al.*, 1985] W.A. Knaus, E.A. Draper, and D.P. Wagner et al. APACHE II: a severity of disease classification system. *Crit Care Med*, 13:818–29, 1985.

[Kohavi *et al.*, 1997] R. Kohavi, P. Langley, and Y. Yun. The utility of feature weighting in nearest-neighbor algorithms. In M. van Someren and G. Widmer, editors, *Proc. 9th Europ. Conf. on Machine Learning (ECML-97)*. Springer, Berlin, 1997.

[Le Gall *et al.*, 1993] J. Le Gall, S. Lemeshow, and F. Saulnier. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA*, 270:2957–63, 1993.

[Lopez and Plaza, 1997] B. Lopez and E. Plaza. Case-based learning of plans and goal states in medical domains. *Artif Intell Med*, 9:29–60, 1997.

[Montani *et al.*, 2000] S. Montani, R. Bellazzi, and L. Portiginale et al. Diabetic patients management exploiting case-based reasoning techniques. *Comput Methods Programs Biomed*, 62:205–218, 2000.

[NICE, 2005] NICE, 2005. National Intensive Care Evaluation (NICE) register. http://www.stichting-nice.nl, also in English, accessed June 20th, 2005.

[Schmidt and Gierl, 2005] R. Schmidt and L. Gierl. A prognostic model for temporal courses that combines temporal abstraction and case-based reasoning. *Int J Med Inform*, 74:307–315, 2005.

[Silverman, 1986] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[Tan, 2005] C.H.K. Tan. Instance-based prognosis in intensive care using severity-of-illness scores. Master's thesis, Faculty of Medicine, University of Amsterdam, 2005. Accesible through: http://dare.uva.nl/scriptie/159502.

[Verduijn, 2002] M. Verduijn. Prognostic tree models in cardiac surgery. Identifying interactions between risk factors in a process-oriented approach. Master's thesis, Faculty of Medicine, University of Amsterdam, 2002.

[Witten and Frank, 2001] I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques with JAVA implementation*. Morgan Kauffmann, San Francisco, 5th edition, 2001.

[Wyatt, 1990] J. Wyatt. Construction of clinical scoring systems. *BMJ*, 300:538–9, 1990.

The Predictive Value of Consecutive Event Episodes in the Intensive Care

Tudor Toma[1], Ameen Abu-Hanna[1], Robert-Jan Bosman[2]
[1] Academic Medical Center, Universiteit van Amsterdam,
Department of Medical Informatics, PO Box 22700, 1100 DE Amsterdam,
The Netherlands
[2] Department of Intensive Care, Onze Lieve Vrouwe
Gasthuis, 1e Oosterparkstraat 279, PO Box 10550, 1090 HM Amsterdam,
The Netherlands

Abstract

Prediction of patient mortality plays a central role in quality of care assessment programs in the Intensive Care (IC). Existing prognostic models for mortality prediction, such as SAPS, are logistic regression models based on medical scores calculated from data obtained only during the first 24 hours of patient admission. They, hence, do not use temporal information collected during the IC stay. In this paper we investigate the added value of the daily sequential organ failure (SOFA) score on predicting mortality after the first, second, and third days after admission. In particular, we first use data-mining techniques to discover frequent patterns in the sequence of SOFA scores. Then, we assess their added value in mortality prediction by considering them as potential covariates in a logistic regression model that includes the SAPS score. We demonstrate that this method results in better models and validate this result on a new test set.

1. Introduction

Prognosis, or the prediction of medical outcomes, plays an important role in Medicine [Abu-Hanna and Lucas, 2001]. In the Intensive Care, hospital mortality is predicted and compared to the actual mortality in order to assess the quality of care of an IC unit. Hospital mortality comprises deaths in the hospital during or after the stay in IC unit (ICU). Current IC prognostic models, such as the SAPS-II [Gall et al., 1993] are logistic regression models that use a relatively small number of severity-of-illness scores as their covariates. These scores are based on data collected during the first 24h after admission. They take positive integer values where higher values indicate more severe conditions of the patient. Recently, modern ICUs started collecting a daily score called the SOFA (Sequential Organ Failure) score [Vincent and Ferreira, 2000], which ranges from 0 to 24. The SOFA score quantifies the degree of derangement and failures of organ systems for each patient. The SOFA score was hence not developed specifically for its utility to help predict mortality but it is believed that its temporal information can contribute to this purpose.

In this paper we describe and evaluate a method for including patterns from SOFA sequences for developing prognostic models for the prediction of hospital mortality. More specifically, we construct three models that predict mortality after one, two, and three days after ICU admission. In general, there are two main approaches for the utilization of temporal data. In the first approach the temporal data is reduced by means of summary statistics such as the maximum value in a sequence. In the second approach, which we shall adopt here, the temporal relationship between values is preserved. We will represent temporal information by patterns of frequently occurring consecutive qualitative values. These patterns will be represented as binary covariates in the logistic regression models. In order to assess the added value of these patterns to existing models, we use the SAPS-II score as a permanent covariate. In our model-building strategy, covariates representing the patterns are added only if they show improvements of statistical significance. We then validate these models on a new test set.

The following section will include a brief description of the data used. Section 3 will describe the methods, their applicability and the results. Discussion and conclusions are presented in Section 4.

2. Data

Data was provided by the ICU of the OLVG teaching hospital in Amsterdam and contains information about all 5160 IC patients from July 1998 until December 2004. The data set contains static and temporal information. The static information includes more than 100 attributes for each patient including demographics, like age and sex; reason for admission to the ICU, like surgery; and physiology, like body temperature and heart rate. Severity of illness scores, most notably the SAPS-II score, are also part of the static data. All static data are collected in the first 24h after admission. The temporal data contains SOFA scores, ranging from 0 to 24, collected on daily basis for each patient. Hence, each patient has a temporal sequence of SOFA scores computed for each day of stay of the patient in the ICU. For example the sequence *14 - 12 - 8* describes the SOFA scores of a patient that stayed 3 days and is recovering. For patients that were readmitted in the ICU, e.g. because of complications, only the last readmission was included due to its relevancy to hospital mortality, which is the outcome one wants to predict. This has resulted in a total of 4771 patients.
Because our dataset is relatively large we randomly divided the data set into a training set (3181 patients) for model development and a test set (1590 patients) for model

validation. Some main characteristics of the patients in the data sets are shown in Table 1.

	Data set	Training set	Test set
No. of patients	4771	3181	1590
#Males / #Females	1.90	1.86	1.97
Age mean ± SD	64.4 ± 14.41	64.7 ± 14.21	63.7 ± 14.79
Age median	67	68	66
SAPS mean ± SD	34.2 ± 15.6	34 ± 15.3	34.4 ± 16.3
Hosp. mortality %	10.7	10.3	11.6
Length of stay median	0.92	0.92	0.92
Length of stay mean	2.11	2.13	2.08

Table 1. *Descriptive statistics for the whole data, training and test sets.*

3. Methods and results

3.1. Data preprocessing

We categorize the SOFA score, which ranges form 0 to 24, into manageable more intuitive qualitative categories. We derive these categories automatically, from the training set, by seeking the cut-off points that minimize the entropy of mortality. This is achieved by fitting a binary classification tree for predicting mortality using the maximum SOFA values for each patient during their length of stay in ICU. The tree is pruned according to minimizing the 10-fold cross-validation error. We obtained 2 cutoff points corresponding to three SOFA categories: L (Low) for SOFA Ò 10, M (Medium) with 11 Ò SOFA Ò 13 and H (High) for SOFA Ó 14. The SOFA scores are now recoded, in the training as well as the test set, according to these categories. A SOFA score sequence of *14 - 12 – 8* will be now recorded qualitatively as H-M-L.

3.2. Frequent episodes

A sequential episode [Mannila et al., 1997], is a form of a temporal pattern that specifies a set of events that occurs in a particular order in a pre-set time window. The sequential episode A-B matches the sequences A-B and A-C-D-B. An event in our case is either L, M, or H. Analysis of non-consecutive sequential episodes has been described in [Toma et al., 2005]. In this paper we only consider episodes in which the events are also consecutive, so in the example above the sequence A-C-D-B would not match the episode A-B. An Apriori-like [Agrawal and Srikant, 1994] efficient algorithm for discovery of sequential episodes is described in [Mannila et al., 1997]. We adapted and implemented this algorithm in Java. The restriction to consecutive events in the sequence is implemented in practice by constraining the window size. Our main adaptation to the algorithm in [Mannila et al., 1997] is related to the way the support of episodes is calculated. In the ICU, patients have markedly different lengths of stay (regardless of their vital status at discharge), and hence, different lengths of qualitative SOFA sequences (hereafter SOFA sequences). If we calculate support of an episode based on its occurrence in the whole patient population, then the longer episodes will have much less support because the longer sequences are much less frequent. Because we want to make predictions for three different cohorts -for those that stayed for at least 1, 2 and 3 days- we must adjust the support based on the respective cohort. For example, when calculating the support for an episode of length 3 we should seek support within sequences of at least length 3. In other words, only patients that stayed for at least 3 days are eligible to be counted in the denominator of the support. This allows longer sequences to emerge in the set of frequent episodes without being negatively biased by the existence of shorter sequences. We require a minimum support of 5% for an episode to be considered frequent.

Applying the frequent episodes discovery algorithm on the data generated 101 frequent episodes. Short episodes have less possible realizations but enjoy high support. Table 2 presents a selection of episodes and their identifiers, which we will encounter below in model development.

Identifier	**S19**	**S34**	**S61**	**S33**	**S3**	**S73**
Temporal sequence	L	L - L	H - H	L-L-L	H-H-H	M- L

Table 2. *Episodes and their identifiers*

In Table 3 the distribution of the frequent episodes over different possible lengths is presented. We discover episodes of maximum length 7 although in the current setup only patterns of length maximum 3 are used.

Length	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
#episodes	**3**	**7**	**13**	**18**	**31**	**19**	**10**	**101**

Table 3. *Distribution of the frequent episodes over length.*

3.3. Logistic regression models

A logistic regression model [Hosmer and Lemeshow, 1989] is a parametric model that specifies the conditional probability of a binary variable *Y ({0, 1})* to have the value 1, given the values of the covariates of the model. Y = 1 indicates the occurrence of an event such as death in our concrete case. The logistic model has the following form:

$$p(Y \mid 1 \mid x) \mid \frac{e^{g(x)}}{1 2 e^{g(x)}} \quad (1)$$

where $x \mid (x_1,...,x_m)$ is the covariate vector. For *m* variables (also called predictors) the *logit function* g(x) has the following form:

$$g(x) \mid \eta_0 2 \sum_{i \mid 1}^{m} \eta_i x_i \quad (2)$$

where η_i, i=1,...,m, denote the coefficients of the *m* predictors. One reason for the popularity of the model is the interpretation that is given to each η_i in terms of an *odds ratio*. Suppose the logit function is η_0, + $\eta_1 sex$ + $\eta_2 age$ where *sex* = 1 for males and 0 for females, and age is calculated in years. The odds of dying for males, *odds*(*sex*=1), is *P(Y=1|sex=1)/P(Y=0|sex=1)* and for females, *odds*(*sex*=0), is *P(Y=1|sex=0)/P(Y=0|sex=0).*

The quantity e^{η_1} turns out to be equal to the odds ratio *odds*(*sex*=1)/*odds*(*sex*=0). If there is no difference between the odds for dying for males compared to females, assuming all other variables (in this case only age) have the same values, the odds ratio will be 1. A higher value indicates higher risk to die for males, and a lower value than 1 indicates higher risk for females. The interpretation of e^{η_2} is similar, it indicates the odds ratio between a group of patients who are one unit, here one year, older than the other group. We will use the episode identifiers as binary dummy variables, like the variable sex above.

The motivation for using logistic regression as our formalism of choice is two-fold. First, it is the most popular model used in medicine for binary classification problems, and is also the *de facto* model used in quality assessment programs in the ICU. Second, the η_is take into account the dependence between the variables used, so the value of any η_i is adjusted for all other variables in the model and there is no assumption that the variables are independent.

Our idea in developing logistic-based prognostic models it to use a dummy (sometimes called design) variable for each episode. Each patient will have a vector of dummy variables, one for each episode. In our case every dummy variable will only have two values: 0, indicating the episode does not match the SOFA sequence of a patient, and 1 if it does match. When predicting mortality for patients on the k^{th} day, one may only use episodes with maximum length of *k*. For example, to predict mortality for patients on the third day of their stay, we only consider patients who stayed for at least three days. That is, we use episodes of a maximum length of 3 which are matched against the SOFA sequences of the first three days. In total we will hence have 3 models to predict mortality on each of the first 3 days. Prediction in these days is relevant for the ICU. Note that the number of patients considered for prediction on day 1 includes all patients, and that this number decreases with each day due to patient discharge (regardless of survival status). To assess the *added* value of the temporal patterns to current logistic models, we develop for each day a model with only the SAPS score as a covariate, and we compare this model with a temporal one in which also dummy variables representing the existence of the episodes are included.

The strategy for fitting the temporal models is as follows. For any given day for prediction, the best dummy variables are included by a simple hill-climbing search process. We first fit a model including only the intercept, η_0, and the term for the SAPS score, η_1SAPS. This is the reference model. We then assess the inclusion of each dummy variable to this model by the log-likelihood test [Hosmer and Lemeshow, 1989]. The dummy variable with the most significant p-value, as long as it is Ò 0.05, is included in the model. We reiterate this process till a maximum of 4 covariates. This restriction is meant to keep the models manageable and also to combat over-fitting.

The last 2 columns in Table 4 show, for each one of the three temporal models, the frequent episodes whose dummy variables were selected in the logistic regression and the corresponding η_i. For example the first temporal model for predicting mortality after observing the first day is:

$$P(Y \mid 1 \mid SAPS, S19) \mid \frac{e^{45.042 0.09 SAPS 41.47 S19}}{12\, e^{45.042 0.09 SAPS 41.47 S19}} \quad (3)$$

When inspecting the selected episodes we note that the models include episodes with consistently high or low values like *H-H* or *L-L-L*. Also one model includes an episode that varies over time like *M-L* which apparently captures trends influential to mortality.

Test set #patients	**SAPS model**		**Temporal models**		**A-temporal+ Temporal covariates**	**η**
	Log	**Brier**	**Log**	**Brier**		
1576 LOS Ó1 day	356.4333	0.1316	348.2468	0.1272	Intercept + SAPS + S19	-5.04 0.09 -1.47
504 LOS Ó2 days	222.2785	0.2838	215.7803	0.273	Intercept + SAPS + S34 + S61	-3.89 0.06 -0.86 2.22
299 LOS Ó3 days	164.9534	0.3708	160.0427	0.356	Intercept + SAPS + S33 + S3 + S73	-2.02 0.04 -1.57 6.22 -0.69

Table 4. *Comparative prediction model performance: SAPS models versus temporal models for day 1, 2 and 3 on the test set. LOS denotes length of stay in the ICU.*

The correspondence between the dummy variables in the temporal models and the temporal are shown in Table 2.

3.4. Validation After model selection we validated the models on the test set. Our performance measures include the Brier score which is:

$$\frac{1}{N}\sum_{i|1}^{N}(P(Y_i \mid 1 \mid x_i) 4\, y_i)^2$$

where N denotes the number of patients, Y_i the random variable of the predicted outcome, and y_i the value of the actual outcome. Also we use the logarithmic score:

$$\sum_{i|1}^{N} LS_i \text{ where } LS_i \mid 4 \ln P(Y_i \mid y_i \mid x_i).$$

Lower values mean better performance for both scores. These performance scores penalize models when they do not provide the true probability, and are more appropriate than purely discriminating measures such as error rate and the area under the ROC curve which might mask under- and over-prediction (see discussion in [Abu-Hanna and Keizer, 2003]). The performance of each temporal model is then compared to its respective reference model on the same part of the test set.

The validation results are presented in the columns labeled ``Log" and ``Brier" in Table 4. Similar to partitioning the training set, three test sets were used having patients staying at least one day for the first data set until at least three days for the third data set. All the temporal models 1, 2 and 3 outperform the reference model based on SAPS alone. This resembles the results from [Toma et al., 2005] where non-consecutive sequential episodes have been considered. We are planning to statistically test the significance of the differences in the Brier score as described in [Redelmeier et al., 1991]. We have however extended the analysis to day 4 and day 5 as well and again the temporal models in these days have outperformed the respective static ones. Hence, in total the temporal approach, as a method, has outperformed the static approach in 5 out of 5 experiments. This provides evidence for the utility of the temporal approach although one should keep in mind that the experiments are not independent. We note however that the consecutive episodes described here have slightly better performance than the non-consecutive episodes.

4. Discussion and Conclusion

In this section we reflect critically on our approach and results, draw conclusions, and provide context and an outlook for further research.

4.1. Discussion on methods and results

The entropy-based categorization method using maximum SOFA score per patient resulted in three categories. Further inspection of these categories shows that these make clinical sense as they correspond to three groups of patients with distinctively different number of multiple organ failures as can be calculated from the mean of the 6 sub-scores (an organ is failing when its SOFA sub-score -which ranges from 0 to 4- is 3 or 4). Alternatives to the usage of the maximum SOFA score include using the last SOFA value in a patient's sequence or the mean in the last 3 days. Further analysis showed that our choice for the maximum value for each patient is quite robust as the cut-off points hardly changed if the procedure is applied on different random training sets. The current cut-off points were obtained on a “frozen” version of the training set.

The frequent consecutive-event episodes that were discovered have an intuitive clinical meaning in terms of improvement or worsening of a patient's condition. When an episode appears statistically significant to be included in a temporal model, the analyst can still judge how much it makes clinical sense by inspecting its corresponding η (either positive or negative). There seems to be a preference for selecting episodes that are as long as the number of days under consideration. A number of reasons could be responsible for this. First, we do not explicitly include trends in the episodes and hence more information in the episode is required. Second, we do not require alignment of an episode to the day of prediction, longer episodes tend to be aligned or at least closer to the last day of prediction.

The results obtained are a proof of concept that consecutive SOFA score episodes, based on our categorization, is beneficial indeed and has an added value compared to static models. Note that this added value is inherent in the patterns themselves and not because the models are developed on cohorts with different lengths of stay: the static ones have been fitted separately for each of these cohorts as well.

4.2. Existing and future work

Our approach aims at understanding the merits of temporal information in prediction and thus the advancement of the state of the art in IC prediction by including the temporal content of the recently developed SOFA scores. With the exception of the work described in [Kayaalp et al. 2000] and [Kayaalp et al., 2001] all approaches known to us that use SOFA scores for prediction, e.g. [Kajdacsy-Balla Amaral et al., 2005], reduce the temporal information into few summary statistics and use them in prediction. In our approach, the temporality of the events is preserved.

The work described in [Kayaalp et al., 2000] and [Kayaalp et al., 2001] investigates SOFA temporal patterns and share important elements with our work. In terms of [Kayaalp et al., 2000], we adopt the stationarity assumption of the process (generating the SOFA scores) in the sense that the frequency of the episodes is calculated independently of when it occurred in time.

In [Kayaalp et al., 2001] temporal patterns using SOFA sub-scores are calculated from the data, then integrated in a Naive-Bayes framework. In our approach the, discovered frequent sequential episodes, based on an adaptation of the algorithm in [Mannila et al., 1997], are integrated in a logistic regression framework. This not only allows for integrating new methods into the established framework in IC prediction, but also takes into account the possible inter-dependence between the sequential episodes, unlike the Naive Bayes approach which assumes conditional independence of patterns. In addition, our approach provides an intuitive interpretation of significance of these episodes, by means of the ηs, by which the episodes can be judged. Another difference is the way models are validated. We use the Brier score and the logarithmic score instead of the area under the ROC curve used there because, in quality assessment programs, it is important to measure discrimination and precision in combination, instead of relying on only the discrimination power of the model.

Unlike in our approach, in [Kayaalp et al., 2001] patterns are always considered backwards from the day of the prediction. These properties might be beneficial and, as future research, we plan to investigate whether they would lead to improvements of our temporal models. Another future research is investigating new temporal event types. The intensivists hypothesize that dealing with the notion of recovery, instead of the SOFA scores themselves, might be useful. For example the sequence could reflect the recovery, or recovery rate, in time. Other possible improvements include more sophisticated search strategies for fitting the logistic regression temporal models.

The idea of integrating logistic regression with other formalisms as we have done here, seems to bear fruit. In [Abu-Hanna and Keizer, 2003] logistic regression models have been

fit to patient sub-groups implied by a decision tree partitioning of the data. Another interesting way to use logistic regression is integrating it during the episode discovery stage: an episode is assessed not only by its frequency but also by its predictive power according to a logistic regression model.

Acknowledgment This research was supported by the Netherlands Organization for Scientific Research (NWO) under the I-Catcher project number 634.000.020. We thank Arno Siebes, Niels Peek and Manuel Campos for their helpful discussions on the topics described in this paper.

Reference:

[Abu-Hanna and Lucas, 2001] A. Abu-Hanna, and P.J.F. Lucas, Editorial: Prognostic models in medicine - AI and statistical approaches, Methods of Information in Medicine **40** (2001) 1—5

[Abu-Hanna and Keizer, 2003] A. Abu-Hanna, and N. de Keizer, Integrating classification trees with local logistic regression in Intensive Care prognosis, Artificial Intelligence in Medicine **29(1-2)** (2003) 5Ь23

[Agrawal and Srikant, 1994] R. Agrawal, and S. Srikant, Fast algorithms for mining association rules, In Proc. of the 20th VLDB Conf., (1994) 487Ь499

[Hosmer and Lemeshow, 1989] D.W. Hosmer, and S. Lemeshow, Applied logistic regression. New York: John Wiley \& Sons, Inc. (1989)

[Kayaalp et al., 2000] M. Kayaalp, G. F. Cooper, and G. Clermont, Predicting ICU mortality: a comparison of stationary and constationary temporal models, Proc. of AMIA, (2000) 418Ь422

[Kayaalp et al., 2001] M. Kayaalp, G F. Cooper, and G. Clermont, Predicting with variables constructed from temporal sequences, Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics. (2001) 220-–225.

[Kajdacsy-Balla Amaral et al., 2005] A.C. Kajdacsy-Balla Amaral, F.M. Andrade, R. Moreno, A. Artigas, F. Cantraine, and J.L Vincent, Use of Sequential Organ Failure Assesment score as a severity score, Intensive Care Med **31** (2005) 243—249

[Gall et al., 1993] J. Le Gall, S. Lemeshow, and F. Saulnier, A new Simplified Acute Physiology Score (SAPS-II) based on a European/North American multicenter study. *JAMA* **270** (1993) 2957 Ь 2963.

[Mannila et al., 1997] H. Mannila, H. Toivonen, and A.I. Verkamo, Discovering frequent episodes in sequences, Data Min. Knowl. Discov., **1(3)** (1997) 259—289

[Vincent and Ferreira, 2000] J.L. Vincent, F.L. Ferreira, Evaluation of organ failure: we are making progress, Intensive Care Med **26** (2000) 1023—1024

[Toma et al., 2005] T. Toma, A. Abu-Hanna, and R. J. Bosman, Predicting mortality in the Intensive Care using episodes, IWINAC, LNCS **3561** (2005), 447

[Redelmeier et al., 1991] D.A. Redelmeier, D.A. Blonch, and D.H. Hickam, Assessing predictive accuracy: How to compare Brier scores, Journal of Clinical Epidemiology **44** (1991) 1141Ь1146

A Dynamic Bayesian Network for Diagnosing Ventilator-Associated Pneumonia in ICU Patients

Theodore Charitos, Linda C. van der Gaag,
Inst. of Inform. and Comp. Sciences
Utrecht University
P.O Box 80.089
3508 TB Utrecht, The Netherlands
theodore, linda@cs.uu.nl

Stefan Visscher, Karin Schurink,
Dept. of Internal Medic. and Infect. Diseases
University Medical Centre Utrecht
Heidelberglaan 100
3584 CX Utrecht, The Netherlands
S.Visscher, K.Schurink@azu.nl

Peter Lucas
Inst. for Comp. and Inform. Sciences
Radboud University, Nijmegen
Toernooiveld 1
6525 ED Nijmegen, The Netherlands
peterl@cs.ru.nl

Abstract

Diagnosing ventilator-associated pneumonia in mechanically ventilated patients in intensive care units is currently seen as a clinical challenge. The difficulty in diagnosing ventilator-associated pneumonia stems from the lack of a simple yet accurate diagnostic test. To assist clinicians in diagnosing and treating patients with pneumonia, a decision-theoretic network was designed with the help of domain experts. A major limitation of this network is its inability to represent pneumonia as a dynamic process that progresses over time. In this paper, we construct a dynamic Bayesian network that explicitly captures the development of the disease through time. We discuss how probability elicitation from domain experts serves to quantify the dynamics involved and show how the nature of patient data helps reduce the computational burden of inference. We evaluate the diagnostic performance of our dynamic model and report promising results.

1 Introduction

Many patients admitted to an intensive care unit (ICU) need respiratory support by a mechanical ventilator; in addition, many of these patients are affected by severe disease which may result in depression of their immune system. Both conditions promote the development of ventilator-associated pneumonia (VAP) in these patients. Because of the wide-spread dissemination of multiresistant bacteria at the ICU, effective and fast treatment of VAP is seen as an issue of major significance. The difficulty of the diagnosis of VAP is in the lack of a gold standard; VAP is therefore diagnosed by taking a number of different clinical features into account [7].

A probabilistic and decision-theoretic network [3], representing the uncertainties and preferences involved in dealing with the treatment of VAP, was constructed by Lucas et al. [4]. The network was developed with the help of two infectious disease experts, who assessed both its qualitative structure and its numerical part. The goal of the network was to prescribe an optimal antimicrobial therapy, and thereby assist clinicians in treating patients with VAP.

A prominent role in the domain of pneumonia is played by two stochastic processes: the *colonisation* of the laryngotracheobronchial tree by pathogens and the onset and development of *pneumonia*. Although both processes evolve dynamically, these dynamics were not explicitly modelled by means of temporal transitions in the network described above. Instead, the dynamics of the processes were implicitly modelled by additional interactions between the duration of stay and the duration of mechanical ventilation of a patient with the colonisation by pathogens. The main motivation for this simplification was the large amount of data needed to specify the probability distribution underlying the stochastic processes and the increase in computational requirements. The network thus constitutes a *static* simplification of the domain which obscures its dynamic nature. In fact, the static network was used for every patient for each day on the ICU separately, without taking into account the patient's characteristics from earlier days. As the development of VAP is a dynamic process, we need to model time in a more explicit way to improve the diagnosis.

In this paper, we ameliorate the problems related with having modelled VAP as a dynamic process. We develop a dynamic Bayesian network that explicitly captures the temporal relationships between the variables [5]; our focus thereby is on the diagnostic part of the network. We use the method of Van der Gaag et al. [9], for the elicitation, from domain experts, of the probability distribution of the underlying stochastic process. This method transcribes probabilities and uses a scale with both numerical and verbal anchors that assists experts to assess many probabilities in little time. Moreover, we discuss how the computational burden of inference in our model can be eased by exploiting the nature of the observations involved, with just a small loss in accuracy [2].

We evaluated our dynamic network on a group of patients, drawn from the files of the ICU of the University Medical Centre Utrecht in the Netherlands. Our results indicate that the dynamic model is capable of distinguishing between patients with VAP and without VAP. By exploiting all available past information of a patient, it in fact yields better predictions than the static model. This occurs specifically for patients without VAP, for whom we notice that the use of previous information leads to much lower estimates for VAP than the ones obtained from the static network.

The remainder of this paper is organised as follows. In the next section, we briefly describe the static probabilistic and decision-theoretic network that had been developed be-

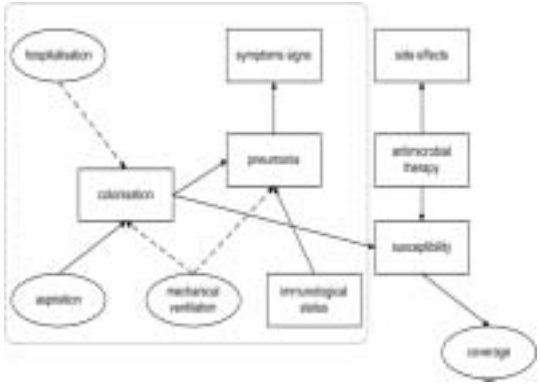


Figure 1: Global structure of the sVAP network. The dashed box indicates the network's diagnostic part.

fore for the management of patients with VAP. In Section 3, we discuss the construction of a dynamic network for VAP. Section 4 presents the results of an experimental evaluation of our network. The paper ends with our conclusions and directions for further research in Section 5.

2 Pathophysiology of VAP

Ventilator-associated pneumonia is a low-prevalence disease occurring in mechanically-ventilated patients in critical care and involves infection of the lower respiratory tract [1]. In contrast to infections of more frequently involved organs (such as the urinary tract), for which mortality is low, ranging from 1 to 4%, the mortality rate for VAP ranges from 24 to 50% and can reach 76% for some high-risk pathogens. Important factors related to the development of VAP include an increased *body temperature*, the use of *antipyretic drugs*, an abnormal amount of coloured *sputum*, *signs* on the chest X-ray, an abnormal ratio between the amount of oxygen in the arterial blood and the fractional inspired oxygen concentration, that is, pO_2/FiO_2, the duration of *mechanical ventilation*, and an abnormal number of *leucocytes*. As diagnosing a disorder in medicine involves reasoning with uncertainty, a decision-theoretic network was constructed as part of a decision-support system to assist clinicians in the diagnosis and treatment of VAP in the ICU [4],[7]. Figure 1 illustrates the network, which we call the static VAP network, os sVAP network for short. The signs and symptoms included in the sVAP network are shown in more detail in Figure 2.

The relationship between the *colonisation* by pathogens and the development of *pneumonia* is captured in the sVAP network as follows. Periodically, a sample of the patient's sputum is cultured at the laboratory. When the culture shows a number of colonies of a particular bacterium that is above a certain threshold, the patient is said to be colonised by this bacterium. The seven groups of microorganisms that occur most frequently in critically ill patients and cause colonisation, are modelled in the therapeutic part of the network. Figure 3 depicts the probabilistic relation between the seven groups of microorganisms of colonisation to pneumonia. Information about which bacterium or bacteria are currently present in a patient and the current signs

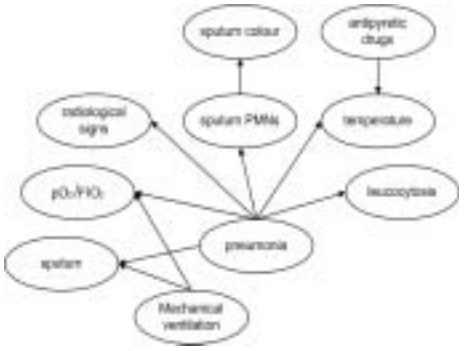


Figure 2: Symptoms and signs of pneumonia.

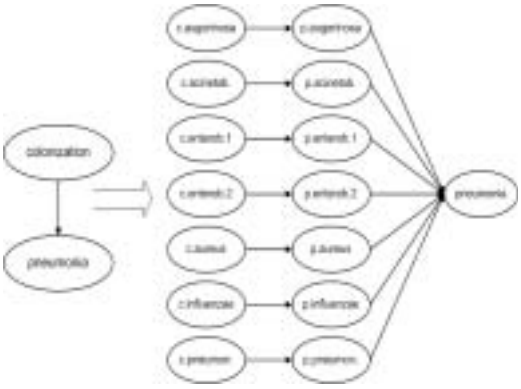


Figure 3: Detailed structure of the influence of colonisation on pneumonia.

and symptoms constitute the basis for choosing optimal antimicrobial treatment on multi-resistant bacteria and is considered best practice.

3 A dynamic Bayesian network for VAP

In this section, we describe the construction of a dynamic Bayesian network that represents explicitly the development of pneumonia. In addition, we address the computational burden of inference with the network.

3.1 Preliminaries

A dynamic Bayesian network is a graphical model that encodes a joint probability distribution on a set of stochastic variables, explicitly capturing the temporal relationships between them. More formally, let $\mathcal{V}_n = (V_n^1, \ldots, V_n^m), m \geq 1$, denote the set of variables at time step n. Then, a dynamic Bayesian network is a tuple (B_1, B_2), where B_1 is a Bayesian network that represents the prior distribution for the variables at the first time step $\mathcal{V}_1$, and B_2 defines the transitional relationships between the variables in two consecutive time steps, so that for every $n \geq 2$

$$p(\mathcal{V}_n \mid \mathcal{V}_{n-1}) = \prod_{i=1}^{m} p(V_n^i \mid \pi(V_n^i))$$

where $\pi(V_n^i)$ denotes the set of parents of V_n^i, for $i = 1, \ldots, m$.

We distinguish between two types of relationship in a dynamic Bayesian network: *transitional* relations that capture a dependence among variables between different time steps, and *local* relations that capture a dependence between variables within the same time step. If a relationship exists between the same variable over different time steps, this variable is called *persistent*. Based on the two types of relationship, per time step, the set of variables $\mathcal{V}_n$ is split into three mutually exclusive and collectively exhaustive sets $\mathcal{I}_n, \mathcal{X}_n, \mathcal{Y}_n$, where the sets $\mathcal{I}_n, \mathcal{Y}_n$ constitute the input and output variables and $\mathcal{X}_n$ consists of the hidden variables for the time step under study. Usually, $\mathcal{I}_n$ includes observable variables that affect the probability distribution of $\mathcal{X}_n$, while $\mathcal{Y}_n$ includes observable variables whose probability distribution is affected by $\mathcal{X}_n$. The set $\mathcal{X}_n$ includes the variables that represent the stochastic processes of the network and whose values are never observed. Later in the paper, we will need the notion of *forward interface* of a dynamic network, which is the set of variables at time step n that affect some variables at time step $n+1$.

Dynamic Bayesian networks are usually assumed to be time invariant, which means that the topology and the parameters of the network per time step and across time steps do not change. Moreover, the Markov property for transitional dependence is assumed, which means that $\pi(V_n^i)$ can include variables either from the same time step n or from the previous step $n-1$, but not from earlier time steps [5]. Then, by unrolling B_2 for N time steps, a joint probability distribution $p(\mathcal{V}_1, \ldots, \mathcal{V}_N)$ is defined for which the following decomposition property holds:

$$p(\mathcal{V}_1, \ldots, \mathcal{V}_N) = \prod_{n=1}^{N} \prod_{i=1}^{m} p(V_n^i \mid \pi(V_n^i))$$

Applying a dynamic Bayesian network usually amounts to computing the marginal probability distributions of the hidden variables at different times. The computations involved constitute the *inference*. Three types of inference are distinguished. *Monitoring* is the task of computing the probability distribution for $\mathcal{X}_n$ at time n given the observations that are available up to and including time n. *Smoothing* is the task of computing the marginal probability distribution for $\mathcal{X}_n$ at time n given the observations available up to time N where $N > n$. Finally, *forecasting* is the task of predicting the probability distribution of $\mathcal{X}_n$ at time n given the observations that are available about the past up to time N where $N < n$.

3.2 Modelling issues

A natural extension of the diagnostic part of the sVAP network is a network that represents time explicitly [4]. Figure 4 gives an overview of the structure of the dynamic network that we constructed for the diagnosis of VAP, which we call the dVAP network. The dVAP network includes two interacting dynamic hidden processes, modelled by the variables *colonisation* and *pneumonia*; there is no transitional influence between them, but both are persistent. The process of colonisation is influenced by three input variables, *hospitalisation*, *mechanical ventilation* and *previous antibiotics*, which in essence control its dynamics. We note

Figure 4: The dVAP network for the diagnosis of VAP; clear nodes are hidden, shaded nodes are observable. The dashed boxes indicate the hidden processes of the network.

that the variables *hospitalisation* and *mechanical ventilation* are observed for a period that is longer than the transition interval of the model. The variables thus are modelled as affecting adjacent time steps. The variable *previous antibiotics* represents the effect of previous medication to the patient on the process of colonisation.

One of the difficulties in constructing the dVAP model, was defining the length of the transition interval. It may seem trivial in general to decide upon the actual interval length, but in our case it proved to be rather difficult since there was no *a-priori* commonly acknowledged interval length that appropriately represents the evolution of the unobserved disease. Also, there was not a standard interval with which observations were collected in our data files. The latter can be attributed to most of the measurements being collected by nurses; for example, observable variables such as *body temperature* and *sputum colour* were measured frequently (approximately every two or three hours), while variables such as *radiological signs* and *leucocytosis* were measured once per day. Based on these insights, we decided to use a transition interval of one day (24 hours) for the dVAP network. Within this interval, the network *aggregates* the observations in a way similar to the previously constructed static network. For each observable variable, the value most frequently observed during the day was chosen as representative for that day; in cases where there was no prevalent value in the data, the worst value observed for the patient was chosen, to allow for *conservative* conclusions from the network. The chosen transition interval appealed to be compatible with the application characteristics and admissible by the domain experts.

A main issue in building the dVAP network was the acquisition of all conditional probabilities required. Although the three ICUs that acted as a setting for this study used the same shared computer-based patient record system, it appeared very hard to select relevant patient cases from the collected data. The main reason was that VAP is always a concomitant disease. As a consequence, clinicians tend to not report the presence of VAP in a patient. We thus found that only in a very small proportion of cases, a patient was reported as having VAP. Since we could not exploit the data for estimating the probabilities for our network, all parameters had to be assessed by experts. Compared to the sVAP network, the new parameters to be assessed concerned the dynamics of the stochastic processes of colonisation and

Suppose a patient has been mechanically ventilated for 48 hours and now has pneumonia caused by *s.aureus*. If this patient after 24 hours is *not mechanically ventilated*, but is *colonized with s.aureus* and has *phagocyte dysfunction*, then how likely is it that the patient will still have pneumonia caused by s.aureus ?

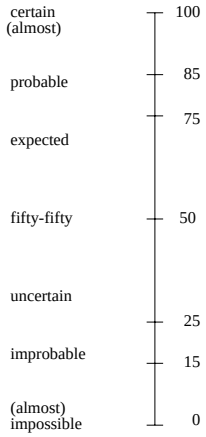


Figure 5: The fragment of text and probability scale for the assessment of the conditional probability $p($*pneum.aureus=yes* $|$ *pneum.aureus* $=$ *yes*, *mech.ventilation* $=$ *no*,*colonisation.aureus* $=$ *yes, phagocytes.dysfunction* $=$ *yes*$)$

pneumonia. To estimate those probabilities from domain experts we used the elicitation method proposed by Van der Gaag et al. [9]. This method is tailored to eliciting a large number of probabilities in a short time. Its main characteristic is the idea of presenting conditional probabilities as fragments of text and of providing a scale for marking assessments with both numerical and verbal anchors; for every conditional probability that needs to be assessed the domain experts are provided with a separate figure with the text and associated scale. Figure 5 shows, as an example, the figure pertaining to the conditional probability

$$p(\textit{pneum.aureus=yes} \,|\, \textit{pneum.aureus} = \textit{yes}, \textit{mech.ventilation} = \textit{no}, \\ \textit{colonisation.aureus} = \textit{yes}, \textit{phagocytes.dysfunction} = \textit{yes})$$

for the dVAP network. In total, 2226 probabilities were elicited from a single domain expert within a few hours.

3.3 Computational issues

The practicability of the dVAP network depends to a large extent on the computational burden of inference with the network. For diagnosing patients with VAP, we monitor them at each time step. For this purpose, we use the *interface algorithm* with the dVAP network [5]. The interface algorithm is an extension of the *junction-tree algorithm* for inference with Bayesian networks in general [3], efficiently exploiting the forward interface of a dynamic network. The algorithm is linear in the total number of time steps and for large time scopes, the computation time can prove to be prohibitive for practical purposes.

Recent results show that, in case consecutive similar observations are obtained, the probability distribution of the hidden process converges to a limit distribution within a given level of accuracy [2]. After some number of time steps, therefore, there is no need for further inference as long as similar observations are obtained. The phenomenon of consecutive similar observations was evident for several patients in the ICU files. For example, for many patients we found that the same configuration of values was observed for all or almost all of the observable variables for a number of consecutive days.

Using the *relative entropy* distance measure for distributions, we can further show that it suffices to use just the most recent data for monitoring. This result depends on the properties of the transition matrix that models the evolution of the process, but a detailed description is out of the scope of the present paper. We define the *forward acceptable window* $\omega^f_{n,\epsilon}$for the present time step n given a specified level of accuracy ϵ, to be the minimal number of time steps that we need to use from the past to compute the probability distribution of the hidden variable at the present time within the level of accuracy ϵ. The scheme below illustrates the concept of the forward acceptable window, whose value can be established based upon the properties reviewed above:

$$\underbrace{\{1, \ldots, n_f, \ldots, n\}}_{\text{total time scope}} \longrightarrow \underbrace{\{n_f, \ldots, n\}}_{\omega^f_{n,\epsilon}}$$

We can now perform inference for time step n by considering only the forward acceptable window $\omega^f_{n,\epsilon}$ without losing too much in accuracy. Note that by doing so, the runtime requirements decrease from $O(n)$ to $O(n - n_f)$.

The main conclusion from the above considerations is that monitoring in the dVAP network can be eased considerably by exploiting the nature of the observations for a patient and by using the forward acceptable window.

4 Diagnostic performance

We evaluated the performance of the dVAP network, focusing on its diagnostic prediction per day. At our disposal we had a temporal database with data from 2233 distinct patients. Each record contains data collected for a patient during a one day stay in the ICU. The source of these data is the clinical management system used at the Intensive Care Units of the University Medical Centre Utrecht in the Netherlands. For 157 of these patients, VAP was established by two infectious-disease specialists. The conclusions obtained by the dVAP network were examined on a group of 20 patients in total, 5 of which were diagnosed with VAP. For these 5 patients we used the data from the day of admission to the ICU until the day they were diagnosed with VAP which was 10 days per patient. For each of these 5 patients, three patients for whom it was known

symptoms	VAP n = 5	no VAP n = 15
abnormal temperature	60%	7%
mech. ventilation (mean)	10d	10d
abnormal leucocytes	80%	53%
abnormal pO_2/FiO_2	60%	27%
abnormal sputum	80%	73%
coloured sputum	60%	60%
colonised	40%	13%
antipyretic drugs	100%	87%
positive X chest	40%	0%

Table 1: Data summary

that they did not develop VAP over time, were matched on three criteria: gender, number of mechanically ventilated days, and ICU ward. Table 1 summarises the data for the 5 patients with VAP and for the 15 patients without VAP at the tenth day of admission.

To compare the diagnostic performance of the dVAP network to that of the original sVAP network, we used the Brier score [6], [8]. We illustrate the Brier score for our dVAP network. For each patient i, the network yields a probability distribution p_i over the two values $j = 1, 2$ (yes, no) of pneumonia. The Brier score B_i for this distribution is defined as

$$B_i = \sum_{j=1,2} (p_{ij} - s_{ij})^2$$

where $s_{ij} = 1$ if the medical record of the patient states the value j, and $s_{ij} = 0$ otherwise. If the network would yield the correct value with certainty for a patient, then the associated Brier score would be equal to 0. For the probability distribution computed for any patient, therefore, the Brier score ranges between 0 and 2, and the better the prediction is, the lower the score. The Brier scores for all patients as well as the probability of VAP at the day it was diagnosed, for the dVAP and the sVAP networks respectively, are shown in Table 3. We note that for 15 patients of the total of 20 the computed Brier score was lower with the dVAP model than with the sVAP network.

The quality of the two networks can be expressed in an overall score that is computed from the scores for our collection of patients. For m patients, the overall Brier score is defined as

$$B = \frac{1}{m} \sum_{i=1,\ldots,m} B_i$$

The overall Brier score for the sVAP network can be readily computed from Table 3 and equals 0.3370, while the overall Brier score for the dVAP network is 0.2376. The lower score for the dVAP network conveys the information that this network is better informed than the sVAP network and can arrive at relatively good estimates for diagnosing VAP.

Compared to the sVAP network, the dVAP network takes into consideration the history of a patient. For the patients $22122, 23844, 24114, 21542, 22736$ for example, who were not diagnosed with VAP, the dVAP network derived low probabilities for the presence of VAP by exploiting all previous information. The sVAP network, in contrast, used just the current information and produced much higher probabilities. For the patients diagnosed with VAP, the two models behave more or less similarly, with the highest absolute discrepancy observed in patient 28393, to whom the sVAP network assigned a probability of VAP of 0.997 and the dVAP network assigned a probability of VAP of 0.904.

patient id.	24528	22303	23505	23844
exact	0.9987	0.0015	0.0005	0.0325
$\omega^f_{10,0.003}$	0.9987	0.0013	0.0005	0.0347

Table 2: Exact and approximate probabilities for VAP for a group of matched patients.

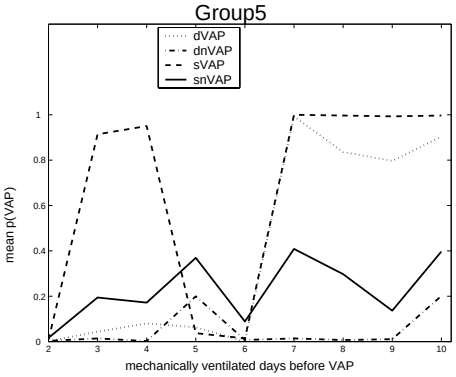


Figure 6: The dVAP and sVAP performance over time for a group of matched patients; dnVAP and snVAP represent the performance for the three patients without VAP combined.

To study the performance of the dVAP network over time, we computed the probability of VAP for each day and compared it to the respective probability from the sVAP network. In Figure 6 we plot, for a single group of four related patients, the probability of VAP for patient 28393 and the mean probability of VAP for the matched patients $21542, 22301, 22736$, from both networks. We observe that for the patient with VAP the trend in both networks is more or less the same after the fifth time step; to the patients without VAP, however, the dVAP network assigns lower probabilities than the sVAP network. The dVAP network thus is better able to distinguish between VAP and non-VAP patients.

To conclude, we performed the computations in the dVAP network using different values for the forward acceptable window $\omega^f_{n,\epsilon}$. We conclude that instead of using the observations for all 10 days in the ICU to compute the probability of VAP, we can use the observations for just the last 5 days with an average error for all patients smaller than $\epsilon = 0.003$. For a particular group of matched patients for example, the exact and approximate probabilities for VAP are showed in Table 2. We can thus use this forward acceptable window to speed up the computations and obtain results with an almost negligible error.

5 Discussion

In this paper, we discussed the construction of a probabilistic model that is aimed at assisting ICU clinicians in diag-

patient id.	VAP	sVAP	sBrier	dVAP	dBrier
22022	yes	0.996913	$1.90591 \cdot 10^{-5}$	0.9987	$3.38 \cdot 10^{-6}$
22563	no	0.0203017	$8.24318 \cdot 10^{-4}$	0.1395	0.0389205
22716	no	0.167208	0.055917031	0.0558	0.00622728
22730	no	0.00276365	$1.52755 \cdot 10^{-5}$	0.0002	$8 \cdot 10^{-8}$
23397	yes	0.00972048	1.961307055	0.0002	1.99920008
22122	no	0.430888	0.371328937	0.0316	0.00199712
22634	no	0.0203017	0.000824318	0.0003	$1.8 \cdot 10^{-7}$
22659	no	0.193411	0.07481563	0.0309	0.00190962
24528	yes	0.999959	$3.362 \cdot 10^{-9}$	0.9987	$3.38 \cdot 10^{-6}$
22303	no	0.0226662	0.001027513	0.0015	$4.5 \cdot 10^{-6}$
23505	no	0.0457446	0.004185137	0.0005	$5 \cdot 10^{-7}$
23844	no	0.297688	0.177236291	0.0325	0.0021125
25724	yes	0.0347989	1.863226327	0.0033	1.98682178
23872	no	0.0203017	0.000824318	0.0005	$5 \cdot 10^{-7}$
24114	no	0.43644	0.380959747	0.099	0.019602
24151	no	0.00999126	$2.22311 \cdot 10^{-5}$	$7 \cdot 10^{-8}$	$9.8 \cdot 10^{-15}$
28393	yes	0.996666	$2.22311 \cdot 10^{-5}$	0.9035	0.0186245
21542	no	0.175202	0.061391482	0.0218	0.00095048
22301	no	0.0740135	0.010955996	0.0013	$3.38 \cdot 10^{-6}$
22736	no	0.942073	1.775003075	0.581	0.675122

Table 3: Brier scores for the sVAP network and for the dVAP network, respectively.

nosing ventilator-associated pneumonia. In contrast to previous approaches that used a static decision-theoretic network for this low-prevalence disease, we focused on its dynamic evolution and used a dynamic Bayesian network as the primary tool for representation and inference.

We detailed various modelling steps in the construction of our dynamic network and described the use of an efficient procedure for expert elicitation of the probabilities required. We further argued that a number of convergence properties of dynamic Bayesian networks can be exploited to arrive at feasible algorithms that restrict the computational burden of inference with such a model. In this way, we ameliorated two important problems that were considered impervious in the past: the specification of the probabilities underlying the stochastic process modelled in the network and the computational burden of inference.

We evaluated our network on a set of ICU patients to examine its diagnostic accuracy. The lower overall Brier score of the dynamic network in comparison to the static one, indicated that representing time explicitly and taking into consideration the history of the patient, increases diagnostic performance. In our evaluation experiments, the dynamic network proved to be better at distinguishing between VAP and non-VAP patients than the static network, especially by assigning lower probabilities of VAP to the non-VAP patients. In the near future, we intend to improve the dVAP network by use of the available data for parameter learning and to test it on more ICU patients with the aim of embedding it in the clinical information system of the ICU.

Acknowledgements

This research was (partly) supported by the Netherlands Organization for Scientific Research (NWO). The authors would like to thank Marc Bonten for his valuable comments on an earlier version of this paper.

References

[1] M.J. Bonten. Prevention of infection in the intensive care unit. *Current Opinion in Critical Care*, 10(5):364-368, 2004.

[2] T. Charitos, P. de Waal, and L.C. van der Gaag. Speeding up inference in Markovian models. *Proceedings of the 18th International FLAIRS conference*, pp. 785-790, 2005.

[3] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, 1999.

[4] P.J.F. Lucas, N.C de Bruijn, C.A.M Schurink, and A. Hoepelman. A probabilistic and decision theoretic approach to the management of infectious disease at the ICU. *Artificial Intelligence in Medicine*, 19(3):251-279, 2000.

[5] K. Murphy. *Dynamic Bayesian networks: Representation, Inference and Learning*. PhD thesis, University of California Berkley, 2002.

[6] H.A. Panofsky and G.W. Brier. *Some applications of Statistics to Meteorology*. The Pennsylvania State University, University Park, Pennsylvania, 1968.

[7] C.A.M. Schurink. *Ventilator Associated Pneumonia: a Diagnostic Challenge*. Ph.D thesis, Utrecht University, 2003.

[8] L.C. van der Gaag and S. Renooij. Probabilistic networks as probabilistic forecasters. *Proceedings of the Ninth Conference on Artificial Intelligence in Medicine in Europe* pp. 294-298, 2003.

[9] L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B. Aleman, and B.G. Taal. How to elicit many probabilities. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pp. 647-654, 1999.

A Probabilistic Method for Multiple-Patient Temporal Abstraction

Michael Ramati and Yuval Shahar
Medical Informatics Research Center
Department of Information Engineering
Ben-Gurion University
P.O.B. 653, 84105 Beer-Sheva, Israel
{ramatim, yshahar}@bgu.ac.il

Abstract

Several systems have been designed to reason about longitudinal patient data in terms of abstract, clinically meaningful concepts derived from raw time-stamped clinical data. All approaches had to some degree severe limitations in their treatment of incompleteness and uncertainty that typically underlie the raw data, on which the temporal reasoning is performed, and have generally narrowed their interest to a single subject. We have designed a new probability-oriented methodology to overcome these conceptual and computational limitations. The new method includes also a practical parallel computational model that is geared specifically for implementing our probabilistic approach in the case of abstraction of a large number of electronic medical records.

1 Introduction

The commonly occurring task of Temporal Abstraction (TA) was originally defined as the problem of converting a series of time-oriented raw data (e.g., a time-stamped series of chemotherapy-administration events and various hematological laboratory tests) into interval-based higher-level concepts (e.g., a pattern of bone-marrow toxicity grades specific to a particular chemotherapy-related context) [Shahar, 1997]. Several of the main objectives involved in solving this task include the need for a formal representation that facilitates acquisition, maintenance, sharing, and reuse of the required temporal abstraction knowledge. Most of these aspects were catered for by the Knowledge-Based Temporal Abstraction (KBTA) method [Shahar, 1997] and its extensions [O'Connor et al., 2001; Spokoiny and Shahar 2001; Balaban et al., 2004]. Nevertheless, these solutions, although being evaluated as fruitful, maintained several unsolved subproblems. These subproblems seem common to some of other methods suggested for solving the TA task as well as closely related systems applied in the clinical domain [De Zegher-Geets, 1987; Kohane, 1987; Russ, 1989; Kahn, 1991; Haimowitz and Kohane, 1993; Miksch et al., 1997; Salatian and Hunter, 1999]. Thus, Considering these challenging subproblems suggests an additional method.

At least three subproblems in the former methods can be pointed out, which we propose to solve through the method discussed in this paper. First, raw clinical data, to which the temporal reasoning is being applied, are assumed as certain – that is, typically no mechanism is suggested for handling the inherent impreciseness of the laboratory tests taken to obtain the clinical data. Second, current mechanisms used for completing missing data in an electronic medical record are typically not sound and are incomplete. For example, in the case of the KBTA method, a knowledge-based interpolation mechanism is used [Shahar, 1999]. However, completion of missing values is supported only for bridging gaps between two intervals, in which the proposition (e.g., anemia level) had the same value (e.g., moderate anemia). Furthermore, the value concluded by inference is too crisp, and a threshold is used for computing it with absolute certainty, eliminating uncertainty and leading to potentially unsound conclusions. Third, no special mechanism has been devised for multiple patient abstraction. That is, so far temporal abstraction was performed on a single patient only.

The proposed method, *Probabilistic Temporal Abstraction* (PTA), decomposes the temporal abstraction task into four subtasks, that solve the case of a single patient, and two more subtasks that solve the case of multiple patients. In addition to overcoming the above mentioned subproblems, we also propose a design for a parallel computational model that implements the method.

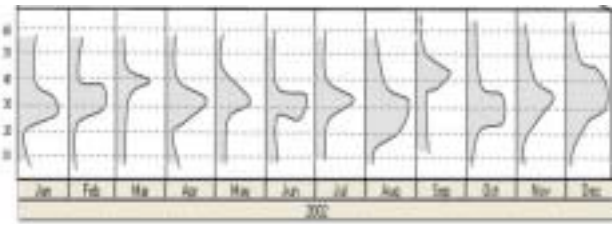


Fig. 1. A typical instance of using the PTA method: the value (*vertical axis*) distribution of a certain medical concept appears for different (in this case consecutive) periods along the time axis. The medical concept, which can be either raw or abstract, and the specification of the set of periods (including the time granularity) are determined by the application using the PTA method.

2 The Subtasks of the PTA Method

Several basic notions in probability theory relate to time, and are important when considering a probabilistic temporal model, task or mechanism. A *stochastic process* $\{X_t : t \in T\}$ is a set of random variables, and may represent a clinical observation, a medical intervention, or an interpretation context of some clinical protocol. The index is often interpreted as time, and thus X_t is referred as the *state* of the process at time t. The set T is called the *index set* of the process. The clinical subtasks specified below are defined in terms of these notions.

2.1 Single-Patient Subtasks

Temporal abstraction for a single patient requires one basic subtask, interpolation, and three interpolation-dependent subtasks – coarsening, transformation and pattern matching.

Temporal Interpolation. Estimating the distribution of a stochastic process state, given the distributions of some of its other states (Fig. 2). For example, estimating the distribution of raw hematological data or derived concepts (such as bone-marrow toxicity grades) during a week in which raw data were not measured, using the distribution of values before and after that week. Applying the interpolation subtask does not increase the abstraction level of the underlying stochastic process, but rather serves the role of a core operation that enables the application of actual temporal abstraction.

Temporal Coarsening. Applying an aggregation function to a stochastic subprocess (Fig. 3). The coarsening subtask abstracts over the time axis and is aimed at the calculation of a stochastic process at a coarser time granularity.

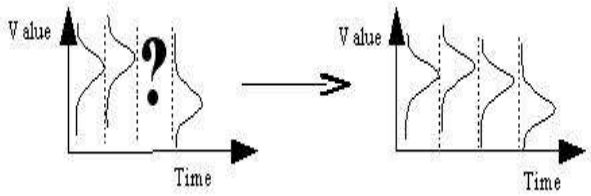


Fig. 2. An illustration of the interpolation subtask. Given the value distribution at several time points, there is a need to calculate an unobserved value distribution. The solution suggested by the PTA considers all value distributions.

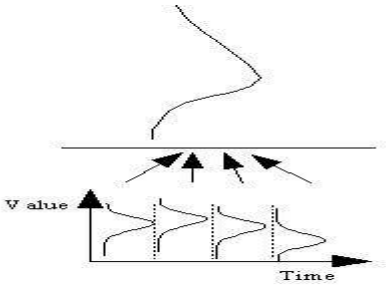


Fig. 3. An illustration of the temporal coarsening subtask. Given the value distribution at several time points, there is a need to calculate an aggregated distribution.

Temporal Transformation. Generating a stochastic process, given stochastic processes of a lower abstraction level. For example, deriving bone-marrow toxicity grade distribution, given the distributions of the raw white blood cell and platelet counts. The transformation subtask abstracts along the (clinical concept) abstraction-level axis.

Temporal Pattern Matching. Locating occurrences of specific values in certain time constraints of high-level predefined temporal variables. As opposed to the temporal transformation subtask, that maps all given data to a higher-level temporal concepts, this subtask is aimed at finding those data sets which complies with the given pattern.

2.2 Multiple-Patient Subtasks

Applying the TA task to multiple patients requires extra subtasks, such as the ones explicated below. However, these subtasks fit also sophisticated needs of abstraction for a single patient.

Temporal Aggregation. Generating an aggregation of stochastic processes. The Aggregation subtask abstracts along the patient axis. This subtask is aimed at the application of aggregation functions, such as minimum, maximum, average, etc. on stochastic processes.

Temporal Correlation. Calculating the correlation between two stochastic processes. The correlation subtask compares two temporal abstractions. This subtask is intended to mainly compare two patient populations, but should work the same when comparing different time periods of the same patient.

3 The PTA Property

The central property of the PTA method is based on the notion of *temporal field*, as defined below. Following this definition, the property states, that each unobserved state of some stochastic process is a linear combination of the temporal fields of the observed states of the process. Thus, the unobserved distribution of bone-marrow toxicity grades is a linear combination of all of the observed distributions, before and after it. A proper basis that will fit the requirements of the PTA property could be found in the following two known definitions.

Let $\{w_{ij}\}_{1 \le i \le m,\, 1 \le j \le n}$ and $\{\mu_i\}_{1 \le i \le m}$ be constants. The random variables X_i are said to have *multivariate normal distribution*, if:

$$X_i = \sum_j w_{ij} \cdot Z_j + \mu_i \quad , Z_j \sim Normal(0,1)$$

A stochastic process $\{X_i : i \ge 0\}$ is called *Gaussian process* if each state X_i of the process has a multivariate normal distribution.

Uncertain Observations. Observed states of stochastic processes are distributed as a function of the clinical test taken and the clinical data itself. Typically, where states of stochastic processes have a normal distribution, the mean (expectation) of the state is the value sampled, and the variance is determined by the reliability or the precision of the test taken.

Temporal Fields. Calculating a dependent variable given the independent variables as they appear in a multivariate distribution may imply a temporal *persistence* of the independent variables. However, allowing the observed states to induce a *field*[1] over its temporal environment could express temporal knowledge about the stochastic process in question, such as a *periodic behavior* or *change* of the observed states. Thus, for each stochastic process, a temporal field is induced by a time index, which formally means a function that maps time points to states of the stochastic process, as follows:

$$field_{\vec{X}}(t): T \to \mathbb{R}^{\Omega} \quad , X_t: \Omega \to \mathbb{R}$$

For example, suppose a stochastic process with a periodic behavior and cycle length *c*. The temporal field of an observed state of such stochastic process could be as follows:

$$(field_{\vec{X}}(t_s))(t_i) = \sin\left(\frac{\pi}{c} \cdot (|t_i - t_s| \bmod c)\right) \cdot X_{t_s}$$

Temporal Weighting. A specific choice for the selection of the weights of the independent variables can be suggested. These weights should express the notion that the closer-in-time the observed state is – the more relevant it is. That is, the absolute time difference between a dependent state and its observed state should be inversely proportional to the weight of the latter when estimating the former. Therefore, there is a need to choose a monotonic decreasing function of absolute time differences between a dependent state and its inducing observed states. The weighting function is of the following form:

$$w_{\vec{X}}: \Delta T \to \mathbb{R}$$

A natural choice for the monotonic decreasing weighting function would be a normal density, where its variance (σ^2) determines the temporal tolerance of observed states of the stochastic process. Thus, *w* may hold:

$$w_{\vec{X}}(\Delta t) = f_W(\Delta t) \quad , W \sim Normal(0, \sigma^2)$$

Prior Knowledge. Each stochastic process may have a prior knowledge of its typical state. Prior distribution is expressed by giving it the $-\infty$ time index for the temporal field inducer argument, as well as the temporal field argument.

4 Mechanisms of the PTA Method

The main computational concept in our methodology is the PTA chain. A *PTA chain* is defined as the application of any subset of the following composition of subtasks, while preserving the relative order among them:

$$Coarsen \circ Correlate \circ Aggregate \circ Transform \circ Interpolate(data)$$

Temporal Interpolation. The subtask of interpolation is solved by the application of the PTA property. Given the temporal weighting function of a stochastic process, its values need to be normalized to ensure they sum to unity. The subset of sampled states which participate in the calculation process of each unobserved state determines the precision of its distribution, and could be determined given the temporal weighting function. If we interpolate in t_i and have all of the points that are known t_s sampled, then:

$$X_{t_i} = \frac{1}{\sum_{t_s} w_{\vec{X}}(t_i - t_s)} \sum_{t_s} w_{\vec{X}}(t_i - t_s) \cdot (field_{\vec{X}}(t_s))(t_i)$$

For each temporal gap between sampled data, the procedure *Interpolate* generally computes the value distribution of missing states starting at one extreme point (an observed state) until either reaching the prior value distribution (and then doing the same in the other direction) or the other extreme. This leaves out states in which prior value distribution is expected, in order to reduce costs in time and space. For the case in which updates to the underlying clinical data occur, we consider a hierarchical system of states, where each unobserved state has a set of observed parent states, as depicted by Pearl [1987]. In case the sample is updated, propagating the new piece of evidence we are viewing as the perturbation that propagated through a Bayesian network via message-passing between neighboring processors.

The knowledge required for the application of the interpolation subtask includes for each type of PTA chain the definitions of temporal fields (the default is set to persistence of the inducer state), temporal weighting (the default is set to normal density function with mean 0), prior distribution of a typical state (no default is set), and a function that maps each pair of clinical test taken and datum (sampled value) to the distribution of the field inducing state (default sets sampled value to the state's mean).

Temporal Coarsening. The procedure *Coarsen* transforms a given PTA chain to one with a coarser time granularity. The value of such application to a subchain in the requested time-granularity length is a stochastic state, according to the following formula:

$$X_{[t_i, t_j]} = \frac{1}{j - i + 1} \cdot \sum_{k=i}^{j} X_{t_k}$$

Temporal Transformation. The procedure *Transform* returns the application of the given transformation function to the given PTA chains according to the following formula:

$$Y_t = (g(\overrightarrow{X_1}, \ldots, \overrightarrow{X_n}))(t)$$

If *g* has the following form, then $|g|$ is called a *rate* transformation, and *sgn(g)* (positive, negative or zero) is called a *gradient* transformation:

$$(g(\vec{X}))(t_i) = \frac{X_{t_i} - X_{t_{i-1}}}{t_i - t_{i-1}}$$

For example, in the case of a contemporaneous transformation of several arguments (e.g., height and weight) into a higher-level abstraction (e.g., body-mass index), the time-series of the arguments are the same as the of the abstraction. However, a context of a Bone-Marrow Transplantation (BMT) is defined as the application of the fol-

[1] In the sense of an electromagnetic field.

lowing transformation function to the Boolean day-granularity stochastic process that represents a BMT:

$$(g(\overrightarrow{BMT}))(t) = BMT_{t-3} \vee \ldots \vee BMT_{t+90}$$

Temporal Pattern Matching. The procedure *Match* returns the probability of the occurrence of the given temporal pattern in each subinterval of the given time interval. The temporal patterns are represented by regular expressions, where the concatenation operator stands for temporal succession, Kleene-closure stands for a temporally unbounded repetition and the alphabet Σ_g is the discrete finite vector space spanned over the sample or transformation spaces of random variables composing the temporal pattern, and *g* is the respective time granularity. That is, a letter $\sigma \in \Sigma_g$ is a vector, which its *i*-th coordinate is some possible discrete value of the *i*-th variable composing the pattern. For example, a pattern of platelet half-life is composed of bone-marrow transplantation (first coordinate) and platelet state (second coordinate), using ϕ to represent all value possibilities, and an hour-granularity:

$$\langle true, \phi \rangle \langle \phi, high \rangle$$
$$(\langle \phi, high \rangle \cup \langle \phi, normal \rangle) * \langle \phi, low \rangle$$

The probabilistic nature of the underlying data requires the temporal matching mechanism to compute the conditional probability of the occurrence of each letter given the occurrence of the subpattern preceding it. In order to identify the data used for the probability computation of the preceding subpattern, one needs to find the time-series of the transformation arguments of each coordinate in the preceding letters. This is accomplished by the definition and application of functions of the following form:

$$h_{\vec{Y}}(t) = \langle \overrightarrow{T}_1, \ldots, \overrightarrow{T}_n \rangle$$

Given these functions, the interpolation mechanism is used only for the resulting time intervals as well as given the already computed subpattern-match probability. The probability the occurrence of some pattern in a given time interval is thus the joint probability of its letters, i.e., the multiplication of their conditional probabilities. Computing the value distribution of some letter coordinate given its conditional distribution (when matching a new interval, that is not conditioned with the time-points given in the former interval matched) is done by removing the weighted temporal fields from the (interpolated) conditional distribution. The matching process continues until the probability for the occurrence of some letter's coordinate equals or lesser than its prior probability, or until the pattern was fully matched.

Temporal Aggregation. Applied to stochastic processes of the same sample space and independent patients, resulting in a new stochastic process. This measure is computed as a new PTA chain, where each of its state is the application of some aggregative function (minimum, maximum, average, etc.) to the corresponding states of the given PTA chains. Suppose *agg* is some aggregation function and t_i is some time-point, then:

$$agg_{t_i}(\overrightarrow{X}_1, \ldots, \overrightarrow{X}_n) = agg(X_{1t_i}, \ldots, X_{nt_i})$$

Temporal Correlation. Applied to stochastic processes of different sample spaces, independent patients or same, resulting in a series of correlation factors. This measure is computed as a time series of correlations between corresponding states of the given PTA chains:

$$\rho(X_{t_i}, Y_{t_j}) = \frac{Cov(X_{t_i}, Y_{t_j})}{\sqrt{Var(X_{t_i}) \cdot Var(Y_{t_j})}}$$

An example for a single patient would be the contemporaneous correlations between height and weight or correlation of height during different periods for the same person.

6 The Parallel Computational Model

The computational model used to compute a PTA chain is *goal-driven, bottom-up* and *knowledge-based* (the pattern matching mechanism is *top-down*, however, as explicated above). The main algorithm is thus required to compute the result of a PTA chain (the goal), given the transformation and interpolation functions (the temporal knowledge) as well as the access to the clinical data, beginning at the raw (lowest abstraction level) clinical data. The computational model is parallelized in three orthogonal aspects: (1) Time, during the calculation of the PTA chains' states; (2) Transformation, during the calculation of the transformation arguments; and (3) Patient, during the calculation of the PTA chains for multiple patients.

The Main Algorithm. A parallel algorithm is typically presented in terms of a theoretical model for parallel computing: the *Parallel Random-Access Machine* (PRAM) [Brent, 1974]. In its basic architecture, the PRAM model includes *p* serial processors that have a shared memory. We shall assume the PRAM supports *concurrent-read*, i.e., multiple processors can read from the same location of shared memory at the same time.

The following procedure computes the PTA chain for the given patient, goal and index set. First, it retrieves the goal's transformation function. In case it does not exist, it retrieves the raw clinical data, interpolates the missing clinical data, and may change in parallel the time granularity. If the transformation function was found, its arguments are retrieved, and the transformation is applied in parallel.

Complexity of the Computation. The results of asymptotic run-time analysis for parallel combinatorial circuits (Brent's theorem) [Brent, 1974] can be applied to such analysis of the overall algorithm. The description of the different PTA mechanisms suggests parallelizing the interpolation subtask and the temporal coarsening subtask on the time axis, and the transformation subtask on its arguments axis. Multiple-patients subtasks are parallelized on the time axis as well as on the patient axis. Let *args* be the the maximal number of arguments in all transformation, let ΔT be the temporal length of the requested PTA chain, let *p* be the number of processors, and let *level* be the number of transformations applied until the requested goal is reached, then the corresponding PTA chain are created in:

$$O(args^{level+1} \cdot \Delta T \cdot |subjects| / p + level)$$

7 Implementation

The PTA architecture is in the process of fully being implemented using the C++ programming language, the Standard Template Library (STL), and the MPICH2 implementation of the Message-Passing Interface (MPI)[2], an international parallel programming standard. The implementation is thus object-oriented and platform-independent. The implementation is in the process of fully integrated into the IDAN system [Boaz and Shahar, 2005], which satisfies the need to access medical knowledge and clinical data sources.

8 Discussion

In this paper, we proposed a probabilistic method to solve the task of abstraction of longitudinal clinical records, and described a *scalable* [Hwang and Xu, 1998] parallel computational model that implements it. The new method has removed several limitations of former methods. First, the use of PTA chains enables the expression of uncertainty in the underlying clinical data. Second, two mechanisms were developed for temporal abstraction of the clinical data of multiple patients. Third, the interpolation mechanism was shown to be sound and complete. However, the previous model's assumptions were replaced with those of the other's: observed clinical data are assumed to be independently distributed. This assumption could be easily removed, provided the extra medical knowledge of conditional distribution functions for the underlying stochastic processes available.

When dealing with probability of events that occur over time, it is not unusual to assume the Markovian property. This property states that the conditional distribution of any future state, given the present state and all past states, depends only on the present state and is independent of the past. Our probabilistic temporal model, however, cannot assume this known property for a couple of reasons. First, the property does not hold for temporal chains, in which past states help in forecasting future states. Second, the assumption that is actually needed is one that would explicitly state the influence of future states on *interpreting* past states, and in particular on interpolating the present state, given past and future states.

Finally, there are two more points that are worth mentioning, while comparing the proposed method to the model used to solve the temporal abstraction task, as part of the KBTA method. First, as it was specified in section 4, the interpolation in the PTA model is performed at the lowest abstraction level only, as opposed to being repeatedly performed at every abstraction level in the former method. Second, the temporal patterns can be acquired in any temporal representation language, such as CAPSUL [Chakravarty and Shahar, 2001] or TAR [Balaban et al., 2004], assuming it is reducible to regular expressions in the temporal semantics attributed above. The expressions of the source language are then compiled to the formal regular expressions beforehand, thus gaining modularity as well as run-time computational speedup.

We are in the process of fully implementing the new architecture and evaluating it on a large longitudinal clinical database.

[2]http://www.mpi-forum.org/

Acknowledgments

This work has been supported in part by NIH award No. LM-06806 and Israeli Ministry of Defense award No. 89357628-01. We would like to thank Denis Klimov and Efrat German of the Medical Informatics Research Center in Ben-Gurion University.

References

[Balaban et al., 2004] Applying Temporal Abstraction in Medical Information Systems. *Annals of Mathematics, Computing & Teleinformatics* 1(1), 54-62.

[Boaz and Shahar] A Framework for Distributed Mediation of Temporal-Abstraction Queries to Clinical Databases. *Artificial Intelligence in Medicine* (in press)

[Brent, 1974] The Parallel Evaluaion of General Arithmetic Expressions. *Journal of the ACM* 21.2: 201-206

[Chakravarty and Shahar, 2001] Specification and Detection of Periodicity in Clinical Data. *Methods of Information in Medicine* 40(5), 410-420.

[Chakravarty and Shahar 2000] A Constraint-Based Specification of Periodic Patterns in Time-Oriented Data. *Annals of Mathematics and Artificial Intelligence* 30 (1-4)

[De Zegher-Geets, 1987] IDEFIX: Intelligent Summarization of a Time-Oriented Medical Database. M.S. Dissertation. *Program in Medical Information Sciences*, Stanford University School of Medicine

[Haimowitz and Kohane, 1993] Automated trend detection with alternate temporal hypotheses. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, San Mateo 146–151.

[Hwang and Xu, 1998] Scalable Parallel Computing, WCB McGraw-Hill.

[Kahn, 1991] Combining physiologic models and symbolic methods to interpret time varying patient data. *Methods of Information in Medicine* **30**(3) 167–178.

[Kohane, 1987] Temporal reasoning in medical expert systems. Technical Report 389, Laboratory of Computer Science, Massachusetts Institute of technology, Cambridge, MA.

[Ladkin, 1986a] Time representation: A Taxonomy of interval relations. In *Proc. of the AAAI*, pages 360-366

[Ladkin, 1986b] Primitives and Units for Time Specification. In *Proc. of the AAAI* pages 354-359

[Miksch et al., 1997] Time-Oriented Analysis of High-Frequency Data in ICU Monitoring. In *Intelligent Data Analysis in Medicine and Pharmacology*, N.

Lavrac, E.T. Keravnou, and B. Zupan, Editors. Kluwer. p. 17-36.

[O'Connor et al., 2001] RASTA: A Distributed Temporal Abstraction System to facilitate Knowledge-Driven Monitoring of Clinical Databases. *MedInfo 2001*, London

[Pearl, 1987] Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers.

[Russ, 1989] Using hindsight in medical decision making. In *Proceedings of Symposium on Computer Applications in Medical Care*. IEEE Computer Society Press, New York NY 38–44.

[Salatian and Hunter, 1999] Deriving trends in historical and real-time continuously sampled medical data. *Journal of intelligent information systems*. **13**: p. 47-71.

[Shahar, 1997] A Framework for Knowledge-Based Temporal Abstraction. *Artificial Intelligence* 90 79-133

[Shahar, 1999] Knowledge-Based Temporal Interpolation. *Journal of Experimental and Theoretical Artificial Intelligence* 11:123-144

[Spokoiny and Shahar 2001] A Knowledge-based Time-oriented Active Database Approach for Intelligent Abstraction, Querying and Continuous Monitoring of Clinical Data. *MedInfo 2004*:84-8.

Short term blood glucose measurements may be severely biased

Jette Randløv and Jonas Kildegård
Novo Nordisk A/S
Brennum Park, DK-3400 Hillerød, Denmark
jttr@novonordisk.com, jkip@novonordisk.com

Abstract

The values of continuous measured blood glucose have little difference when measured at short time intervals. As time increases so does the difference in average. For discretely measured values the picture is quite different: measurements made at short time intervals display a surprising higher difference than continuous measurements. We have not seen this effect reported before.

1 Introduction

Today most people with diabetes measure their blood glucose (BG) by sampling a drop of capillary blood—typically from the finger tip—and measuring with a BG-meter. Continuous measuring devices also exist like the Minimed Continuous Glucose Monitoring (CGM) that measures the BG value every 5 minutes. Here a needle has to be inserted and replaced subcutaneously every third day.

One of the main problems in the management of diabetes is to balance the dose of insulin with the near future values of the BG concentration. Being able to predict the BG level would simplify the management. Many attempts have been made to predict the value of the BG from historical data [Arita *et al.*, 1999; Hejlesen, 1998; Lehmann and Deutsch, 1998; Liszka-Hackzell, 1999; Mougiakakou and Nikita, 2002; Tresp *et al.*, 1999]. No model has shown good prediction power for more than one data set—like for instance an error rate of less than 1mM/hour. The mentioned publications involve only strip based BG measurements. Attempts to predict BG values from continuously measured BG shows a clear connection between how far into the future the prediction reaches and prediction error [Hovorka and others, 2004; Prank *et al.*, 1998]. No attempts we know of have been made to examine whether this is also true for predictions based on strip based measurements. The general assumptions seem to be that strip measurements are equal to continuous measurements, only less frequent and that the BG measurement is a sampling of the underlying reality. The present paper shows that this assumption is not sound.

2 Strip versus CGM

The accuracy of strip measurements is slightly better than continuous measurements. The accuracy is defined as the percent of measurements that are within 20% of the reference value or in the hypoglycaemic area—the zone A in the Clarke Error Grid [Clarke *et al.*, 1987].) The accuracy of measurements with a handheld meter is somewhere between 73.9% [Clarke *et al.*, 1987] and 83.5% [Alto *et al.*, 2002] and the accuracy of Minimed CGM is 70.2% [Gross *et al.*, 2000]. From these accuracies, one might think that the strip based BG measurements would be as good as CGM, just less frequent.

However, there is a very strong correlation between why people measure their BG and the actual value they measure. Consider, for example, a Modal Day plot (see Figure 1). The plot often shows low values in the small hours of the night. Is it because the BG is always low at that time? Or if the person wakes to make a measurement, is it because BG is low? In the latter case the average of the measured values have little to do with the real average.

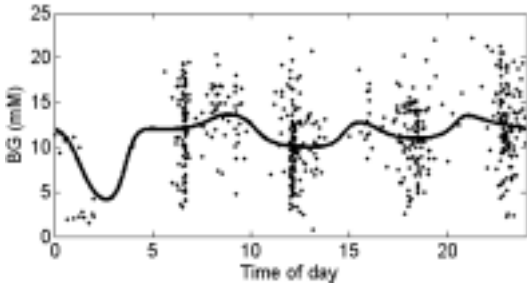


Figure 1: Modal day for data set number 20 from the AAAI 1994 Spring Symposium. The data set is one person's normal diary. Dots represent single measurements and the curve is a Gaussian smoothed average with $\sigma = 0.2$.

The Gaussian smoothed average used in the Figure is calculated by summing the value of a single BG measurement (made at time t_s) times the influence of that value at time t. The influence is calculated as $\exp\left(-\frac{(t_s-t)}{2\sigma^2}\right)\frac{1}{\sigma\sqrt{2\pi}}$.

We want to examine the degree to which BG values are related by plotting the absolute difference as a function of the time between the measured values—see Figure 2. The short term difference is high and decreases as time increases. The BG data covers 79 normal diaries including the 70 data sets from the AAAI 1994 Spring Symposium, a total of 13 615 strip measurements.

For continuously measured values the picture is quite

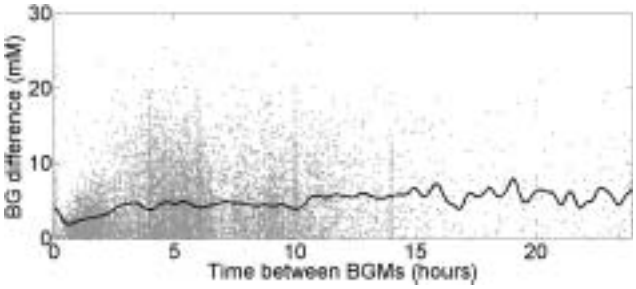


Figure 2: The dots represent the difference for single predictions and the line is a Gaussian smoothed average with $\sigma = 0.2$.

different. The deviation from one value to the next was observed at different time intervals. Summing the deviations through the complete dataset it is possible to plot the deviation as a function of the time interval—see Figure 3. Here, the deviation increases with time as would be expected.

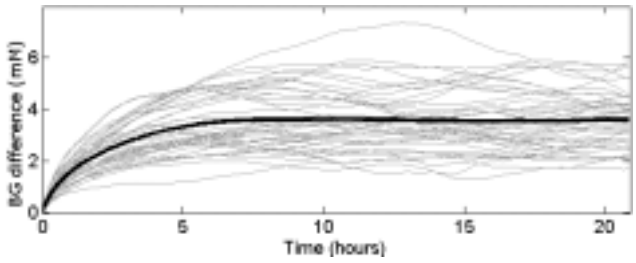


Figure 3: The grey curves show data from 42 patients in a Novo Nordisk study with 72 hours Minimed CGM measurements. The black curve is the average of the curves.

Figure 4 shows the two averages curves in the interesting area for small time intervals. The difference for small time intervals is obvious.

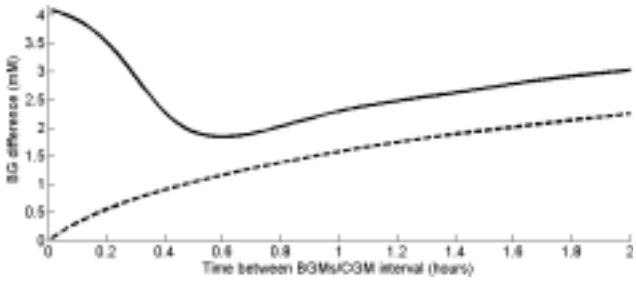


Figure 4: Strip-based (solid) and CGM-based (dashed).

Why this difference for small time intervals? Consider strip based measurements: If a person just made a measurement, why perform one more, less than half an hour later? It is likely that this occurs when there is suspicion that the first measurement was not correct, or when the person has a sensation of undergoing dramatic changes in BG. Our calculations show that these suspicions are often correct—measurements made shortly after each other are less correlated than, for instance, measurements with half an hour between them. This effect makes strip based short term prediction difficult. This is not an issue with continuously measured BG values as they are measured independently of circumstances.

3 Conclusion

We have shown that strip measurements display a strong dependence on the circumstances for measurements when made at short time intervals, making prediction of these blood glucose values relatively difficult. This is not an issue with continuously measured BG values.

References

[Alto *et al.*, 2002] William A. Alto, Daniel Meyer, James Schneid, Paul Bryson, and Jon Kindig. Assuring the accuracy of home glucose monitoring. *Journal of Am. Board Fam Pract.*, 15(1):1–6, 2002.

[Arita *et al.*, 1999] Arita, Yoneda, and Iokibe. System and method for predicting blood glucose level. US Patent 5971922, October 26, 1999.

[Clarke *et al.*, 1987] Clarke, Cox, Gonder-Frederick, Carter, and Pohl. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care*, 10:622–628, 1987.

[Gross *et al.*, 2000] Gross, Bode, Einhorn, Kayne, Reed, White, and Mastrototaro. Performance evaluation of the Minimed continuous glucose monitoring system during patient home use. *Diabetes technology & therapeutics*, 2(1), 2000.

[Hejlesen, 1998] Hejlesen. DIAS—the diabetes advisory system. Technical report, Department of Medical Informatics, The University Aalborg, 1998.

[Hovorka and others, 2004] Hovorka et al. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological Measurement*, 25:905–920, 2004.

[Lehmann and Deutsch, 1998] Lehmann and Deutsch. Compartmental models for glycaemic prediction and decision-support in clinical diabetes care: promise and reality. *Computer methods and programs in biomedicine*, 56(2):193–204, 1998.

[Liszka-Hackzell, 1999] JJ Liszka-Hackzell. Prediction of blood glucose levels in diabetic patients using a hybrid AI technique. *Computers and biomedical research, an international journal*, 32(2):132–144, 1999.

[Mougiakakou and Nikita, 2002] Mougiakakou and K.S. Nikita. Blood glucose profile prediction for type 1 diabetes patients using a hybrid approach. In *EMBEC*, pages 1160–1161, 2002.

[Prank *et al.*, 1998] Prank, Jurgens, Muhlen, and Brabant. Predictive neural networks for learning the time course of blood glucose levels from the complex interaction of counterregulatory hormones. *Neural Computation*, 10(4):941–954, 1998.

[Tresp *et al.*, 1999] Tresp, Briegel, and Moody. Neural-network models for the blood glucose metabolism of a diabetic. *IEEE-NN*, 10(5):1204, September 1999.

Profiling Examiners using Intelligent Subgroup Mining

Martin Atzmueller and **Frank Puppe**
Department of Computer Science
University of Würzburg
97074 Würzburg, Germany
{atzmueller, puppe}@informatik.uni-wuerzburg.de

Hans-Peter Buscher
DRK-Kliniken Berlin-Köpenick,
Clinic for Internal Medicine II,
12559 Berlin, Germany
buscher.dhp@t-online.de

Abstract

The demand for effective knowledge discovery methods in a clinical setting is growing: the number of hospital information systems and medical documentation systems in routine-use increases rapidly. Then, often high-quality collections of electronic patient records are available for statistical analysis. One interesting issue concerns the quality of the examinations records which depends both on the examination quality and the documentation habits of the individual examiners. We apply a subgroup mining approach for explorative and descriptive data mining to tackle this issue, and we provide a case study of the proposed approach using data from a fielded system in the medical domain.

Purely automatic data mining methods often suffer from the limitation that too many uninteresting results are presented to the user. In order to improve upon this situation, we propose two strategies: we use background knowledge, if available, and provide suitable visualizations for guiding the discovery process. The context of the presented approach is a knowledge-based documentation and consultation system.

1 Introduction

The available data in clinical settings is growing with a rapid pace. More and more hospitals use medical information systems and/or (knowledge-based) documentation systems that enable the storage of electronic patient records (EPRs). Then, subsequent analysis of high-quality EPRs is a promising option. The quality of the stored examination records is determined by the documentation habits of the examiners, i.e., depending on the experience and training of the individual examiners. Therefore, the identification and analysis of documentation patterns of different examiners is a crucial task to improve the quality of the examinations and therefore of the whole database of patient records.

We propose a subgroup mining approach to analyze the inter-individual documentation quality of the examiners. Subgroup mining or subgroup discovery [Wrobel, 1997; Klösgen, 2002] is a promising technique for explorative and descriptive medical data mining that aims to discover "interesting" subgroups of individuals. Then, the subgroups can be defined as a subset of the target population with a distributional unusualness concerning a certain property we are interested in, e.g., in the subgroup of smokers with a positive family history the risk of coronary heart disease is significantly higher than in the general population.

Subgroup mining is especially suited for the sketched analysis task in the medical domain, since it does not necessarily focus on finding complete relations between the specific target concept and the explaining variables; instead, interesting partial relations are sufficient. Due to this criterion the discovered patterns do not necessarily fulfill high support criteria, which are necessary for other prominent data mining approaches, e.g., methods for association rule discovery [Agrawal and Srikant, 1994]. Furthermore, subgroup discovery methods do not depend on support measures, but on a quality function which is flexibly defined according to the criteria of the user.

Usually the ultimate goal of knowledge discovery methods is to identify novel, potentially useful, and interesting knowledge. However, in real-world settings novelty and interestingness criteria of the user often cannot be fully satisfied: quite similar to a search query submitted to a web search engine, (e.g., Google), the application of purely automatic methods can yield a huge number of (possibly uninteresting) results which are hard to handle. Then, a 'query refinement' needs to be considered. In order to perform the discovery process more intelligently, we propose the combination of a semi-automatic subgroup mining method guided by visualization and background knowledge.

We exemplify the approach in a case study based on the knowledge-based documentation and consultation system for sonography SONOCONSULT [Huettig *et al.*, 2004], which is in routine use in the DRK-hospital in Berlin/Köpenick: we identify profiles of examiners concerning their documentation habits for general quality control and management.

The rest of the paper is organized as follows: In Section 2 we introduce our method, i.e., a process model for knowledge-intensive subgroup mining. We describe suitable background knowledge for integration into the mining method and a visualization method to guide the user in the interactive discovery process. Finally, we provide the results of a case study of the presented approach with a fielded system in the medical domain in Section 3. We conclude with a summary of the paper in Section 4.

2 Methods: The Semi-Automatic Process for Knowledge-Intensive Subgroup Mining

Subgroup mining aims to discover "interesting" subgroups of individuals that are described by relations between independent (explaining) variables and a dependent (target) variable, rated by a certain interestingness measure. For example, two possible criteria are the difference in the distribution of the target variable concerning the subgroup and the general population, and the subgroup size. Subgroup mining does not necessarily focus on finding complete relations; instead partial relations, i.e., (small) subgroups with "interesting" characteristics can be sufficient.

In this section we first describe the process model for intelligent subgroup mining. After that, we define the subgroup mining task, and discuss the elements of the proposed process model in detail, i.e., helpful background knowledge applied for subgroup mining, and the core visualization method to guide the subgroup mining process. Finally, we discuss related work.

2.1 Process Model

The general goal of a subgroup mining task is to identify a set of highly interesting, diverse subgroups. Both the quality measures for the subgroup and the redundancy criteria heavily depend on the goals of the user. A purely automatic approach is often appropriate, if the analysis goals of the user are fixed during the search process. However, if the user wants to test specific hypotheses or already has a lot of background knowledge and experiences about the analysis domain, then an automatic search method may not always be transparent enough.

In the proposed mining process both interactive and automatic elements are combined: the automatic methods can be used to identify useful starting points for analysis, or for a quick "what if" analysis of the current situation. The presented approach includes the background knowledge and experiences of the user in order to focus the mining method on the interesting patterns, and to restrict the search space. Then, direct user interaction enables an *active mining* approach (e.g., [Gamberger *et al.*, 2003]). In this approach, the user is directly integrated into the subgroup discovery process and can manipulate the subgroup descriptions interactively. The process model is depicted in Figure 1.

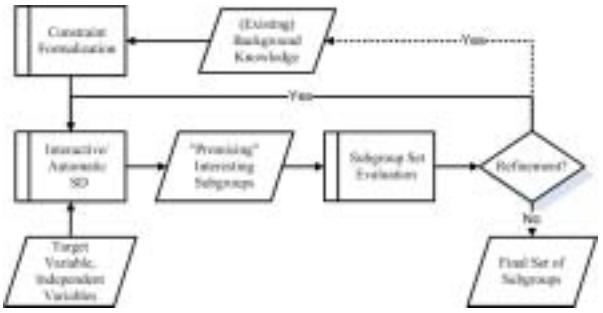


Figure 1: The Knowledge-Intensive Semi-Automatic Subgroup Mining Process

2.2 Subgroup Mining

We first introduce our knowledge representation schema before defining the subgroup mining task. After that, we describe the background knowledge and the visualization method used in the proposed subgroup mining process.

General Definitions

Let Ω_A the set of all attributes. For each attribute $a \in \Omega_A$ a range $dom(a)$ of values is defined. Furthermore, we assume $\mathcal{V}_A$ to be the (universal) set of attribute values of the form $(a : v)$, where $a \in \Omega_A$ is an attribute and $v \in dom(a)$ is an assignable value. A diagnosis attribute is represented by a binary attribute, i.e., for a diagnosis attribute $d \in \Omega_D, \Omega_D \subseteq \Omega_A$ we define a (boolean) range $dom(d) = \{established, not\ established\}$. Let CB be the case base containing all available cases. A case $c \in CB$ is defined as a tuple $c = (\mathcal{V}_c, \mathcal{D}_c)$, where $\mathcal{V}_c \subseteq \mathcal{V}_A$ is the set of attribute values observed in the case c. The set $\mathcal{D}_c \subseteq \mathcal{V}_A$ is the set of diagnoses describing the *solution* of this case.

Basic Subgroup Mining A subgroup mining task mainly relies on the following four main properties: the target variable, the subgroup description language, the quality function, and the search strategy. The target variable may be binary, nominal or numeric. Depending on its type, there are different analytic questions, e.g., for a numeric target variable we can search for significant deviations of the mean of the target variable.

A subgroup mining problem encapsulates the target variable, the search space of independent variables, the general population, and additional constraints.

Definition 1 (Subgroup Mining Problem). *A subgroup mining problem SP is defined as the tuple*

$$SP = (T, A, C, CB)\,,$$

where $T \in \Omega_A \cup \mathcal{V}_A$ is a target variable. $A \subseteq \Omega_A$ is the set of attributes to be included in the subgroup discovery process. CB is the case base representing the general population used for subgroup mining. C specifies (optional) constraints for the discovery method. We define Ω_{SP} as the set of all possible subgroup mining problems.

The definition above allows for arbitrary target variables. However, for our analytic questions we will focus on binary target variables, i.e., $T \in \mathcal{V}_A$.
The description language specifies the individuals from the reference population belonging to the subgroup.

Definition 2 (Subgroup Description). *A subgroup description $sd = \{e_i\}$ consists of a set of selection expressions (selectors) $e_i = (a_i, V_i)$ which are selections on domains of attributes, i.e., $a_i \in \Omega_A, V_i \subseteq dom(a_i)$. A subgroup description is defined as the conjunction of its contained selection expressions. We define Ω_{sd} as the set of all possible subgroup descriptions.*

A quality function measures the interestingness of the subgroup (c.f., [Klösgen, 2002] for examples).

Definition 3 (Quality Function). *A quality function*

$$q : \Omega_{sd} \times \Omega_{SP} \rightarrow R$$

evaluates a subgroup description $sd \in \Omega_{sd}$ given a subgroup mining problem $SP \in \Omega_{SP}$. It is used by the search method to rank the discovered subgroups during search.

For binary target variables, examples for quality functions are given by

$$q_{BT} = \frac{(p - p_0) \cdot \sqrt{n}}{\sqrt{p_0 \cdot (1 - p_0)}} \cdot \sqrt{\frac{N}{N - n}}, \quad q_{RG} = \frac{p - p_0}{p_0 \cdot (1 - p_0)},$$

where p is the relative frequency of the target variable in the subgroup, p_0 is the relative frequency of the target variable in the total population, $N = |CB|$ is the size of the total population, and n denotes the size of the subgroup. In contrast to the quality function q_{BT} (**B**inomial **T**est), the quality function q_{RG} (**R**elative **G**ain) only compares the target shares of the subgroup and the total population measuring the relative gain. Therefore, a suitable support thresholds is necessary to discover significant subgroups.

An efficient subgroup search strategy is necessary, since the search space is exponential concerning all the possible selectors of a subgroup description: commonly, a beam search strategy is used because of its efficiency [Klösgen, 2002]. We apply a modified beam search method, where an initial subgroup description can be selected as the initial value for the beam. Beam search iteratively expands the k best subgroup descriptions by adding the selector that provides the best quality improvement. Iteration stops, if the quality as evaluated by the quality function q does not improve any further.

For the characterization of the discovered subgroups we have two alternatives: Besides the principal factors contained in the subgroup description there are also supporting factors. These are attribute values $supp \subseteq \mathcal{V}_A$, which are characteristic for the containing subgroup, i.e., the value distributions of their corresponding attributes (supporting attributes) differ significantly comparing two populations: the true positive cases contained in the subgroup and non-target class cases contained in the total population. In addition to the principal factors the supporting factors can also be used to statistically characterize a discovered subgroup, as described, e.g. in [Gamberger and Lavrac, 2002].

Background Knowledge for Subgroup Mining

There are different classes of background knowledge which can be used in the knowledge-intensive process for subgroup mining, e.g., constraints, ontological knowledge, and abstraction knowledge. Knowledge acquisition is always expensive, so its costs should be minimized. Sometimes knowledge can be derived from already formalized knowledge, e.g., we can derive constraints from ontological knowledge, and thus reduce its acquisition costs. In the following, we summarize the individual knowledge elements; we refer to [Atzmueller *et al.*, 2005] for a more detailed discussion.

Constraints restrict the search process/space by specifying the attributes and attribute values of interest. In addition, a set of attribute values can be used to define additional meta values specific to the application domain. For example, for the diagnosis *cirrhosis of the liver* the values *possible* and *probable* can be defined as a disjunctive attribute value. Furthermore, constraints can also include quality and syntactical constraints that filter the mined patterns during the discovery process.

Ontological knowledge includes information about the domain ontology, e.g., abnormality information/normality information about attribute values indicating either abnormal/pathological states, or the normal state. For example, consider the attribute temperature with the value range $dom(\text{temperature}) = \{\text{normal}, \text{marginal}, \text{high}, \text{very high}\}$. The values *normal* and *marginal* denote normal states of the attribute, while the values *high* and *very high* describe abnormal states. Using abnormality information, we can define meta values containing several attribute values with certain abnormality categories.

Similarity information about attribute values relates to the relative similarity between attribute values. Significant similarities between attribute values can indicate that the respective values can be combined into a new value. Then, appropriate meta values need to be defined. A high attribute weight specifies, that an attribute is relatively important.

Ordinality information is used to indicate the ordinal attributes which can be used to construct certain 'ordinal groups', e.g., summarizing certain consecutive age groups. In general, specifying appropriate meta values can significantly increase the interpretability of mined subgroup patterns for the domain specialist (c.f. Section 3).

Derived attributes (abstraction knowledge) play a special role in the mining process. These attributes are constructed according to the needs of the user, e.g., intermediate concepts which are not contained in the set of basic attributes can be modelled, or attributes can be constructed such that missing values are minimized.

In Table 1, we summarize the different classes and types of background knowledge (CK = constraint knowledge, OK = ontological knowledge, AK = abstraction knowledge). We show their characteristics in terms of the 'derivable knowledge' if applicable, their costs, and their potential contribution to restricting the search space and/or focusing the search process for a qualitative comparison. The individual ratings are based upon our experiences and feedback provided by the domain specialists, e.g., during the case study in Section 3. Considering the costs/impact of the knowledge elements for subgroup mining, the label - indicates no cost/impact; the labels +, ++, and +++ indicate increasing costs and impact. A +(+) signifies, that the respective element has low costs if it can be derived/learned, and moderate costs otherwise. Similarly ++(+) indicates this for moderate and high costs, respectively.

Knowledge		*Derivable*	*Cost*	Search Space	
Class	*Type*	*Knowledge*		Restr.	Focus
CK	Syntactical Constr.	–	+	+	+
CK	**Quality Constr.**	–	+	++	++
CK	Attr. Values Constr.	–	+(+)	+	+
CK	Meta Values Constr.	–	+(+)	–	++
CK	**Attributes Constr.**	–	+(+)	++	++
OK	**Normality** Info	Attr. Val. Constr.	+	+	++
OK	Abnormality Info	Attr. Val. Constr.	++	+	++
		Meta Val. Constr.		–	++
OK	Similarity Info	Meta Val. Constr.	++(+)	–	++
OK	**Ordinality Info**	Meta Val. Constr.	+	++	+++
OK	Attr. Weights	Attr. Constraints	+(+)	+	++
AK	***Derived Attributes***	Derived Attributes	+++	+++	+++

Table 1: Background Knowledge for Subgroup Mining

The most important types of background knowledge with an especially good cost/benefit ratio concerning the subgroup mining task are indicated in bold type.

Figure 2: The Zoomtable

Guiding Subgroup Mining by Visualization Techniques

In this section we present the main visualization for subgroup discovery, i.e., the *zoomtable* depicted in Figure 2. This visualization is associated with the *current subgroup view* (Annotation I) showing the target variable and the selectors of the current subgroup. The bars (Annotation II) depict the target distributions in the whole population (upper bar), and in the subgroup. The left part of a bar shows the positives, the right part the negative instances. The zoomtable (Annotation III) shows the distribution of the data restricted to the currently selected subgroup. Each row of the zoomtable shows the value distribution of a specific attribute. The width of a cell relates to the frequency of an attribute value. The zoomtable is updated when the user modifies the current subgroup, e.g., by adding a selector from the zoomtable.

Figure 3 shows a row of the zoomtable concerning a binary attribute with the values *yes* and *no*. The important parameters for subgroup mining w.r.t. a "future" subgroup are the *subgroup size* – given by the width of a specific selector cell, and the *target share* (precision), i.e., the share of subgroup instances containing the target variable (positive instances). In the current subgroup SG_c, (a) indicates the

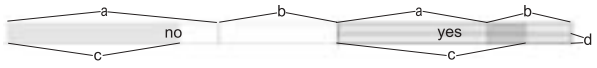


Figure 3: The Zoomtable – Detail View

(currently) positive instances, and (b) denotes the negative ones. In the 'next' subgroup SG_n, i.e., including the particular attribute value, (c) shows the positive instances for this subgroup, which can be compared to (a). So, if (c) is larger than (a), then the precision increases adding this selector. Finally, (d) shows the gain in precision, comparing the subgroups SG_c and SG_n: if the height of (d) is zero, the precision does not increase. If it fills the entire bar, then the precision reaches 100%.

The zoomtable enables the user to directly manipulate the subgroup and to estimate the effects of individual selectors. Furthermore, interesting attributes and their values are easy to spot due to the visual markers in the respective cells. Then, an active subgroup mining approach (c.f., [Gamberger *et al.*, 2003]) can be implemented quite easily.

2.3 Related Work

The application of subgroup mining especially for the medical domain using the guidance of an expert is described in [Gamberger and Lavrac, 2002; Gamberger *et al.*, 2003]. This active approach stresses the interaction between the expert and the system to identify interesting subgroups. However, in the semi-automatic process mainly the parameters of the search process can be adapted. In our semi-automatic process, the domain specialist can adapt the subgroup mining problem by including background knowledge, and modifying the search process directly guided by interactive visualizations.

The proposed interactive core component, i.e., the zoomtable visualization was inspired by the *InfoZoom* system [Spenke, 2001]. InfoZoom also visualizes the value distributions of attributes in single rows of a table, and also allows the user to zoom in on individual values. However, our approach extends this idea significantly, since we also guide the user during the subgroup mining process by visualizing additional quality parameters directly in the zoomtable, e.g., the future target share or the gain of a specific selector. Changes in the zoomtable, e.g., adding/removing selectors to the current subgroup (description) are also visualized dynamically.

Using background knowledge to constrain the search space and pruning hypotheses during the search process has been proposed in ILP approaches. [Weber, 2000] proposes *require*- and *exclude*-constraints for attribute – value pairs, in order to prune the search space. [Zelezny *et al.*, 2003] integrate constraints into an ILP approach as well; the used constraints are mainly concerned with syntactical and quality constraints w.r.t. the discovered subgroups.

The main difference between the presented approach and the existing approaches is the fact, that we are able to integrate several new types of additional background knowledge. This knowledge can be refined incrementally according to the requirements of the mining task. In our process model for semi-automatic and knowledge-intensive subgroup mining we aim to focus the discovery method on the interesting patterns using background knowledge. Then, interactive exploration is made more convenient, since mostly interesting patterns/factors are presented. Furthermore, we apply a novel visualization technique in an active and user-centered approach that is usually more transparent for (experienced) users.

3 Results: Case Study

In this section we describe a case study for the application of the proposed subgroup mining process.

We first introduce the analysis task w.r.t. its clinical relevance. Then, we describe the documentation and consultation system SONOCONSULT. After that, we present and discuss the results of the case study.

3.1 Profiling Examiners for Quality Control

Our application domain is the domain of sonography. Sonographic examination and documentation is highly dependent on the skills of the examiners. Individual examiners rotate according to a defined schedule (e.g., every 6 months). Before performing the examinations, they get special training and can always consult experienced colleagues. However, while performing the examination they are on their own. Then, it is easy to see that the quality of the examinations is dependent on the individual experience and skills of the examiners. Therefore, documentation and interpretation habits of examiners may differ significantly, which is problematic considering the consistency and quality of the documented examinations; e.g., some examiners may be more competent in identifying specific symptoms concerning certain diagnoses or organ systems than others.

While a gold standard for the correct examination and documentation is not available in sonography, the detection of systematic discrepancies among different examiners is clinically important in itself. To identify deviations regarding the documentation habits of examiners, subgroup mining is used to discover novel and unexpected (documentation) patterns, i.e., certain symptom combinations that are observed significantly more (in-)frequently in conjunction with certain examiners.

3.2 The Documentation and Consultation System SonoConsult

We use cases taken from the SONOCONSULT system [Huettig *et al.*, 2004] – a medical documentation and consultation system for sonography – which has been developed with the knowledge system D3 [Puppe, 1998]. The system is in routine use in the DRK-hospital in Berlin/Köpenick and documents an average of about 300 cases per month. These are detailed descriptions of findings of the examination(s), together with the inferred diagnoses (binary attributes). The derived diagnoses are usually correct as shown in a medical evaluation (c.f. [Huettig *et al.*, 2004]), resulting in a high-quality case base with detailed case descriptions. The applied SONOCONSULT case base contains 7096 cases. The domain ontology contains 427 basic attributes with about 5 symbolic values on average, 133 symptom interpretations, which are rule-based abstractions of the basic attributes, and 221 diagnoses.

3.3 Results

The domain specialist performed subgroup mining considering individual diagnostic areas and organ systems, e.g., liver and kidney diseases, using the [VIKAMINE, 2005] system. Then, the relevant factors that were important for deriving the diagnoses of a certain area were identified; these were then provided to the subgroup method in order to constrain the search space and to focus the search method. Furthermore, the domain specialist provided normality information to filter out some uninteresting *normal* values, e.g., *liver vessels = normal*. Meta Values were defined to build disjunctive meta values, e.g., *liver plasticity = moderately or strongly reduced*. Additionally, several derived attributes (abstractions) were defined to limit missing values. For example, diagnostic attributes like *cirrhosis of the liver* were either defined or tuned in order to minimize missing values by providing a default *normal* value. After that, the proposed process model was applied, using beam-search as the automatic component for subgroup mining.

We show examples of the results in Table 2, considering liver diseases, especially focussing on *cirrhosis of the liver*. The cases that were used in the case study were acquired by 8 different examiners (E1 - E8). Concerning liver examinations, each examiner contributed 200-600 cases, resulting in a total population of 3931 cases where an examination of the liver was performed. Then, we analyzed the individual factors concerning the individual examiners as the target variable (column *E*). We used the relative gain quality function q_{RG} (c.f., Section 2.2), which was easy to interpret for the experts. Then, deviations concerning findings or combinations of findings were measured.

Each row of the table depicts a subgroup with the subgroup parameters *Size* (subgroup size), *TP* (true positives), *FP* (false positives), *Pop.* (the defined population), the default and subgroup target share p_0 and p, respectively, and *RG*, i.e., the value of the relative gain quality function q_{RG}.

#	E	LP		LS		LE		LV		LC		Subgroup Parameters						
		mr	sr	uk	kn	mi	si	rp	tp	po	pr	Size	TP	FP	Pop.	p0	p	RG
1	E1			X								221	44	177	2295	0.164	0.199	0.24
2	E1						X					435	41	394	2295	0.164	0.094	-0.51
3	E1	X	X									420	28	392	2295	0.164	0.066	-0.71
4	E1				X							13	0	13	2295	0.164	0	-1.19
5	E2		X									248	19	229	2295	0.123	0.076	-0.43
6	E2							X				689	25	664	2294	0.123	0.036	-0.8
7	E3	X	X					X		X		129	91	38	2294	0.129	0.705	5.12
8	E3	X										248	116	132	2295	0.128	0.467	3.01
9	E3							X		X	X	385	131	254	2294	0.129	0.34	1.87
10	E3	X	X									420	132	288	2295	0.128	0.314	1.64
11	E3			X								13	4	9	2295	0.128	0.307	1.59
12	E3								X		X	102	0	102	2294	0.13	0	-1.14
13	E5			X								13	9	4	2295	0.057	0.692	11.8
14	E5	X				X	X					227	85	142	2295	0.057	0.374	5.89
15	E5	X										248	87	161	2295	0.057	0.35	5.45
16	E5	X	X									420	96	324	2295	0.057	0.228	3.18
17	E5									X	X	440	56	384	3918	0.053	0.127	1.46
18	E5	X	X			X	X	X		X	X	271	39	232	2294	0.057	0.143	1.61
19	E5			X								221	6	215	2295	0.057	0.027	-0.55
20	E5			X		X		X		X	X	109	0	109	2294	0.058	0	-1.06

LP = Liver Plasticity
mr = moderately reduced
sr = strongly reduced

LS = Liver surface
uk = uneven, knotty
kn = knaggy

LC = Cirrhosis of the liver
po = possible
pr = probable

LE = Liver Echogenicity
mi = moderately increased
si = strongly increased

LV = Liver Vessels
rp = rarefication of portal branches
tp = tapering of portal branches

Table 2: Interesting subgroups and individual factors concerning liver diseases. The first line depicts the subgroup (target variable *Examiner=E1*) described by *Liver surface = uneven, knotty* with a target share of 19.9% (p) in the subgroup compared to 16.4% (p_0) in the total population with a relative gain of 24% (RG).

Applying the process model, the domain specialist considered the visualization component very helpful, since it enabled an easy step by step analysis: single factors could be identified first, and then subgroups were refined. Furthermore, subgroups discovered by the automatic search method were also validated and refined interactively.

3.4 Discussion

The results in Table 2 show significant differences in the documentation habits of the individual examiners. Negative relative gain (RG) values indicate that the examiner documented/interpreted certain findings less frequently than his colleagues, while a positive relative gain indicates the opposite. For a comprehensive overview, we also show some single factors in addition to significant combinations, which were also very interesting for the domain specialist. Especially significant deviations are shown in lines 7, 14 and 15, which are very descriptive for the respective examiners. Line 7 also shows a significant correlation with the diagnosis *cirrhosis of the liver* combined with the relevant findings.

Lines 4, 11, and 13 show a surprising result: the examiners E3 and E5 are the only examiners that document a specific finding, i.e., *Liver surface = knaggy* in comparison to their colleagues. Further investigation turned up that the specific attribute value was added to the consultation system in a later step. Therefore, only some examiners had the opportunity to use this finding.

Furthermore, as shown in the table, examiner E5 (lines 14-20) deserves special attention, since the shown documentation habits differed most significantly compared to the peer examiners. Especially interesting were the subgroups depicted in line 17, 18 and 20: it is easy to see that examiner E5 documents a *cirrhosis of the liver = probable or possible* more frequently than his peers. An even more significant subgroup is shown in line 18 that shows a specialization of the subgroup in line 17. For the very indicative finding combination in line 20 (regarding the diagnosis cirrhosis of the liver) even no case of E5 could be identified. It is striking that E5 uses very special patterns for inferring the diagnosis *cirrhosis of the liver* compared to his colleagues: e.g., symptoms of plasticity are much more frequent (lines 14-16) whereas *liver surface = uneven, knotty* is significantly infrequent (lines 19, 20).

In summary, these results show a high variability of documentation and interpretation habits of the different examiners. They indicate the need for further prospective studies. These results are a starting point for initiating a discussion on training or standardization actions to increase the inter-examiner homogeneity of the sonographic reports.

4 Conclusion and Outlook

In this paper we presented an approach for semi-automatic and knowledge-intensive subgroup mining. We exemplified the approach in a case study in the medical domain of sonography, where we were able to extract interesting profiles of examiners concerning their documentation habits. The proposed approach applies background knowledge and visualization to guide the subgroup mining process, which was regarded as extremely important by the domain specialist. The obtained results are a first step toward surveying the documentation performance of individual examiners, and to support their learning phase.

In the future, we are planning to embed a component for subgroup analysis in knowledge-based documentation systems directly. A prerequisite is a comprehensive analysis applying the presented method to identify interesting patterns. Then, using these patterns, the completeness of findings regarding specific examiners can be checked instantly. This provides a transparent survey of general documentation habits and the potential for training certain examiners.

References

[Agrawal and Srikant, 1994] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.

[Atzmueller *et al.*, 2005] Martin Atzmueller, Frank Puppe, and Hans-Peter Buscher. Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In *Proc. 19th International Joint Conference on Artificial Intelligence (IJCAI-05), to appear*, 2005.

[Gamberger and Lavrac, 2002] Dragan Gamberger and Nada Lavrac. Expert-Guided Subgroup Discovery: Methodology and Application. *Journal of Artificial Intelligence Research*, 17:501–527, 2002.

[Gamberger *et al.*, 2003] Dragan Gamberger, Nada Lavrac, and Goran Krstacic. Active Subgroup Mining: a Case Study in Coronary Heart Disease Risk Group Detection. *Artificial Intelligence in Medicine*, 28:27–57, 2003.

[Huettig *et al.*, 2004] Matthias Huettig, Georg Buscher, Thomas Menzel, Wolfgang Scheppach, Frank Puppe, and Hans-Peter Buscher. A Diagnostic Expert System for Structured Reports, Quality Assessment, and Training of Residents in Sonography. *Medizinische Klinik*, 99(3):117–122, 2004.

[Klösgen, 2002] Willi Klösgen. *Handbook of Data Mining and Knowledge Discovery*, chapter 16.3: Subgroup Discovery. Oxford University Press, New York, 2002.

[Puppe, 1998] Frank Puppe. Knowledge Reuse among Diagnostic Problem-Solving Methods in the Shell-Kit D3. *Intl. Journal of Human-Computer Studies*, 49:627–649, 1998.

[Spenke, 2001] Michael Spenke. Visualization and Interactive Analysis of Blood Parameters with InfoZoom. *Artificial Intelligence In Medicine*, 22(2):159–172, 2001.

[VIKAMINE, 2005] http://ki.informatik.uni-wuerzburg.de/projects/dm, 2005.

[Weber, 2000] Irene Weber. Levelwise search and Pruning Strategies for First-Order Hypothesis Spaces. *Journal of Intelligent Information Systems*, 14:217–239, 2000.

[Wrobel, 1997] Stefan Wrobel. An Algorithm for Multi-Relational Discovery of Subgroups. In *Proc. 1st European Symposion on Principles of Data Mining and Knowledge Discovery*, pages 78–87, Berlin, 1997. Springer Verlag.

[Zelezny *et al.*, 2003] Filip Zelezny, Nada Lavrac, and Saso Dzeroski. Using Constraints in Relational Subgroup Discovery. In *International Conference on Methodology and Statistics*, pages 78–81. University of Ljubljana, 2003.

Data analysis based on subgroup discovery: Experiments in brain ischaemia domain

Dragan Gamberger
Rudjer Bošković Institute
Zagreb, Croatia
dragan.gamberger@irb.hr

Antonija Krstačić
Univ. Hospital of Traumatology
Dept. of Neurology,
Zagreb, Croatia

Goran Krstačić
Institute for Cardiovascular Prevention and Rehabilitation,
Zagreb, Croatia

Nada Lavrač
Jožef Stefan Institute
Ljubljana, Slovenia

Michèle Sebag
Université Paris-Sud
Orsay, France

Abstract

This paper presents insightful analysis of medical data collected in regular hospital practice. The domain consists of patients suffering from brain ischaemia, either permanent as brain attack (stroke) with positive computer tomography (CT) or reversible ischaemia with normal brain CT test. The goal of the analysis is the extraction of useful knowledge that can help in diagnosis, prevention and better understanding of vascular brain disease. The work demonstrates the applicability of subgroup rule induction as the basis for insightful data analysis and describes intellectual process of converting rules into reasonable medical concepts. Detection of coexisting risk factors, selection of relevant discriminative points for numerical descriptors, as well detection and description of characteristic patient subpopulations are important results of the analysis. Graphical representation is extensively used to illustrate the detected regularities.

1 Introduction

Data analysis in medical applications is characterized by the ambitious goal of extracting potentially new relationships from the data, and providing insightful representations of detected relationships. Applications of quantitative statistical methods seldom lead to insightful results, leaving a large workload on the human experts who have to provide appropriate interpretations of results, with no guarantees that—due to a huge search space of possible solutions—the most relevant combinations have been tested at all [Fayyad *et al.*, 1996]. The goal of intelligent data analysis is to effectively detect most relevant dependencies in an explicit qualitative form and to enable that quantitative analysis and human expert interpretation can concentrate on a relatively small set of potentially relevant hypotheses. This approach is specially suited for medical data analysis, as large amounts of available medical expert knowledge allow for appropriate interpretation of detected relations.

This work demonstrates that rules induced by the existing methodology of supervised subgroup discovery [Gamberger *et al.*, 2003] can serve as an appropriate basis for data analysis, if supplemented by the sufficient intellectual effort of medical experts, willing to convert machine-induced rules into adequate medical interpretations. The proposed approach, applied to a typical database collected in regular hospital practice describing brain ischaemia patients, is used to illustrate this expert-guided approach to knowledge discovery. The next section presents the problem domain. Section 3 presents the proposed data analysis approach leading to insightful knowledge, interpreted by medical specialists in Section 4.

2 Brain ischaemia data

The database consists of records of patients who have been treated in the Intensive Care Unit of the Department of Neurology, University Hospital Center "Zagreb", in Zagreb, Croatia during the year 2003. In total, 300 patients are included in the database: 209 with the confirmed diagnosis of brain attack (stroke), and 91 patients who entered the same department with adequate neurological symptoms and disorders, but were diagnosed (based on the outcomes of neurological tests) as patients with transition ischaemic brain attack (TIA, 33 patients), reversible ischaemic neurological deficit (RIND, 12 patients), and serious headache or cervical spine syndrom (46 patients). In this paper, the goal of data analysis experiments is to discover regularities that characterize brain stroke patients.

Patients are described with 27 different descriptors representing anamnestic data, physical examination data, laboratory test data, ECG data, CT test result and information about previous hospital therapies. Descriptors used in the analyses are listed in Table 1.

It must be noted that the control group does not consist of healthy persons but patients with serious neurological symptoms and disorders. In this sense, the available database is particularly appropriate for studying specific characteristics and subtle differences that distinguish patients with stroke. The detected relationships can be accepted as true characteristics for these patients. However, the computed evaluation measures—including probability, specificity and sensitivity of induced rules—only reflect characteristics specific to the available data, not necessarily holding for the general population or other medical institutions [Victor and Ropper, 2001].

3 Data analysis process

This section presents the data analysis process, using rules induced by the SD subgroup discovery algorithm [Gam-

Descriptor	Abbreviation
sex (f,m)	sex
age (years)	age
family anamnesis (n,p)	fhis
present smoking (y,n)	smok
stress (y,n)	str
alcohol consumption (y,n)	alcoh
systolic blood pressure cont. (mmHg) normal value < 139 mmHg	sys
diastolic blood pressure continuous (mmHg) normal value < 89 mmHg	dya
uric acid continuous ($\mu mol\ L^{-1}$) ref. value for men < 412 ref. value for women < 380	ua
fibrinogen continuous ($g\ L^{-1}$) ref. value 2.0-3.7	fibr
glucose continuous ($mmol\ L^{-1}$) ref. value 3.6-5.8	gluc
heart rate continuous ref. value $60 - 100$ /min	ecgfr
atrial fibrillation (y,n)	af
left ventricular hypertrophy (y,n)	ecghlv
aspirin therapy (y,n)	asp
anticoagulant therapy (y,n)	acoag
antihypertensive therapy (y,n)	ahyp
antiarrhytmic therapy (y,n)	aarrh
statins (antihyperlipoproteinaemic t.) yes, no	stat
hypoglycemic therapy none, yesO (oral), yesI (insulin)	hypo

Table 1: List of most relevant descriptors in the brain ischaemia domain with abbreviations used in induced rules. Included are also reference values representing the range typically accepted as normal in the medical practice.

berger *et al.*, 2003]. The process begins with a set of rules that are obtained by repetitively applying the SD algorithm with different generalization parameter values. In the experimental setting determined for the ischemia domain, the process of expert-guided subgroup discovery was performed as follows. The SD algorithm was run for values g in the range 5 to 100, and a fixed number of selected output rules equal to 3. The rules induced in this iterative process were shown to the expert for selection and interpretation. The intention of this paper is to illustrate what type of insights are possible by the analysis based on individual rules and what can be additionally obtained if rules are analysed in groups. The SD algorithm,[1] described in detail in [Gamberger *et al.*, 2003] is—due to paper length restrictions—out of scope of this paper.

The basic characteristic of the presented approach is supervised learning of subgroup defining rules that characterize the target (positive) class cases (in this domain stroke cases) in contrast to cases in the non-target (negative or control) class (in this domain transitory ischaemia cases). This means that examples of two classes have to be available. Sometimes the decision about what is the target class is not simple and the complete data analysis process can have a few task definitions with different choices of target and non-target classes. For example, in the same brain ischaemia domain the target class could be also patients with stroke taking some therapy, and the non-target class being stroke patients not taking the therapy. In this setting, the process of data analysis is far from completely automatic. Moreover, the process should be sometimes repeated for different subpopulations with specific properties, like sex or age range, or with different subsets of descriptors. In this section we demonstrate only the process performed for the complete database with patients who experienced stroke selected as the target class. We have performed a series of experiments also with patients separated in different age and sex groups, some of them also with reduced descriptor sets. Although the results are very interesting, specially due to the possibility of the comparative analysis of rules, they are not included in this paper due to space restrictions.

[1]The algorithm is available as part of the publicly available Data Mining Server at http://dms.irb.hr, and can be used to induce rules for domains with up to 250 cases.

4 Results of rule analysis

Table 2 presents rules generated for the class stroke. There are in total 15 rules, three for each of the five selected g-values in the range $5 \leq g \leq 100$. By selecting a low g-value, the subgroup discovery algorithm tends to construct very specific rules with relative low sensitivity. With the increase of the g parameter the sensitivity typically improves at the cost of decreased specificity. The sensitivity and the specificity values for each rule are given in columns 3 and 4, respectively. The last column indicates the overlap between the current rule and one/two rules induced previously for the same g-value. The overlap value is defined as the number of positive cases that are covered both by the current rule and the previously generated rule(s) divided by the number of positive cases covered by either the current rule or the previosly generated rule(s), whichever is the smaller. Low overlap values mean relative independence between the rules.

Because inductions with different generalization parameters are independent, there is a possibility that the same rule (e.g. ahyp=yes) is induced with different generalization parameter values. The order of rules in each group is the order selected by the algorithm and it is determined by the rule quality value and the rule covering properties.

4.1 Analysis of individual rules

The interpretation of induced rules starts by independent interpretation of each individual rule. There is no apriori preference of either more specific or more sensitive rules. Highly sensitive rules, like those induced with parameter $g = 100$ describe general characteristics of the target class. In the given domain we see that stroke is characteristic for middle aged or elderly population $(age > 52.00)$, that people with the stroke typically have normal or increased dyastolic blood pressure $(dya > 75.00)$, and that they have already detected hypertension problems and take some therapy (anti-hypertension therapy yes). We also see that the

Ref.	Rule		
	Sens.	**Spec.**	**Overlap**
generalization parameter value 5			
g5a	$(fibr > 4.55) and (str = no)$		
	25%	100%	-
g5b	$(fibr > 4.45) and (age > 64.00)$		
	41%	100%	94%
g5c	$(af = yes) and (ahyp = yes)$		
	28%	95%	36%
generalization parameter value 10			
g10a	$(fibr > 4.45) and (age > 64.00)$		
	41%	100%	-
g10b	$(af = yes) and (ahyp = yes)$		
	28%	95%	34%
g10c	$(str = no) and (alcoh = yes)$		
	28%	95%	67%
generalization parameter value 20			
g20a	$(fibr > 4.55)$		
	46%	97%	-
g20b	$(ahyp = yes) and (fibr > 3.35)$		
	65%	73%	71%
g20c	$(sys > 153.00) and (age > 57.00) and (asp = no)$		
	45%	88%	80%
generalization parameter value 50			
g50a	$(ahyp = yes)$		
	74%	54%	-
g50b	$(fibr > 3.35) and (age > 58.00)$		
	79%	63%	76%
g50c	$(age > 52.00) and (asp = no)$		
	64%	63%	96%
generalization parameter value 100			
g100a	$(age > 52.00)$		
	96%	20%	-
g100b	$(dya > 75.00)$		
	98%	8%	98%
g100c	$(ahyp = yes)$		
	74%	54%	100%

Table 2: Rules induced for generalization parameter g values in the range[5,100]. Presented are their sensitivity and specificity values measured on the available data set as well as their overlap with previouly induced rule(s) in the same g-value group.

selected boundary values are relative low (52 years for the age and 75 mmHg for the dyastolic pressure) which is due to the fact that the rules should satisfy a large number of cases. This is the reason why the rules are not applicable as decision rules but they give useful descriptive information about the target class.

Expert interpretation of each individual rule is essential for the generation of useful knowledge. For example, the interpretation of rules like $(age > 52.00)$ or $(dya > 75.00)$ is straightforward. In contrast, the interpretation of the rule $(ahyp = yes)$ could lead to the conclusion that antihypertensive therapy itself is dangerous for the incidence of stroke. A much better interpretation is that hypertension is dangerous and because of that people with detected hypertension problems, characterized but the fact that they already take antihypertensive therapy, have larger probability of having a stroke. Indirectly, this rule also means that we have little chance to recognize the danger of high

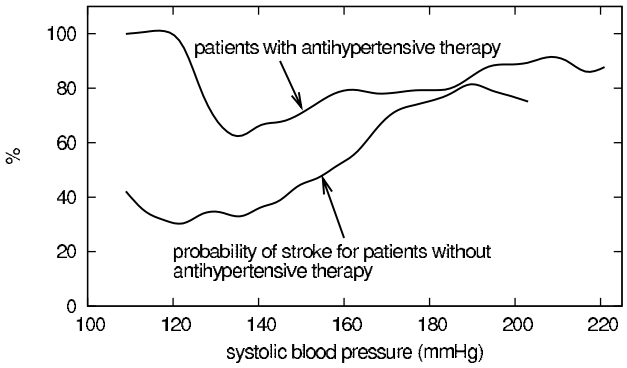


Figure 1: The proportion of patients with brain attack (stroke) in dependence of the total number of patients in the hospital department presented separately for patients with and without antihypertensive therapy for different systolic blood pressure values.

blood pressure directly from their measured values because many serious patients have these values artificially low due to a previously prescribed therapy. This is a good example of expert reasoning stimulated by an induced rule. In this situation we try to answer the question how the probability of stroke with respect to the transitory ischemia cases changes with the increasing systolic blood pressure. From the rule we have learned that we should compare only patients without anti-hypertension therapy. The result is presented in Figure 1. It can be noticed that the probability of stroke grows significantly with the increase of systolic blood pressure. The same dependency can be drawn also for the patients with the therapy. The differences between the two curves are significant and a few potentially relevant conclusions can be made. The first is that antihypertensive therapy helps in reducing the risk of stroke: this can be concluded from the fact that the probability of stroke is decreasing with the decrease of systolic blood pressure also for the patients with the therapy (as long as the systolic blood pressure is not lower than 130 mmHg). But it is also true that for systolic blood pressure between 130 and 170 mmHg the probability of stroke is significantly higher for patients with recognized hypertension problems than for other patients. The interpretation is that also in cases when successful treatement of hypertension is possible, the risk of stroke still remains relatively high and it is higher than for patients without hypertension problems.

As noticed earlier, very general rules are good for extracting general properties of the target class. In contrast to that, very specific rules induced by generalization parameter values 5 or 10 are good as classification rules for the target class. For example rule g5c $(af = yes) and (ahyp = yes)$ well reflects the existing expert knowledge that hypertension and atrial fibrillation are important risk factors for the stroke. The rule is significant as it emphasizes the importance of the combination of these two risk factors, what is not a generally known fact. The relevancy of detected correlation is illustrated in Figure 2. It shows that the probability of stroke is at least 85% in the age range 55 - 80 years for persons with both risk factors measured on the available hospital population. We can not estimate this

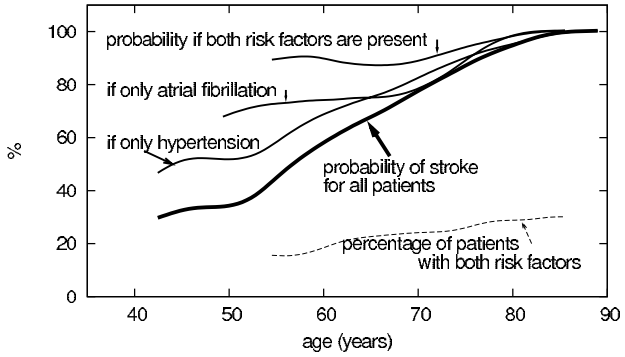


Figure 2: Probability of stroke in dependence of patient age presented for all patients in the available hospital population (thick line), probability of stroke for persons with hypertension problems, with atrial fibrillation problems, and with both hypertension and atrial fibrillation problems (thin solid lines). The percentage of patients with both risk factors is about 20-25% for the available hospital population (dashed line). The curves are drawn only for the range with a sufficiently large numbers of patients in the database.

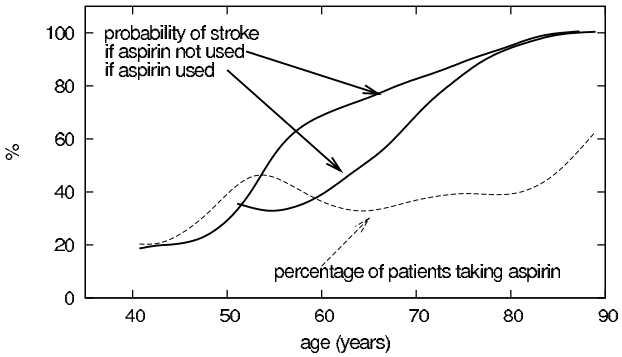


Figure 3: The probability of stroke in dependence of patient age presented for patients taking aspirin as the prevention therapy, and the probability of stroke for patients without this therapy. The percentage of patients with the aspirin therapy is presented by a dashed line.

probability on the general population but we can assume that it is even larger. The observation might be important for prevention purposes in general medical practice, especially because both factors can be easily detected.

Other two rules induced for g-value equal 5 contain conditions based on the fibrinogen values about 4.5 or more (reference values for negative fibrinogen finding are in the range 2.0 - 3.7 $g \cdot L^{-1}$). The rules without doubt demonstrate the importance of high fibrinogen values for the stroke patients. In the first rule the second necessary condition is the absence of stress, while in the second rule the second condition is age over 64 years. The interpretation of the second rule is relatively easy, leading to the conclusion that fibrinogen above 4.5 is itself very dangerous, which is confirmed also by rule g20a, being especially dangerous for elderly people. The interpretation of rule $(fibr > 4.55) and (stres = no)$ is not so easy because it includes contradictory elements 'high fibrinogen value' and 'no stress', knowing the fact that stress increases fibrinogen values and increases the risk of stroke. The first part of the interpretation is that 'no stress' is characteristic of elderly people and this conclusion is confirmed by the high overlap value of rules g5a and g5b (see the last column for the g5b rule). The second part of the interpretation is that high fibrinogen values can be the result of stress and such fibrinogen is not as dangerous for stroke as fibrinogen resulting from other changes in the organism.

From the rules induced with generalization parameter values 10–50 we notice that conditions on age and fibrinogen values repeat often, confirming already made conclusions about their importance. Also they suggest much more reasonable boundary values for the numerical descriptors (age over 57 or 58 years, fibrinogen over 3.3, systolic blood pressure over 153) which, if different from generally accepted reference values, can initialize research in the direction of accepting them as new decision points in medical decision making practice.

Also rules in this middle range of parameter g stress relevant relations among different descriptors like $(ahyp = yes) and (fibr > 3.35)$ or $(age > 52.00) and (asp = no)$. The later rule stimulated the analysis presented in Figure 3 which seems as excellent motivation for patients to accept prevention based on aspirin therapy. It can be easily noticed that the inductive learning approach correctly recognized the importance of the therapy for persons older than 52 years.

4.2 Analysis of rule groups

Besides the possibility to analyse each rule separately, combinations of co-occurring rules can give some additional information. In this respect it is useful to look at the overlap values of rules. A good example is a group of three rules induced for g-value 10. These rules have low overlap values, meaning that they describe relative diverse subpopulations of the target class. Their analysis enables global understanding of the hospital population in the Intensive Care Unit of the Neurology Department. Results of the analysis are presented in Figure 4.

The figure graphically and numerically illustrates the importance of each population subgroup and its overlap with other subgroups. The textual description is also important, reflecting the results of basic statistical analysis (mean values of age and fibrinogen, as well as sex distribution) for the subpopulation described by the rule, followed by the so-called supporting factors. The supporting factors are those descriptor values that are characteristic for the subpopulation in contrast to the cases in the negative class. The importance of these factors lies in the fact that they can help to confirm that a patient is a member of a subpopulation, also giving a better description of a typical member of a subgroup. The results show that the induced subgroups describe three relatively different types of stroke among elderly people (mean age between 70 and 75 years).

The largest subgroup can be called *elderly patients*; it is characterized by extremely high fibrinogen values (mean value 5.5) and increased glucose values (mean value 8.4). In most cases these are women (about 70%) that do not smoke, do not suffer from stress, and do not have problems with lipoproteins. Very different is the subpopula-

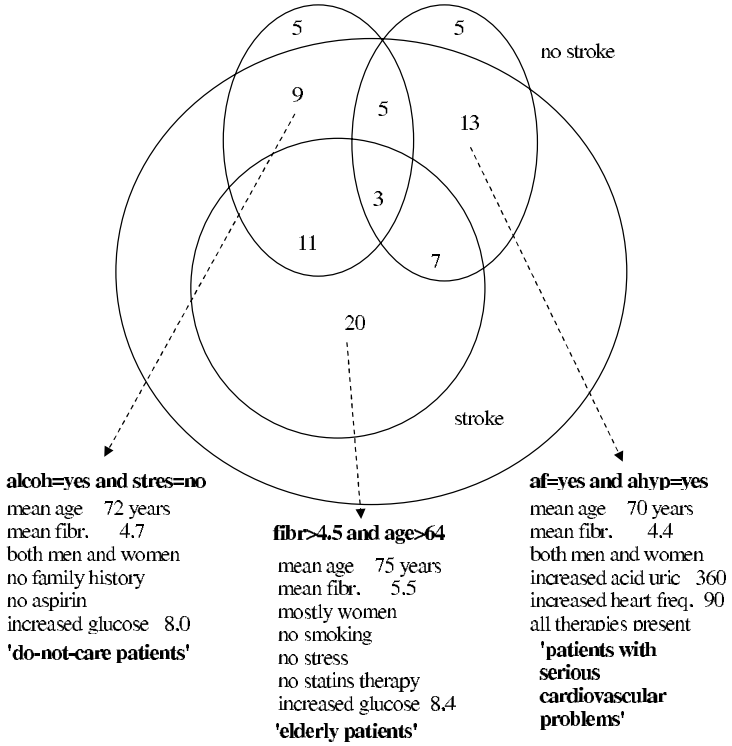


Figure 4: Comparative study of three important subgroups of stroke patients detected by rules induced with generalization parameter value 10. The large circle presents the stroke patients, negative cases are outside the large circle. Small circles present three detected subgroups. One of them includes only positive cases while the other two include also a small part of negative cases. The numbers present the percentages of patients that satisfy the conditions of one, two, or all three rules. In total, 68% of positive cases are included in at least one subgroup. The definitions of patient groups (in bold-face letters) are followed by a list of most relevant properties that characterize the patient group. The list ends with the expert's name given to the group (in bold-face letters).

tion that can be called *patients with serious cardiovascular problems* characterized with diagnosed hypertension and atrial fibrillation. It is a mixed male-female population. Its main characteristic is that they typically receive many different therapies but still they have increased—but inside reference—heart rate frequency (about 90) and acid uric (about 360). In between these two populations—in terms of age—is a subpopulation that can be called *do-not-care patients* characterized by alcohol consumption and no stress. It also a mixed male-female population characterized only by the increased glucose values of laboratory tests. It seems as these people would have the largest chance not to be among patients with stroke because their relevant property is negative family history. Their do-not-care attitude is visible also from not taking aspirin as the prevention therapy.

Conclusions

This work demonstrates that rules induced by the subgroup discovery methodology can be an appropriate starting point for data analysis leading to insightful descriptions of the available data. The extensive presentation of the analysis process intends to illustrate the intellectual effort necessary to convert the induced rules into reasonable medical knowledge. Special attention was devoted to the selection of appropriate visualization, enabling effective and convincing presentation of obtained results. The paper demonstrates that this type of data analysis, besides expert knowledge, requires also a lot of human imagination. Further work is expected in developing the methodology which could be used for semi-automated insightful data analysis.

References

[Fayyad *et al.*, 1996] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

[Gamberger *et al.*, 2003] Dragan Gamberger, Nada Lavrač, and Goran Krstačić. Active subgroup mining: A case study in a coronary heart disease risk group detection. *Artificial Intelligence in Medicine*, 28:27–57, 2003.

[Victor and Ropper, 2001] Maurice Victor and Allan H. Ropper. *Principles of Neurology*, chapter Cerebrovascular Disease, 821–924. McGraw-Hill New York, 2001.

Probabilistic Medical Record Linkage incorporating close agreement

M Tromp, N Méray, ACJ Ravelli, JB Reitsma, GJ Bonsel
Department of Medical Informatics; Department of Public Health Methods
University of Amsterdam
P.O. Box 22700, 1100 DE Amsterdam, The Netherlands
m.tromp@amc.uva.nl

Abstract

Probabilistic close medical record linkage techniques have been used to link the Dutch perinatal registries. Close agreement weights further enhance the probabilistic procedure. External validation showed that the developed procedure is highly reliable.

1 Introduction

In the Netherlands, four different caregivers (midwifes, general practitioners, obstetricians and paediatricians) are involved in perinatal care. Each caregiver collects data into their own national registry. These registries are anonymous and because of privacy laws in the Netherlands, no unique personal identifier exists.

A combined dataset of the separate registries is needed to produce valid reports on outcome measures of perinatal healthcare and to enable further data analysis. Medical record linkage is a technique to identify records belonging to the same individual in the absence of a unique identifier [Newcombe, 1988]. Partly identifying variables present in both datasets are combined to create a powerful discriminating linking key.

There are two different approaches to medical record linkage; deterministic medical record linkage and probabilistic medical record linkage. Deterministic strategies only look at (dis)agreement on the linking variables. For full deterministic linkage all linking variables have to agree, while N-1 deterministic linkage allows one of the linking variables to disagree [Newcombe, 1988]. Probabilistic medical record linkage strategies use the information value of the different linking variables by assigning a weight for agreement and disagreement for every variable. This weight is calculated using two probabilities: the probability that a variable agrees among matches (m_i) and the probability that a variable agrees among non-matches (u_i). The m_i value reflects the reliability of the variable, while the u_i value reflects the discriminating power of the variable [Newcombe, 1988; Bell et al.,1994].

This paper in short describes the applied probabilistic medical record linkage algorithm incorporating close agreement to link the registry of midwifes ("MR") and the registry of obstetricians ("OR").

2 Methods

The datasets of the perinatal registries were linked for the years 2001 - 2003. First, both datasets for one year were internally linked to detect administrative doubles using a deterministic N-1 approach. In the next step, the MR and OR were linked using a probabilistic medical record linkage algorithm. The two datasets were separated for singletons and twins. Records of multiple births require a different, stricter, approach because these records have a lot of variables in common. U_i probabilities were calculated from the marginal distribution in the two files as true non-matches make up the largest part of the total number of pairs. Because the matching status is unknown, m_i values were estimated using the Expectation Maximization (EM) algorithm with the observed patterns of agreement and disagreements of the singleton files without missing values [Felligi and Sunter, 1969; Reitsma, 1999]. Missing values influence the calculation of weights in an undesirable way because a missing value on a variable in both records is seen as agreement by the EM algorithm. Besides full agreement, close agreement was defined for certain variables. Even in case of agreement the value of a variable in two records can differ because of different calculation methods or rounding off of figures. Identification of these variables, and the definition of the close range was established with help of caregivers combined with information from the data (Table 1).

Because of large file sizes (about 160.000 records for MR, 125.000 for OR) blocking was applied in two steps. Records were only compared if they agreed on the blocking variable in the first step and in the second step.

Table 2 list the variables there were compared in the linking procedure. For (DOB), and close ranges were defined.

Table 1 Effect of choice of close range on linking weight for the variable

Close range	Weight agree	Weight disagree
Birth weight (full)	7,99	
Birth weight (± 5g)	1,44	-4,16
Birth weight (± 10g)	**0,91**	**-4,44**
Birth weight (± 25g)	0,17	-4,70
Birth weight (± 50g)	-0,45	-5,17
Birth weight (± 100g)	-1,12	-5,76

Table 2 Linking variables with m_i and u_i value and linking weight for 2002 data

	m_i value	u_i value	weight agree	weight close agree	weight dis-agree
DOB mother	Blocking				
ZIP code mother	0,9573	0,0006	10,74		-4.55
DOB child	0,9780	0,0028	8,47	1,50	-7,28
(± 1 day)	0,0156	0,0055			
Exp. DOB child	0,8877	0,0027	8,36	1,35	-5,79
(± 7 day)	0,0949	0,0371			
Birth weight	0,9356	0,0037	7,98	0,91	-4,44
(± 10 gram)	0,0191	0,0102			
Place of birth	0,8818	0,0064	7,11		-3,07
Minute of birth	0,9173	0,0180	5,67		-3,57
Hour of birth	0,9701	0,9701	4,50		-5,00
Gravidity	0,9457	0,3016	1,65		-3,69
Gender	0,9918	0,5006	0,99		-5,93

Linkage weights of the different variables were calculated using the m_i and u_i values (Table 2):
Full agreement weight of the i^{th} variable: $\log_2 (m_{if}/u_{if})$
Close agreement weight of the i^{th} variable: $\log_2 (m_{ic}/u_{ic})$
Disagreement weight of the i^{th} variable:
$\log_2 (1-(m_{if}+m_{ic})/ 1-(u_{if}+u_{ic}))$

For every record pair a total linkage weight was calculated by adding up the individual variable weights. Linking weight was set to zero if a variable was missing in one or both records compared. All pairs were sorted by this total weight and a threshold value was determined separating links from non-links based on the estimated match rate by the EM algorithm and by reviewing pairs around this estimated threshold value.

A double blind external validation with the medical record as gold standard was carried out for the MR^OR linkage, focussing mainly on the uncertain area of the linkage (around the threshold value).

3 Results

Duplicates were removed from the separate registry files; 0.5% in the MR and 0.05% in the OR. Linking of the MR with the OR showed that 41% of all pregnancies were present in both files. Figure 1 shows all pairs sorted by total linkage weight together with the threshold value (15.4) above which value pairs are considered a link. External validation revealed no errors outside the uncertain region (weight of ±5 around threshold value) and a false margin of 13% in the uncertain region. Only 0.35% of the linked pairs are in the uncertain region and only 0.055 % of the non-linked pairs, which means that our total linkage procedure has a margin of error of less than 1 %.

4 Discussion

A perinatal healthcare data file now exists with combined data from the different involved disciplines, which can be used to produce valid tables on outcome measures and offers new possibilities for further data analysis. As external validation showed, the developed probabilistic close medical record linkage procedure is highly reliable.

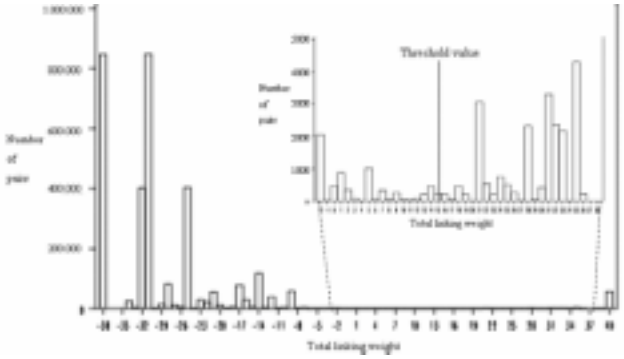


Figure 1 MR^OR singleton pairs sorted by total linking weight

The weights calculated for the linking variables proved to be very stable for variations in the number of linking variables and close ranges. Blocking had to be applied because of large file sizes, but we believe two step blocking minimizes the number of false negative links. Our decision the separate the linkage of singleton and multiple births worked well, although multiple births remain difficult to link. Choices made on handling missing values and defining close ranges should be further founded by simulation studies. SAS was used for all linking procedures and proved to be a flexible and powerful tool.

5 Further research

Additional simulation studies will further ground choices made so far in particular the dependency of their optimality on dataset characteristics (number of records, the ratio of possible links to file sizes and error rates of variables). Simulation studies conducted to refine the linking strategy will focus on the range of close agreement, the handling of missing values and dependencies between linking variables by using a simulated dataset with known match status. Yet a valid probabilistic linking procedure is now available and can be used for similar problems.

Acknowledgments

We gratefully acknowledge the support and funding of the SPRN (Foundation of the Dutch Perinatal Registry) and the investment of numerous caregivers.

References

[Bell ., 1994] Bell RM, Keesey J, Richards T. The urge to merge: linking vital statistics records and Medicaid claims. 32: 1004-1018, 1994.

[Felligi and Sunter, 1969] Felligi IP, Sunter AB. A theory for record linkage.
64: 1183-1210, 1969.

[Newcombe, 1988] Newcombe HB.

Oxford: Oxford University Press, 1988.

[Reitsma, 1999] Reitsma JB.
. PhD thesis Academic Medical Center, University of Amsterdam, 1999. ISBN 90-9013206-6.

Diagnosis of Dysmorphic Syndromes Using Prototypes and Adaptation Rules

Tina Waligora, Rainer Schmidt
Institute for Medical Informatics and Biometry
University of Rostock
Rembrandtstr. 16/17, D-18055 Rostock, Germany
[tina.waligora / rainer.schmidt] @medizin.uni-rostock.de

Abstract

Since diagnosis of dysmorphic syndromes is a domain with incomplete knowledge and where even experts have seen only few syndromes themselves during their lifetime, documentation of cases and the use of case-oriented techniques are popular. In dysmorphic systems, diagnosis usually is performed as a classification task, where a prototypicality measure is applied to determine the most probable syndrome. Our system additionally applies adaptation rules. These rules do not only consider single symptoms but combinations of them, which indicate high or low probabilities of specific syndromes.

1 Introduction

When a child is born with dysmorphic features or with multiple congenital malformations or if mental retardation is observed at a later stage, finding the correct diagnosis is extremely important. Knowledge of the nature and the etiology of the disease enables the paediatrician to predict the patient's future course. So, an initial goal for medical specialists is to diagnose a patient to a recognised syndrome. Genetic counselling and a course of treatments may then be established.

A dysmorphic syndrome describes a morphological disorder. It is characterised by a combination of various symptoms, which form a pattern of morphologic defects. The main problems of diagnosing dysmorphic syndromes are as follows [Gierl ., 1994]:

- existence of more than 200 syndromes,
- many cases remain undiagnosed with respect to known syndromes,
- usually many symptoms are used to describe a case (between 40 and 130),
- every dysmorphic syndrome is characterised by nearly as many symptoms.

Furthermore, knowledge about dysmorphic disorders is continuously modified, new cases are observed that cannot be diagnosed, and sometimes even new syndromes are discovered.

We have developed a diagnostic system that uses a large case base. Starting point to build-up the case base was a large case collection of the paediatric genetics of the University of Munich, which consists of nearly 2,000 cases and 229 prototypes. A prototype (prototypical case) represents a dysmorphic syndrome by its typical symptoms. Many dysmorphic syndromes have been defined in literature. Additionally, nearly one third of our case base was determined by semiautomatic knowledge acquisition, where an expert selects cases that should belong to the same syndrome and subsequently a prototype, characterised by the most frequent symptoms of it's cases, is generated.

In our system the user can choose between two measures of dissimilarity between concepts, namely one measure proposed by Tversky [Tversky, 1977], the other one by Rosch and Mervis [Rosch ., 1975]. However, the novelty of our approach is that we do not only perform classification but subsequently apply adaptation rules. These rules do not only consider single symptoms but specific combinations of them, which indicate high or low probabilities of specific syndromes.

2 Diagnosis of Dysmorphic Syndromes

Our system performs four steps. At first the user has to select symptoms that describe a new patient. This selection is strenuous and time consuming, because more than 800 symptoms are considered. However, diagnosis of dysmorphic syndromes is not a task requiring great speed, but it usually requires thorough reasoning and is followed by a long-term therapy. Since our system is still in the evaluation phase, the user can select a prototypicality measure. In routine use, this step shall be dropped and instead the measure with better evaluation results shall be used automatically. There are two choices.

As humans look upon cases as more typical for a query case with increasing numbers of common features [Rosch ., 1975], distances between prototypes and cases usually mainly consider the shared features. The first measure was developed by Tversky [Tversky, 1977]. It is a measure of dissimilarity of concepts. From the number of features shared by the query case and the prototype two numbers are subtracted. Firstly, the number of symptoms that are observed for the patient but are not used to characterise the prototype (X-Y), and secondly the

number of symptoms used for the prototype but are not observed for the patient (Y-X) is subtracted.

$$D_{Tversky}(X,Y) = \frac{f(X + Y) - f(X - Y) - f(Y - X)}{f(Y)}$$

The second prototypicality measure was proposed by Rosch and Mervis [Rosch ., 1975]. It differs from Tversky's measure only in one point: the factor X-Y is not considered:

$$D_{Rosch,\ Mervis}(X,Y) = \frac{f(X + Y) - f(Y - X)}{f(Y)}$$

In the third step to diagnosis dysmorphic syndromes, the chosen measure is sequentially applied on all prototypes (syndromes). Since the syndrome with maximal similarity is not always the right diagnosis, the 20 syndromes with highest similarity are presented ranked according similarity.

2.1 Application of Adaptation Rules

In the fourth and final step, the user can optionally choose to apply adaptation rules on the syndromes. These rules state that specific combinations of symptoms favour or disfavour specific dysmorphic syndromes. For example, this is an adaptation rule favouring Lenz-Syndrome:

```
IF medial diffuse hypoplast brows
AND IF prominent Corpus-Anthelicis
THEN the Lenz-Syndrome is probable
```

Unfortunately, the acquisition of these adaptation rules is very difficult, because they cannot be found in textbooks but have to be defined by experts of paediatric genetics. So far, we have got only 10 of them and so far it is not possible that a syndrome can be favoured by one adaptation rule and disfavoured by another one at the same time. When we, hopefully, acquire more rules such a situation should in principle be possible but would indicate some sort of inconsistency of the rule set.

The question is how shall adaptation rules alter the results. Our first idea was that the similarity values should be changed. A syndrome that is favoured by an adaptation rule might get a higher similarity. But we had no idea how much an adaptation rule shall increase a similarity. Of course no medical expert can help here and a general value for favoured or disfavoured syndromes by adaptation rules would be arbitrary. So, instead the result after applying adaptation rules is a menu that contains up to three lists. On top the favoured syndromes are depicted, then those neither favoured nor disfavoured, and at the bottom the disfavoured ones. Additionally, the user can get information about the specific rules that have been applied on a particular syndrome.

3 Results

Cases are difficult to diagnose when patients suffer from a very rare dymorphic syndrome for which neither detailed information can be found in literature nor many cases are stored in our case base. This makes evaluation difficult. If test cases are randomly chosen, frequently observed syndromes will be frequently selected and the results will probably be fine, because these syndromes are well-known. However, the main idea of our system is to support diagnosis of rare syndromes. So, we have chosen our test cases randomly but under the condition that every syndrome can be chosen only once. For 100 cases we have compared the results obtained by both prototypicality measures, before and after applying adaptation rules (table 1).

Table 1. Comparison of prototypicality measures

			With Adaptation	With Adaptation
Right Syndrome	Rosch, Mervis	Tversky	Rosch, Mervis	Tversky
on Top	29	40	32	42
among top 3	57	57	59	59
among top 10	76	69	77	71

Obviously, the measure of Tversky provides just very slightly better results, especially when the right syndrome should be on top of the list of probable syndromes. Since the acquisition of adaptation rules is very difficult and time consuming, the number of acquired rules is rather limited, namely 10 rules. Furthermore, again holds: the better a syndrome is known, the easier adaptation rules can be generated. So, the improvement mainly depends on the question how many syndromes involved by adaptation rules are among the test set. In our experiment this was the case only with five syndromes. Since some of them had already been diagnosed correctly without adaptation, the improvement by adaptation rules is very small.

References

[Gierl ., 1994] Lothar Gierl and Sabine Stengel-Rutkowski,. Integrating Consultation and Semi-automatic Knowledge Acquisition in a Prototype-based Architecture: Experiences with Dysmorphic Syndromes. , 6:29-49, 1994

[Rosch ., 1975] Elenor Rosch and C.B. Mervis. Family Resemblance: Studies in the Internal Structures of Categories. 7:573-605, 1975

[Tversky, 1977] A Tversky. Features of Similarity. 84:327-352, 1977

FreeViz - An Intelligent Visualization Approach for Class-Labeled Multidimensional Data Sets

Janez Demšar[1], **Gregor Leban**[1], **Blaž Zupan**[1,2]
Faculty of Computer and Information Science, University of Ljubljana, Slovenia
Department Molecular and Human Genetics, Baylor College of Medicine, Houston, TX
janez.demsar@fri.uni-lj.si

Abstract

Within biomedical data analysis, visualization can greatly improve data understanding and support various data mining tasks. The paper presents FreeViz, a visualization technique for analysis of class-labelled, multi-dimensional data. FreeViz visualizations can present data on many features in the same graph, but through optimization procedure choose a projection that best separates instances of different class. The paper gives mathematical foundations of FreeViz, and presents its utility on various biomedical data sets, including those with thousands of features from cancer gene expression studies.

1 Introduction

Medical data analysis may largely benefit from visualization. The *right* visualization may outline which factors govern the data and uncover their interactions. In the paper, we will be concerned with predictive data mining tasks, where each data instance (case) is described with a set of features (predictive variables) and labelled with a class (*e.g.* outcome, diagnosis). Despite many visualization techniques available, there are not too many of those that can visualize several features in the same graph, and, for instance, include scatterplot (two or three features, the later if plotted in 3D), parallel coordinates and RadViz (both for presentation of data using many features) [Keim, 2002].

When considering data sets with many features, which are typical in the domain of biomedicine, the principal problem to solve is which features to visualize and which projection to use, that is, how to order the selected features in the graph. With increasing number of features, any manual search for good projections becomes unfeasible. In principle, we would then prefer to use some automatic search for *good* projections, that would optimize some criteria for quality of interestingness. For a single-class (unsupervised) data, a well-known technique of projection pursuit is available for the task [Huber, 1985]. But interestingly, for class-labelled data, such *intelligent data analysis* approaches are at best rare, while the task is somehow better defined: interesting visualization is the one that well separates data instances of different class. We are aware of two approaches in this category, McCarthy *et al.*'s RadViz projections that place correlated features in RadViz close to each other and thus try to improve on class separation, and Leban *et al.*'s Vizrank [Leban *et al.*, 2005] that directly optimizes class separation and uses the heuristic search through projection space [McCarthy *et al.*, 2004].

In the paper, we propose an iterative algorithm that optimizes class separation in visualization of class-labelled data sets. The visualization it uses is based on RadViz [Brunsdon *et al.*, 1998], and is called FreeViz since it relaxes the constraints of placement of feature anchors; in RadViz, these are placed on the boundary of a circle. FreeViz is fast, can propose good visualizations even in the case of highly-dimensional data sets such as those from cancer genomics within seconds, and can be further used for feature subset selection and feature interaction discovery.

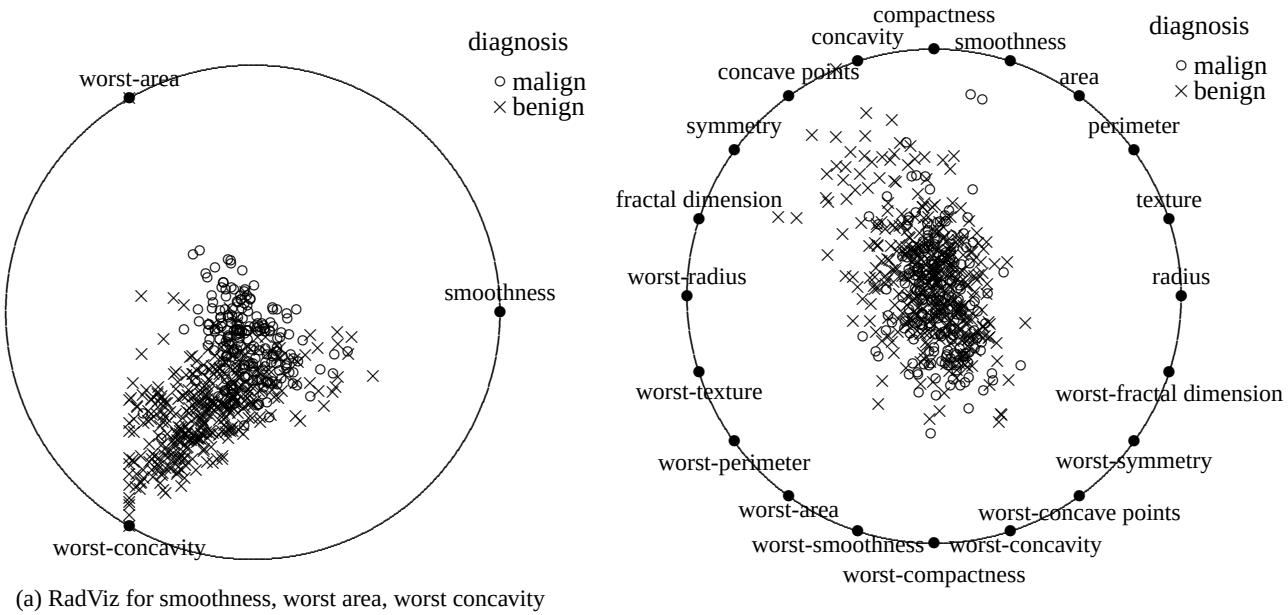
We first give the background on RadViz and its intelligent visualization counterpart VizRank. We formally describe FreeViz, present a mathematical derivation of its fitness (quality) function describe the corresponding implementation of the optimization algorithm. We then give several cases that show a utility of FreeViz in biomedical data analysis, also including examples that use large cancer gene expression data set. We conclude with discussion and ideas for further work.

Before we go on, notice that any modern visualization can largely benefit from colored display. Figures in the paper are printed in black and white, which at places significantly decreases their clarity. The reader is invited to visit a supplemental web page (www.ailab.si/supp/freeviz-idamap) for better images.

2 Background

RadViz [Brunsdon *et al.*, 1998] is a visualization that is suitable for data described with a set of continuous features scaled to the interval [0, 1]; discrete features can be visualized through first transforming them to continuous. The features are represented by anchors placed evenly on the unit circle. The data instances are plotted inside the circle; the position of each is determined by its features and the positions of the corresponding anchors. Informally, each anchor pulls the instance towards itself with a strength proportional to the value of the corresponding feature, so the position of an example depends upon the relative values of features (*e.g.* if all features have equal values, the instance is placed in the center).

Figure 1(a) shows a RadViz for three features (smoothness, worst area, worst concavity) of the Wisconsin Di-



(a) RadViz for smoothness, worst area, worst concavity

(b) RadViz using all twenty features

Figure 1: Two RadViz graphs for Wisconsin Breast Cancer Data

agnostic Breast Cancer data (WDBC) from the UCI ML repository [Blake and Merz, 1998]. The interpretation of such a graph is rather obvious: tissues with a large "worst area" tend to be malign and tissues with a large "worst concavity" are benign, while the role of smoothness is not clear. The problem arises when (or, better, because) the data instances are described by more than a few features. The actual WDBC data has 20 features and the corresponding RadViz looks as shown in Figure 1(b); the order of features is the same as in the data.

RadViz can be truly useful only when used with some methods for optimizing it. The features for Figure 1(a) were chosen using the algorithm VizRank developed by [Leban *et al.*, 2005], which exhaustively searches through all combinations of features within the specified parameters (usually we set the upper number of features to four or five) and evaluates the projection using a k nearest neighbors classifier. A projection is good if each instance is surrounded mostly by instances of its own class. To avoid overfitting, cross-validation is used instead of computing the quality of the graph directly. Since the number of combinations rises exponentially with the number of features, VizRank checks the projections ordered by the quality of the features they use, where the features are evaluated with a common measure such as ReliefF or information gain. Despite the huge number of combinations which can on microarray data easily reach 10^{20}, RadViz can most often find good projections within minutes of runtime.

By placing the anchors evenly around the circle and letting each pull in its own direction, Radviz assumes that the features are not correlated. Placing the anchors corresponding to strongly correlated features closer together would be potentially beneficial in conquering the noise and would, at the same time, offer a cleaner and more informative visualization. This idea is successfully exploited by McCarthy *et al.* [2004], but where a limitation with respect to RadViz is that visualization includes all available features.

The other limitation of RadViz is that it in principle assumes that all features are equally important. Since this is usually not the case, the quality of the projection is decreased since the pull of a less important feature(s) is as strong as those of the important ones.

The visualization we propose, FreeViz, overcomes both limitations by allowing the anchors to be placed anywhere in the circle. The correlated features can thus be placed together and the less important features can be put nearer to the circle's center to lower their impact. Even more than with RadViz, the usefulness of FreeViz depends upon the methods for optimizing it.

3 Formal Description and Optimization

Let $A^i = [A^i_x, A^i_y]$ be the i-th anchor, and $\mathbf{A}$ be a matrix of anchors. Each instance is described by a vector of feature values, $\mathbf{e} = [e^1, e^2, \ldots, e^n]$. The position of instance e in the circle is computed as $e_x = \sum_i e^i A^i_x$, $e_y = \sum_i e^i A^i_y$ or, in matrix notation, $\mathbf{e}' = \mathbf{eA}$. $\mathbf{A}$ thus represents a linear transformation that projects from the original feature space to a two-dimensional FreeViz.

Instead of using k-nearest neighbours, as VizRank does, we will optimize the projection by minimizing its potential energy, vaguely following the real-world physics of gravitational/electric fields [Halliday and Resnick, 1978]. Let $\mathbf{F}_{f \to e}$ be the force acting on instance e due to instance f. The force will depend on the distance between the two instances, their charges (weights of instances) and the type of their charges (instances' class – instances of the same class will attract and instances of different classes

will repel each other). When a particle e is moved by $\mathbf{de}'$ the work and the change of the potential energy equals $dE = A = -\mathbf{F}_{f \to e}\mathbf{de}'$.

In a system of multiple particles, the force acting on a particle equals the sum of forces exerted by all other particles,

$$\mathbf{F}_e = \sum_{f \neq e} \mathbf{F}_{f \to e}$$

and the change of potential energy when moving e is $dE_e = \mathbf{F}_e\mathbf{de}'$. When multiple particles are moved at once (as they will be in our case), the change of energy equals the sum of changes,

$$dE = -\sum_e \mathbf{F}_e \, \mathbf{de}'$$

We shall use the gradient method to optimize the system, *i.e.* to minimize its potential energy by moving the anchors. For this, we need to compute the gradient of the energy as a function of the anchors' position. Consider that $\mathbf{e}' = \mathbf{eA}$ and so $\mathbf{e}' = \mathbf{e}\,\mathbf{dA}$. When anchors are moved, the change in energy equals

$$dE = -\sum_e \mathbf{F}_e \, (\mathbf{e}\,\mathbf{dA})$$

For moving the x-coordinate of the i-th anchor, the related change in energy is $dE = \sum_e \mathbf{F}_{e,x} e^i dA^i_x$, where $F_{e,x}$ is the x-component of the force F_e, therefore

$$\frac{dE}{dA^i_x} = -\sum_e \mathbf{F}_{e,x} e^i$$

The computation of the y-coordinate is analogous. The formula is consistent with our intuition and with the nature which (at least on grand scale) minimizes the potential energy by accelerating the objects in the direction opposite to the energy gradient (that is, in the direction of the force). Instances are attracted or repelled from each other, but since they are held in place by the anchors, the forces between them are transmitted to the anchors. The force acting on each particle is distributed between the anchors proportionally with the values of corresponding features, e^i.

The formula is independent of the definition of the force. Its sign should depend upon whether the two instances are from the same class or not, so the force is attractive in the former and repulsive in the latter case. If instances are weighted, the force should rise linearly with the instance's weight. As for the distance, in our three dimensional space the usual large scale forces decrease by the inverse-square law, $F \sim 1/r^2$. In the two-dimensional world of FreeViz, the density of the field lines decreases linearly with the distance, so the force should be proportional to $1/r$. On the other hand, we can borrow the idea of Gaussian kernels from the statistics and let the force be proportional to e^{-r^2}. After some testing we found that the inverse-square law works best, while with linear or Gaussian kernels the force decrease with distance seems too slow.

A more important consideration regarding the force is whether it needs to decrease or increase with the distance. When separating instances of different classes, we are most concerned with those that are close together, while we do not need to push the groups that are already well separated even further apart. The repulsive force must therefore fall with the distance. On the other hand, the attractive force would try to squeeze the well-defined groups of instances from the same class into a point, and this unneeded effect would rise as the instances come closer together. For a contrast, if an instance is far from other instances of its own class and surrounded by instances of another, the former will not attract it, due to a large distance, while the latter will push it around and, in the best case, throw it out in a random direction. The attractive force should therefore increase with the distance.

```
Input:  number of instances N
        number of features A
        instance projections P
        a table of instances E
        classes of instances C
Output: a vector of gradients G

initialize F to 0
for e := 1 to N
    for f := e+1 to N
        dx := P[e].x - P[f].x
        dy := P[e].y - P[f].y
        r := sqrt(sqr(dx) + sqr(dy))

        if C(e) = C(f)
            then F_ef := -r^2
            else F_ef := 1/r^2

        Fefx := F_ef * dx/r
        F[e].x += Fefx
        F[f].x -= Fefx

        Fefy := F_ef * dy/r
        F[e].y += Fefy
        F[f].y -= Fefy

initialize G to 0
for e := 1 to N
    for i := 1 to A
        G[e].x += F[e].x * E[e][i]
        G[e].y += F[e].y * E[e][i]
```

Figure 2: Computation of gradients for FreeViz optimization

In a sense, the repulsive forces act like the electromagnetic or gravitational forces which decrease by the distance, while the attractive forces resemble the strong force that binds quarks and which increases by the distance, like a rubber band.

In the algorithm for computation of gradients (Figure 2) we make use of the action-reaction symmetry: the force between each pair of instances is computed only once and added to the sum of forces for both instances, but with different directions ($F_{f \to e} = -F_{e \to f}$). The force (`F_ef`) is separated into its x and y components (`Fefx` and `Fefy`) by multiplying it by projections to x and y axis, dx/r and dy/r, respectively.

The algorithm is rather simple and relatively fast: its

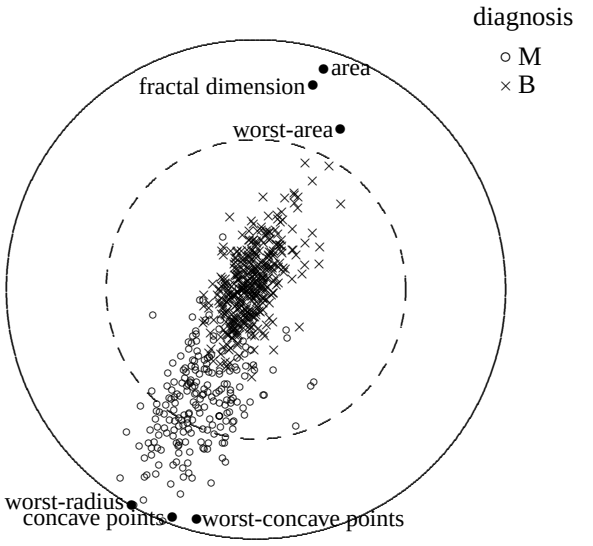


Figure 3: FreeViz for Wisconsin Breast Cancer data

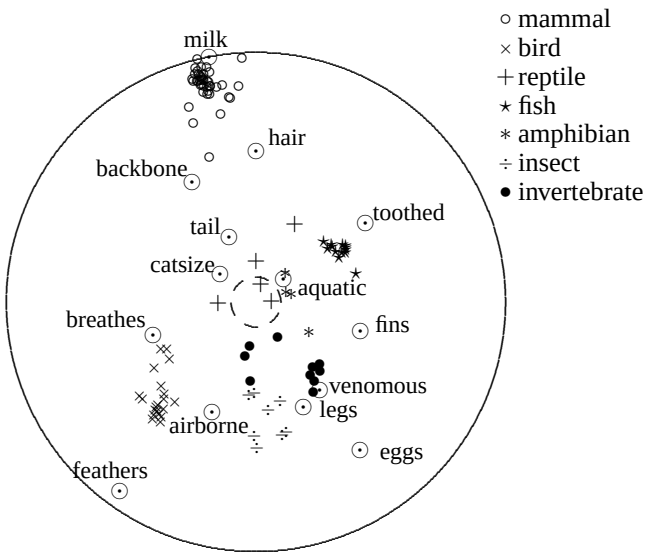


Figure 4: FreeViz for zoology database

time complexity is $O(N^2 + NA)$, where N is the number of instances and A the number of features; the first term comes from computation of forces between particles and the second from the loop that distributes the forces acting on each instance between the anchors. Although the operations performed by the algorithms are rather elementary, the squared number of instances suggests that the algorithm may be less useful when the number of instances is large.

The computed gradients can be used in optimization with the ordinary gradient method; at each step, the gradient vector is subtracted from the vector of anchors, the anchors are centered and renormalized (the farthest anchor should lie on the unit circle), and the projections are recomputed. The procedure is repeated until there is no considerable decrease (*e.g. 1 %*) of the potential energy for few consecutive steps.

Gradient method of optimization could be replaced with more advanced methods, but we found it fit for our purpose: it is fast and does not seem to stop in local minima.

Figure 3 shows a FreeViz for WDBC optimized by the proposed algorithm. For a clearer picture, we did not plot the features whose anchors are less than $0.5r$ from the center (marked with a dashed circle). The "area", "fractal-dimension" and "worst-area" listed in order of importance, seem to be correlated evidences for benignity, while the other three features speak for malignity of the tumor.

An important note about the algorithm is that it should not be used when the number of features exceeds the number of instances. Formally, if **E** is a matrix of instances and its rank equals the number of instances, the system $\mathbf{EA} = \mathbf{P}$ can be solved for any matrix of instances' positions **P**. In other words, if we have more features than instances, there exists a matrix of anchor positions for any prescribed positions of instances. The described algorithm is in this case able to overfit the data, resulting in meaningless projections.

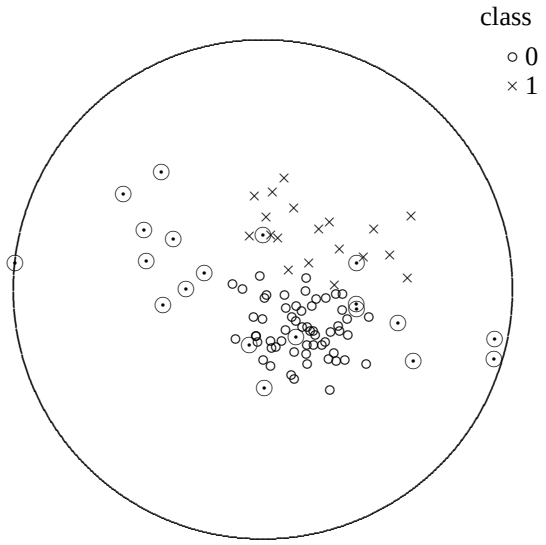
4 Case Studies and Discussion

We start with an example on a zoology data set which contains 101 animals described by their properties (lay eggs, breath, have hair...) and classified into seven groups (mammals, birds, reptiles...). As Figure 4 shows, the animals can be separated using the FreeViz projection and the corresponding positions of features make sense. For instance, mammals (○) have hair, backbone, and as the most important feature, milk. Being airborn is typical of birds (×) and insects (÷); the former are distinguished by feathers, and the latter have more legs (this feature can have values 0, 2 and 4). Amphibians are put between fish and reptiles.

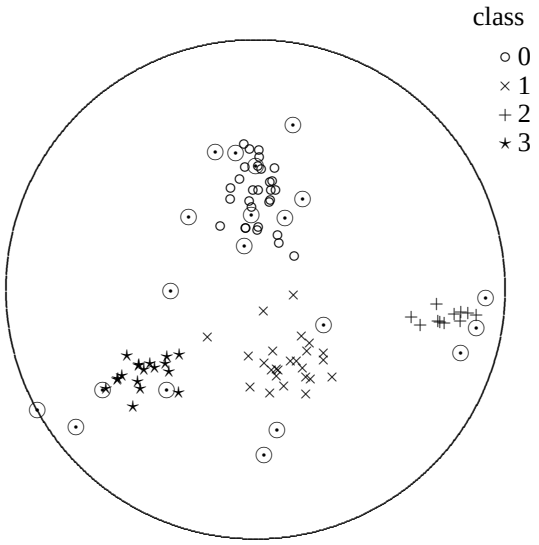
To test the visualization on a more complex data, we have tried FreeViz on several microarray cancer data sets. The resulting visualizations are shown in Figure 5. The feature names are intentionally uninformative (paper focuses on the study of class-separability, and while biomedical interpretation would be useful, it is beyond the scope of our reported study) and we have hidden them for the sake of clarity. The legend is omitted for the same reasons. To limit the number of features well below the number of instances, we have used ReliefF [Kononenko *et al.*, 1997] to select 20 most important genes for each data set (except for Lung cancer which has somewhat larger number of instances, where we have chosen a subset of 40 features).

Figure 5(a) shows the visualization of the data set that studies the outcome for the diffuse large B-cell lymphoma (DLBCL) [Shipp *et al.*, 2002], where the selected 20 features are well able to separate between the two classes. In another example, the data on four types of tumors in childhood (SRBCT) [Wang *et al.*, 2003], see Figure 5(b), the optimization yielded an even clearer separation.

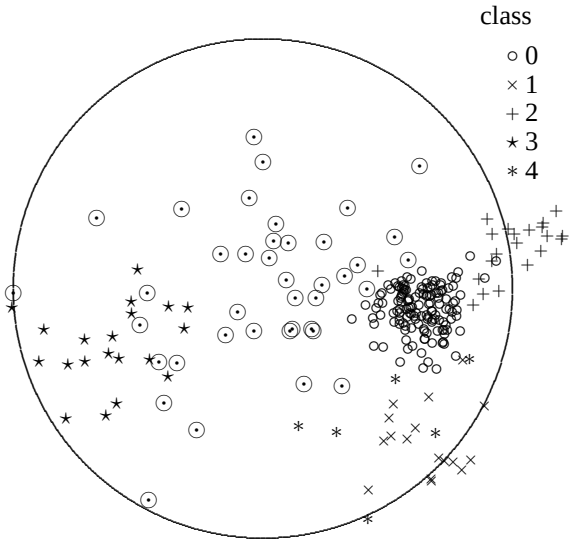
The largest data set we tackled is that on a lung cancer [Bhattacharjee *et al.*, 2001] with 203 instances, 12600 genes and five classes (Figure 5(c)). The separation is generally good, except for the class ∗, which is apparently too small, so the total force that its instances exert on anchors



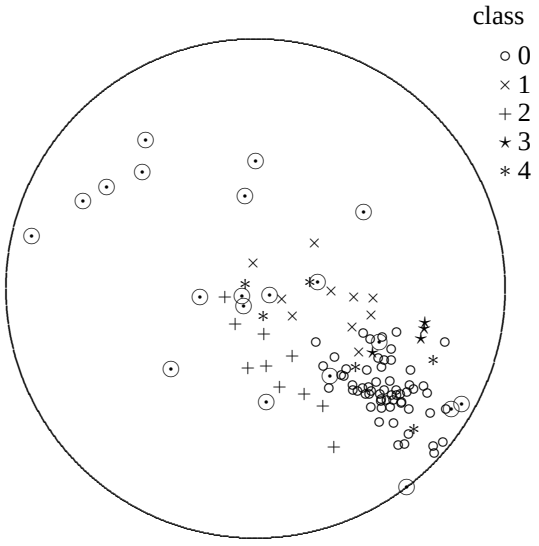
(a) DLBCL: 77 instances, 20 (out of 5469) genes



(b) SRBCT: 82 instances, 20 (out of 2308) genes



(c) Lung cancer: 203 instances, 40 (out of 12600) genes



(d) Brain tumor: 90 instances, 20 (out of 5920) genes

Figure 5: FreeViz on cancer microarray data

is incomparable to the forces by instances of the larger classes. In such cases, the algorithm could be augmented by adjusting the strength of forces according to the size of classes.

For the brain tumor data with 5920 genes and 90 instances (Figure 5(d)), separation was somewhat worse. Again, the instances belonging to the smaller classes ($\star$ and $*$) are lost between those of the large classes.

In all cases, running ReliefF took up to half a minute, while FreeViz optimization took a few seconds on a mediocre PC (Pentium IV, 1800 MHz).

5 Conclusion and Future Work

The paper presents a new method for intelligent visualization of class-labelled, multi-dimensional data sets. We have presented its utility on a number of biomedical data sets. Results of these preliminary studies are very encouraging: FreeViz is very fast and in all presented cases found visualizations of high quality with clear class separation.

There are many ideas that we have on how FreeViz can be exploited further. Some most important include:

- Visualization of probabilities. By computing the potential fields for a grid of points in the circle, it is possible to color the inside of the circle so that the color corresponds to the most probable class for an instance projected to that point and the color's saturation to the probability. We have implemented this functionality, but presenting it in the proceedings would require a color print, so we show it only on supplemental web pages (www.ailab.si/supp/freeviz-idamap).
- Classification. FreeViz visualization can be employed in classification of new cases. The simplest method, for instance, to produce a classifier from those pictures is to project the instance which is to be classified into the FreeViz space and observe its k nearest neighbors. Our experiments (not published here) with this are very encouraging and show that obtained classification accuracy, AUC and Brier scores are in the same range as those from logistic regression, naive Bayesian classifier and SVM.
- Misclassification costs. With the current implementation of the algorithm, the strength of repulsive forces depends upon the distance between the instances but not on their classes. By modifying it so that different combinations of classes would repel with different strengths, misclassification costs could easly be incorporated within analysis.

FreeViz is a available as a part of RadViz visualization widget in open-source data mining suite Orange (www.ailab.si/orange, [Demšar and Zupan, 2004; Zupan *et al.*, 2004]. As such it also offers other functionality, such as manual placement of anchors, selection of subsets of examples and similar, which is not described in this paper. See also supplemental web page (www.ailab.si/supp/freeviz-idamap) for additional material and figures from the paper in color.

References

[Bhattacharjee *et al.*, 2001] A. Bhattacharjee, W. G. Richards, and J. Staunton *et al*. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma sub-classes. *Proc. Natl. Acad. Sci. USA*, 98 (24), 2001.

[Blake and Merz, 1998] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[Brunsdon *et al.*, 1998] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton. An investigation of methods for visualising highly multivariate datasets. In *Case Studies of Visualization in the Social Sciences*, pages 55–80. Joint Information Systems Committee / ESRC, 1998.

[Demšar and Zupan, 2004] J. Demšar and B. Zupan. *Orange: From Experimental Machine Learning to Interactive Data Mining, A White Paper*. Faculty of Computer and Information Science, Ljubljana, Slovenia, 2004.

[Halliday and Resnick, 1978] D. Halliday and R. Resnick. *Physics*. John Wiley and Sons, New York, 3rd edition, 1978.

[Huber, 1985] P. J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–474, 1985.

[Keim, 2002] D. A. Keim. Information visualization and visual data mining. *Transactions on Visualization and Computer Graphics*, 6(1):59–78, 2002.

[Kononenko *et al.*, 1997] I. Kononenko, E. Šimec, and M. Robnik Šikonja. Overcoming the myopia of inductive learning algorithms with ReliefF. *Applied Intelligence Journal*, 7(1):39–56, 1997.

[Leban *et al.*, 2005] G. Leban, I. Bratko, U. Petrovic, T. Curk, and B. Zupan. Vizrank: finding informative data projections in functional genomics by machine learning. *Bioinformatics*, 21(3):413–414, 2005.

[McCarthy *et al.*, 2004] J. F. McCarthy, K. A. Marx, P. E. Hoffman, A. G. Gee, P. O'Neil, M L. Ujwal, and J. Hotchkiss. Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. *Annals of New York Academy of Sciences*, 1020:239–262, 2004.

[Shipp *et al.*, 2002] M. A. Shipp, K. N. Ross, and P. Tamayo *et al*. Diffuse large b-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nature Medicine*, 8:68–74, 2002.

[Wang *et al.*, 2003] Zuyi Wang, Yue Wang, and Jianping Lu *et al*. Discriminatory mining of gene expression microarray data. *J. VLSI Signal Process. Syst.*, 35(3):255–272, 2003.

[Zupan *et al.*, 2004] B. Zupan, G. Leban, and J. Demšar. *Orange: Widgets and visual programming, A White Paper*. Faculty of Computer and Information Science, Ljubljana, Slovenia, 2004.

Gravi++: Interactive Information Visualization of Highly Structured Temporal Data

Klaus Hinum[1], Silvia Miksch[1], Wolfgang Aigner[1], Susanne Ohmann[2], Christian Popow[2], Margit Pohl[3], and Markus Rester[3]

[1]Institute of Software Technology & Interactive Systems, Vienna University of Technology

[2]Department of Child and Adolescent Neuropsychiatry, Medical University of Vienna

[3]Design and Assessment of Technologies Institute, Vienna University of Technology

Vienna, Austria, http://ieg.ifs.tuwien.ac.at/projects/in2vis/

Abstract

Tracking and comparing psychotherapeutic data derived from questionnaires involves a large number of highly structured, time-oriented parameters. Descriptive and other statistical methods are only suited for partial analysis. Therefore, we invented a spring-based interactive Information Visualization method for analysing these data more in-depth. With our method the user is able to find new predictors for a positive or negative course of the therapy due to the combination of various visualization and interaction methods.

1 Introduction

Visualization tools have been used in the medical domain for a long time. The majority of methods and tools was developed for the field of scientific visualization, for example 3D volume visualization tasks or computer tomography visualizations. In the field of medical treatment planning different kinds of data need to be explored in the reasoning process, such as time-oriented patient data or the course of the patient state. Information Visualization (InfoVis) techniques can be used to support this exploration process and promote a deeper level of understanding of the data and information under investigation. To complement them, task-specific interaction methods have to be developed.

We are aiming for supporting a psychotherapeutic study on anorexic girls where high dimensional, abstract, time-oriented medical data is collected. The analysis of these data is a challenging process. One way to deal with this problem would be to apply statistical methods. However, these methods are limited to prove known hypotheses and can hardly capture the complex process of therapeutic interventions with not yet discovered hypotheses. To overcome this limitation, we are investigating interactive information exploration techniques.

We have developed a new interactive InfoVis method, called Gravi++[1], to observe new interdependencies between various kinds of parameters. Gravi++ uses the capabilities of the human perceptual system by displaying moving icons on the screen following a spring-based model to facilitate the reasoning process. Furthermore, this process is strongly supported by task-specific interaction methods.

[1]The name Gravi++ is a mixture of two metaphors. “Gravi” stands for gravitation and the two summation signs stand for two magnetic plus poles.

2 The Medical Problem

Gravi++ is intended to analyze questionnaires of girls with eating disorders (Anorexia Nervosa (AN)). At the Department of Child and Adolescent Neuropsychiatry, Medical University of Vienna, a study is taking place, in which alternative therapeutic processes (cognitive behavioral therapy) of anorexic girls are compared. The psychologists and physicians who are working with the girls need to explore the data in an experimental way in order to derive time-oriented quantitative and qualitative information on the states of the patients. The important features of the involved data structures are their data types, complexity (highly structured), and temporal dimensions.

Data Characteristics

Every patient, their parents, and their therapists have to answer an extensive set of questionnaires before, during (three-monthly), and after the therapy (each time 20 questionnaires per patient, four per parent and one per therapist). Each questionnaire consists of about 40 questions. An additional challenge is that the questionnaires are not all asked in the same interval. Some are even used only once. After a normalization process the data derived from these questionnaires ranges from ’0’ to ’6’. This range can mean ’yes’ to ’no’, but can also stand for nominal values like ’feeling guilty after eating a meal’.

Example questions of different questionnaires are:
FAMOS14: ’To relax is’: (1) totally unimportant (2) unimportant, (3) does not matter, (4) rather important to me, (5) extremely important to me.
MRFSF1: ’I treat myself to tranquility and recreation’: (1) no, (2) rather no, (3) rather yes, (4) yes.
EAT13: ’I feel sick after eating’: (0) never, (1) seldom, (2) sometimes, (3) often, (4) very often, (5) always.

Some questions are concatenated to so called predictors. These predictors should give an idea about a specific status of a patient and is used to predict the further development of the therapy.

The following predictors have been identified during our analysis: (1) Predictors for a negative therapeutic course are ’lacking close friendships’, ’comorbid personality disorder’, ’low self directedness’, ’lacking sincerity in answering questionnaires because of highly social desirable an-

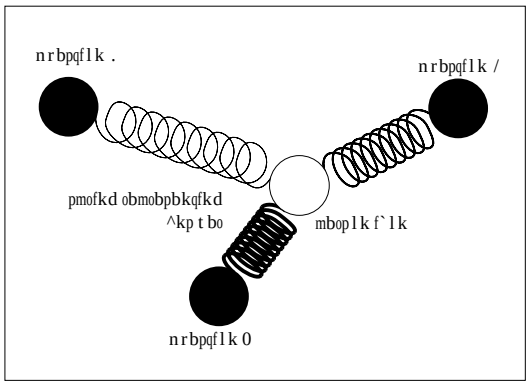


Figure 1: General Principle: A person is attracted by three questions. The answer is modelled with an invisible spring. The higher the question is answered, the stronger the spring pulls to the black disk. We plotted the springs to illustrate this general principle.

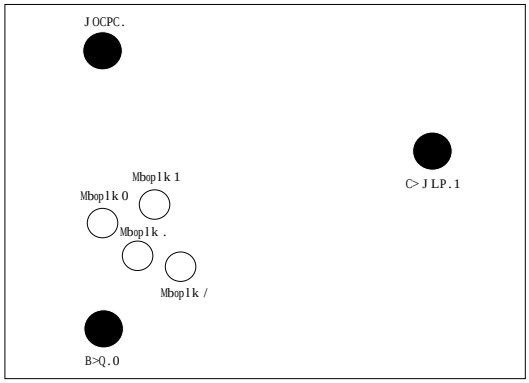


Figure 2: Simple example of finding a cluster: All persons gave similar answers to the three questions. Therefore, a cluster of all person's icons can be seen.

swering style', and 'denying disease'; (2) Predictors for positive course are a 'decreasing depression score after half a year of cognitive behavioral therapy (CBT)' or an 'increasing value of feelings of pleasure by doing favoring activities after 3 months of CBT'.

The task of finding new predictors, like the ones mentioned above, lead to the development of the new visualization technique Gravi++.

3 Related Work

Different kinds of InfoVis methods were developed in the last years. Important features of InfoVis methods are to support the exploration process of complex, heterogeneous data promoting a deeper level of understanding of the data and information and to foster new insights into the underlying exploration process and the data themselves [Ware, 2000; K. Card, 1999].

Medical highly structured data, such as psychotherapeutic data, impose an additional challenge due to their complexity and their temporal dimensions. Worm Plots [Matthews and Roze, 1997], the Zoom Star solution [Noirhomme-Fraiture, 2002], the TimeWheel [Tominski *et al.*, 2003], the Table Lens [Rao and Card, 1994], Stardinates [Lanzenberger *et al.*, 2003b], and LinkVis [Lanzenberger *et al.*, 2003a] are techniques that try to visualize and explore such kinds of data. However, none of these techniques provide enough interaction possibilities to find new interdependencies (such as predictors) and explore the data thoroughly. Furthermore, we are dealing with a huge amount of highly structured time-oriented data, which needs appropriate methods to analyze and to discover patterns. Therefore, we created Gravi++ with the help of the following concepts.

We were inspired by the outstanding idea of the Vibe System [Hendley *et al.*, 1995] to position document icons according to the occurrence of keywords with a spring-based system and adapted that idea for our core visualization. A similar approach was proposed in the RADVIZ method [Brunsdon *et al.*, 1998] to map a set of m-dimensional points onto two dimensional space. The idea for the different metaphors was inspired by the work of [McGinn and Picking, 2003]. Animation is one way to effectively visualize temporal changes, which is shown in [Nakakoji *et al.*, 2001]. Furthermore, to present another view of multidimensional data, we used Star Glyphs such as those presented in the XmdvTool [Ward, 1994].

4 Gravi++

4.1 Concept

The human perceptual system has a remarkable ability to organize and locate things spatially, judge comparative sizes, distinguish between a large range of colors and patterns and perceive motion [Olsen *et al.*, 1993]. Gravi++ tries to utilize these human capabilities by positioning icons on the screen. There are two kinds of icons representing (1) patients and (2) questions from the questionnaires respectively (compare Figure 3). Every patient is attracted by the questions according to the answer she gave. This is modeled with a spring-based system. Every question is connected with every person by a spring. This is illustrated in Figure 1 showing a person who is attracted by three questions.

The strength of the individual springs depends on the answer the patient gave. This way, every persons' icon position on the screen identifies how she answered the involved questions. This leads to the formation of clusters of persons who gave similar answers. Because of the fact that things that are close together are perceived as a group, the finding and differentiation of clusters is an easy task for the human perceptual system according to the Gestalt Laws [Koffka, 1935]. In Figure 2 all persons gave similar answers to the questions MFRS1, FAMOS14, and EAT13. Therefore, their icons form a cluster near EAT13. Furthermore, this tells us, that all persons have answered EAT13 with a higher value than the other questions.

The size of a person's icon can be mapped to any additional parameter (for example to the body mass index of the patients) or to the attraction force. In the second case the sphere is larger if it is attracted by higher values. This fea-

ture helps discriminating different icons that are attracted by the same values with a different coefficient. For example, if a person has answered the questions one, two, and three with answer number one, the icon is on the same place as a person that has answered each question with answer number five.

To visualize the changing values over time, Gravi++ uses animation. The position of each person's icon change over time allowing to trace, compare and analyze the changing values. Alternatively the change over time can also be represented by traces. The size and path of the person's icon is shown corresponding to all time steps or only to a restricted subset like the previous and the next time step.

To visualize the exact values of each question, rings around the question's icon can be drawn. The ring size corresponds with the attraction to the question. To avoid overlapping rings with the same value, they are put closely side by side.

In addition, Star Glyphs [Ward, 1994] can be shown, which communicate the exact values. The edges of the Star Glyph are connected with the corresponding question rings and both are drawn in the same color as the person icon. This helps the user to identify the corresponding person.

Missing data is handled by the system in two ways. If a person has answered no questions at a specific time step, the icon of the person is transparent. If a value of the size of a person's icon is missing, the icon is shown with a special marking.

Gravi++ is intended for a restricted parameter space. The more questions are selected, the smaller is the influence of a single question on the position of the person's icons. Furthermore, too many person icons lead to clutter because of overlapping icons. To select a suitable subset of parameters, we have implemented a list-based overview visualization.

The main aim of the visualization is to derive predictors. In the following, different kinds of interactions are explained to support that task.

4.2 Interactions with Gravi++

Gravi++'s main intention is to provide functionalities to get new insights into the data, like finding clusters or similarities in the movement over time. Furthermore, icons that are drifting apart can give important clues regarding the data. For this purpose Gravi++ provides a set of interactions.

All interactions can be classified in three categories: (1) interactions on question's icons, (2) interactions on person's icons, and (3) general interactions. In the following these categories are explained in detail.

Interactions on Question's Icons

- *Add or remove a question.*
- *Change the position of a question:* The position can be changed either freely or arranged on a circle. Furthermore, the user can choose to evenly space the distance between the questions on the circle. When changing the position of the questions, the corresponding positions of the person's icons change simultaneously. With this feature the user can interactively change the positions of the questions to find new clusters or other interesting visual structures. The traces and the Star Glyph also change their form automatically when dragging a question around. This feature is called 'live preview'.
- *Change the influence of a question* (see $strength_j$ in subsection 4.3): This enables the user to emphasize the influence of a chosen question.
- *Add complementary question:* By executing this option on a question, a new question with complementary values is created.
- *Highlighting a question:* By highlighting a question, only the highlighted question's rings are shown. If no question is highlighted, all rings are shown.
- *Hide question icons:* All question icons can be hidden to reduce clutter and facilitate the analysis of traces or the Star Glyph.

Interactions on Person's Icons

- *Add or remove a person's icon.*
- *Change the parameter representing the sphere size:* You can map the parameters of a single question, the power of attraction, or an increasing size over time to the sphere size.
- *Change the speed of movement:* Adjust how fast the person icons are moving to their destination.
- *Hide person icons:* All person icons can be hidden to reduce clutter and facilitate the analysis of traces or Star Glyphs.
- *Show traces:* The user has the option to show traces ranging over all time steps or only a restricted subset.

General Interactions

- *Save and load current settings and visualization:* The position of all elements and all settings can be saved for later analysis or logging purpose.
- *Show Star Glyph:* The user can superimpose a Star Glyph of the currently displayed data set to clarify the actual values. The corners of the Star Glyph correspond to the positions of the questions.
- *Inverse Star Glyph:* The edges of the Star Glyph are painted either to the center of the visualization or outwards.
- *Next and previous time step function:* This changes the time parameter to the next or previous one.
- *Direct selection of time steps:* The user can directly choose the time step on a time line or in a list.
- *Highlighting:* Persons and questions can be highlighted.
- *Show grid:* The user may fade in a grid in the background. This helps judging distances and sizes.

4.3 Algorithm

As we explained the idea and concept of Gravi++ in section 4.1, now we present the algorithm for positioning the person's icons. The force from each person i and each question j for each axis (in our case for the x and y axis) is $xforce_i = \sum_{j=0}^{n} strength_j \cdot value_{ij} \cdot f(qx)$

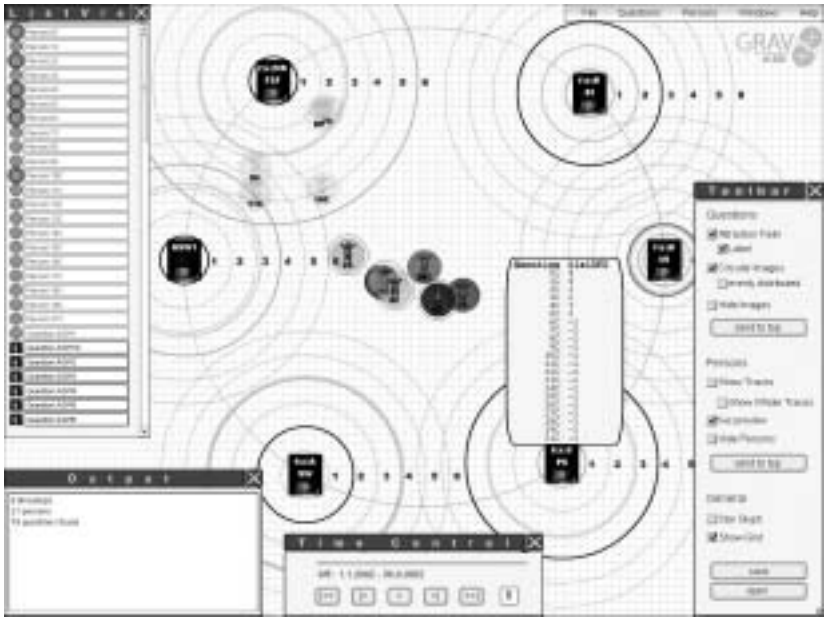


Figure 3: Typical application case with Gravi++.

and $yforce_i = \sum_{j=0}^{n} strength_j \cdot value_{ij} \cdot f(qy)$, where $strength_j$ stands for the manually set strength multiplicator of the question j (can be altered in the user interface and ranges from 0 = no attraction to n = an arbitrary value). $value_{ij}$ stands for the answer of person i to question j. $f(qx)$ and $f(qy)$ stands for the attraction over the distance for each axis. Currently, we use a linear function that grows stronger over the distance: $f(qx) = qx_j - px_i$, $f(qy) = qy_j - py_i$. Here qx_j and qy_j stand for the x, y coordinates of the question j and px_i and py_i stand for the x, y coordinates of the person i.

Now we can analytically solve the equations by substituting the force with zero ($xforce_i = 0$ and $yforce_i = 0$), to find the final position for each person (px_i, py_i).

4.4 Implementation

Gravi++ was implemented as a prototype during the in2vis project[2]. The prototype was implemented in Macromedia Flash MX 2004 because of its rapid visual development possibility.

The system consists of two visualizations that work closely together. There is an overview visualization (ListVis), to select a subset of a large data set. This subset can then be analyzed with the main visualization Gravi++. The data exchange between these two modules is implemented by drag and drop. You can simply drag a person or question to Gravi++ and explore it further. Both modules support linking and brushing.

Gravi++ was implemented with different metaphors which can be exchanged. In the first of the currently two implemented metaphors the persons are symbolized by iron spheres that are attracted by magnets standing for the questions. The circles around the magnets stand for the magnetic fields and visualize the concrete answers. The second metaphor shows the persons as people and maps the questions to paintings in an art exhibition. Here the metaphor is explained by persons who are attracted by beautiful paintings. The person moves closer to those paintings she likes. The circles around the images stand for how much the person likes the image or not. Other metaphors can be implemented by exchanging the icons representing persons and questions.

The persons are color-coded with twelve distinct colors as Colin Ware proposed in [Ware, 2000]. To enable the user to recognize icons that lie on top of each other, all icons are drawn slightly transparent. To control the time steps, a time line was implemented showing the current time step and providing the "video-like" controls *play, rewind, stop, go to first/last frame, go to next/previous frame*. Tool tips give complete information on the data item pointed at with the mouse.

In Figure 3 you can see a screenshot of a typical application case.

5 Example: How to Find a Predictor

Predictors allow to assess the development of parameters after a certain treatment step had been performed. Therefore, we have to analyse the differences in the first answers and compare them with the differences over all timesteps. This can be done by watching the animations over time or analysing the plotted traces of the persons's icons. Furthermore, the change over all timesteps can also show some sort of pattern that indicates a new predictor.

Experimenting sessions with our partners of the Department of Child and Adolescent Neuropsychiatry identified the following steps to find predictors. We will illustrate these steps using an example of our dataset consisting of nine patients and 73 questions in six time steps:

[2]Interactive Information Visualization: Exploring and Supporting Human Reasoning Processes

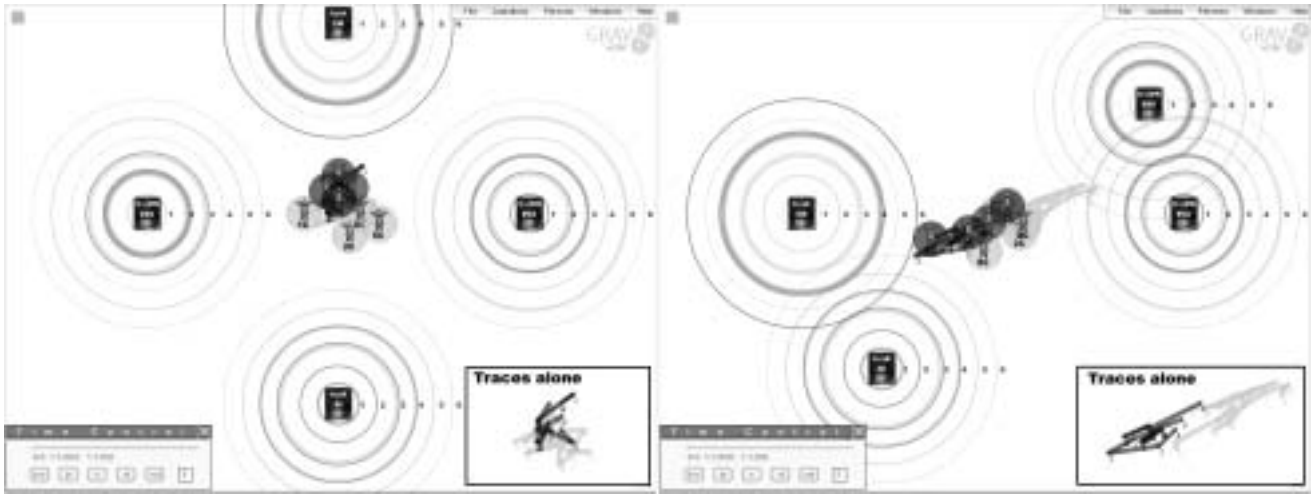


Figure 4: The left-hand screenshot shows four questions and no discernable pattern in the traces. The right-hand screenshot shows the same questions ordered differently with two easyly distinguishable clusters in the traces.

1. We choose the questionnaires ASW (inverted)[3], BDI[4], MR EVA[5], MR SOC[6] and all persons from the overview visualization and add them to Gravi++.
2. With the enabled option "Full Traces" we can analyse the traced paths of the persons over time.
3. By moving around the questions we try to find clusters of traces. In our example we have nine different persons. They are color coded the following way: The icons and traces with dark grey shades had not a favourable outcome in the therapy. The icons and traces with light grey shades had a positive therapy progress. This means we have to find clusters of evenly coloured traces.
4. By positioning MR EVA and MR SOC on the one side and ASW (inverted) and BDI on the other side we can build two clusters of light and dark traces. In the left screenshot on Figure 4 you can see the four questions before positioning (with no identifyable clustering of the traces) and on the right screenshot afterwards (with two identifyable clusters of light and dark traces).
5. By moving through the timesteps we can see that the randomly located persons in the first timestep, compose two clusters in the second and all later timesteps. Therefore, we can conclude that high values of BDI and the inverted ASW after three months therapy predict a bad outcome of the therapy, whereas high values of MR EVA and MR SOC predict a good outcome. Furthermore, we verify this claim by watching the movement over all remaining timesteps. In our example the persons with a positive therapy outcome still move further to the pole with MR EVA and the MR SOC and the persons with a negative outcome to the other pole. This means we can conclude that low BDI values and high MR EVA, MR SOC, and ASW values could be a predictor for a positive therapy outcome. This is confirmed by the clinical impression that patients who repsond more rapidly to the primary therapeutic goals (enjoying pleasure and being in a good mood), have a better outcome.

6 Benefits and Limitations

The visual elements of Gravi++ use several advantages from a cognitive perspective [K. Card, 1999]. Especially, the interactive manipulation can help the user get new insights through the data. This can be used to formulate and test a hypothesis on the data e.g., to find a new predictor. The combination of different visualization techniques, like Star Glyph, traces, an overview visualization, and the Gravi++ core itself increases the possibilities to find new insights. What is special about Gravi++ is the combination of these advantages and its orientation on medical data.

Nevertheless, some problems and shortcomings of Gravi++ are still not solved. Incomplete data leads to incomparable person icon positions because there is no attraction from questions that were not answered. A solution to this problem could be to use the value from the last time step, to use an average value, or a default value. A restriction of Gravi++ is the parameter space. Too many questions lead to clutter and make the interactive change very difficult. The impact of one question on the person's position declines the more questions are shown. Furthermore, if too many persons are analyzed, many icons would overlap and it would be difficult to interact with the representation. The values of the rings surrounding each question would be in-

[3] ASW: The Self-Efficacy Scale Index is a summary of the test items assessing self-esteem and strategies to cope with difficult situations independently.

[4] BDI: The Beck-Depression-Index describes the severity of depression.

[5] MR EVA: The Marburg Inventory Index specifies the feeling of pleasure in hedonistic behavior.

[6] MR SOC: The Sense of Coherence Scale Index indicates the intensity of comprehensibility, manageability and meaningfulness as the internal psychological mechanism mediating the effects of external stressors and resources on psychological dysfunction.

distinguishable. This happens because rings representing the same value are not shown on top of each other but next to each other. The solution for such situations ist to use highlighting of subgroups for detailed exploration.

7 Conclusion and Future Work

We have presented an interactive InfoVis method, called Gravi++, which addresses the particular features of abstract, highly structured data which are acquired during cognitive behavioral treatment (CBT) of anorexia nervosa in adolescent girls. This data is difficult to explore by descriptive and other statistical methods. Because of Gravi++'s various visualization and interaction techniques it is an appropriate method for finding new predictors in the data. Our cooperating psychologists see various application areas for this InfoVis technique within their clinical study of anorexic girls.

A possible extension to Gravi++ could be an algorithm that automatically positions the questions to find clusters of persons.

In the next step we plan to integrate the results of an in-depth user interface study with about 20 participants. Afterwards, a large number of subjects will be involved in a study, that will compare Gravi++ with supervised machine learning and exploratory data analysis to get new insights on the impact on the human reasoning process.

Acknowledgements

This project is supported by The Vienna Science and Technology Fund ('Wiener Wissenschafts-, Forschungs- und Technologiefonds' - WWTF), grant WWTF CI038. Many thanks to the anonymous reviewers who helped us to improve our paper.

References

[Brunsdon *et al.*, 1998] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton. An Investigation of Methods for Visualising Highly Multivariate Datasets. *Case Studies of Visualization in the Social Sciences. Joint Information Systems Committee / ESRC*, pages 55–80, 1998.

[Hendley *et al.*, 1995] R. J. Hendley, N. S. Drew, A. M. Wood, and R. Beale. Narcissus: Visualising Information. In Nahum D. Gershon and Steve Eick, editors, *Proceedings of the IEEE Symp. Information Visualization (InfoVis '95)*, pages 90–96. IEEE Computer Society Press, 30–31 1995.

[K. Card, 1999] B. Shneiderman K. Card, J. D. Mackinlay. *Readings in Information Visualization*. Morgan Kaufman, 1999. Chapter 1.

[Koffka, 1935] Kurt Koffka. *Principles of Gestalt Psychology*. Harcourt-Brace, 1935.

[Lanzenberger *et al.*, 2003a] Monika Lanzenberger, Silvia Miksch, Susanne Ohmann, and Christian Popow. Applying Information Visualization Techniques to Capture and Explore the Course of Cognitive Behavioral Therapy. In *Proceedings of the 2003 ACM symposium on Applied computing (SAC '03)*, pages 268–274. ACM Press, 2003.

[Lanzenberger *et al.*, 2003b] Monika Lanzenberger, Silvia Miksch, and Margit Pohl. The Stardinates - Visualizing Highly Structured Data. In *Proceedings of the Seventh International Conference on Information Visualization (InfoVis '03)*, pages 47–52. IEEE Computer Society, 2003.

[Matthews and Roze, 1997] Geoffrey Matthews and Mike Roze. Worm Plots. *Computer Graphics and Applications, IEEE*, 17(6):17–20, 1997.

[McGinn and Picking, 2003] John McGinn and Richard Picking. The Argument-as Metaphor in Decision-making Visualisation. volume Seventh International Conference on Information Visualization (InfoVis '03), pages 596–599. IEEE Computer Society, 2003.

[Nakakoji *et al.*, 2001] Kumiyo Nakakoji, Akio Takashima, and Ysuhiro Yamamoto. Cognitive Effects of Animated Visualization in Exploratory Visual Data Analysis. *Fifth International Conference on Information Visualisation, IEEE Computing Society (InfoVis '01)*, pages 77–84, 2001.

[Noirhomme-Fraiture, 2002] M. Noirhomme-Fraiture. Visualization of Large Data Sets: The Zoom Star Solution. *International Electronic Journal of Symbolic Data Analysis*, 0(0), 2002.

[Olsen *et al.*, 1993] Kai A. Olsen, Robert R. Korfhage, Kenneth M. Sochats, Michael B. Spring, and James G. Williams. Visualization of a Document Collection:The VIBE System. *Information Processing & Management*, 29(1):69–81, 1993.

[Rao and Card, 1994] Ramana Rao and Stuart K. Card. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+Context Visualization for Tabular Information. In *Proceedings of the ACM Conference Human Factors in Computing Systems (CHI '94)*, pages 318–322. ACM, 1994.

[Tominski *et al.*, 2003] Christian Tominski, James Abello, and Heidrun Schumann. Interactive Poster: Axes-Based Visualizations for Time Series Data. In *Poster Compendium of IEEE Symposium on Information Visualization (InfoVis '03)*, pages 68–69. IEEE, 2003.

[Ward, 1994] Matthew O. Ward. XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data. In *Proceedings of the conference on Visualization (Vis '94)*, pages 326–333. IEEE Computer Society Press, 1994.

[Ware, 2000] Colin Ware. *Information Visualization*. Morgan Kaufmann Publishers, San Francisco, 2000.

INTELLIGENT INTERFACE FOR ADJUVANT TREATMENT PLANNING IN BREAST CANCER

I.H. Jarman, T.A. Etchells and P.J.G. Lisboa
Liverpool John Moores University, UK

Abstract

The primary course of therapy for breast cancer patients, following surgery, depends on the expected prognosis together with the key clinical indicators. An interface for use by clinical oncologists is proposed, which addresses three fundamental questions, namely; evidence that the currently used Nottingham Prognostic Index can be enhanced by additional clinical features, prognostic inference for individual patients with quantified confidence levels, and visualisation of the patient database by clinical indicator of adjuvant treatment. This interface is underpinned by detailed prognostic analysis validated through longitudinal cohort studies of mortality with 931 TNM stage I/II patients recruited between 1990 and 1993 at Christie Hospital, Wilmslow. The data shown in the interface are Kaplan-Meier curves from prognostic risk groups inferred by cross validation.

1. Introduction

The starting point for this paper is the commonly used clinical prognostic index for breast cancer, the Nottingham Prognostic Index (NPI) [1]. While widely applied to inform the choice of adjuvant therapy, advances in therapy, detection technologies and health policy, such as the introduction of breast cancer screening for women aged 50 and over in the UK, has skewed the patient population and has added potential prognostic indicators. This study proposes an interface for clinical oncologists to show the added value of expanding the covariate basis for prognostic inference. It is important to note that the basis of our approach is to keep NPI and expand rather than replace current practice.

Furthermore, there is now an interest in predictive inference of prognosis for individual patients, witness the web-based prognostic interface www.Adjuvantonline.com [2]. This model is gaining clinical support in part because it infers the potential effect of different treatment choices. It also points towards a visualisation format that appears to be readily accepted by practicing clinicians. However, the predictions made do not include confidence estimates, yet are likely subject to substantial uncertainties for particular groups of patients, notably in NPI group 3, which is known to be heterogeneous in its composition. Moreover, it is not clear that the development of this interface has followed the recommended staged process of the continuum of evidence, which is modelled on the development of medicinal drugs and is intended to assure the accuracy and generalisability of clinical inferences [3-4].

In addition to prognostic inference, a previous study of the prevalence of adjuvant treatment, typically hormone therapy e.g. tamoxifen, chemotherapy, or both in combination, showed that the different treatments are clustered primarily by key clinical important indicators of the likely response to treatment, namely oestrogen receptor count, lymph nodes affected and menopausal status [5]. The proposed interface switches between survival modelling and treatment allocation profiles.

In summary, there are two aspects of novelty presented, firstly a methodology for an individual prediction of survival with confidence intervals using neural networks and Monte Carlo methods. Secondly, implementing an interface that shows the added value of a new prognostic model over the current clinical standard prognostic model supported by a personal prognosis and data-based rules for treatment allocation.

The next section explains what the NPI is and how it was extended by modelling with additional covariates using Cox regression with the proportionality of hazards' assumption. This leads to the derivation of a cross-matching framework to discriminate between the survival of patients in each NPI risk group. Section 3 summarises the derivation of prognostic models with confidence intervals for individual patients, using Monte Carlo methods. This is followed, in section 4, by a brief overview of the rule extraction algorithm used to explain treatment allocation. Finally, the interface is described in its entirety.

2. Extended prognostic indices of survival

Survival analysis is an important field in medical statistics where the proportional hazard model [6], also known as Cox regression, is the most widely used method.

The form of the Cox regression model is:

$$h_0(t)\exp\left[\sum_{i=1}^{n}\beta_i x_i\right] \qquad (1)$$

where $h_o(t)$ is the baseline hazard function. $h_o(t)$ is called the baseline function because when all the x variables are equal to zero the formula reduces to this form, hence the 'baseline' of the model. β are the coefficients of x, which are the explanatory variables.

Cox regression is a semi-parametric model that incorporates censored data, which arises when an individual drops out of a study for reasons other than the event of interest, death due to breast cancer in this study. Omitting these data from a survival analysis can introduce significant bias to any results [7].

This forms the basis for a prognostic index that is clinically widely accepted, the Nottingham Prognostic Index (NPI), which uses 3 variables identified as being significant in the prediction of survival, namely; *pathological size of tumour, histological grade of tumour and the number of axillary nodes affected* and requires a calculation in the form of a simple equation, which for a clinician makes it easy to use and understand. In the case of NPI:

$$0.2*\text{pathological size} + \text{histological grade} + \text{nodes involved} \quad (2)$$

From this index, using the log-rank statistic, patients are allocated into 4 prognostic risk groups, ranging from very good to poor, at cut-off points < 2.41, < 3.41, < 5.41 and ≥ 5.41 respectively.

A further Cox regression model using six variables; *age, clinical stage nodes, histology, node ratio, pathological size and ER status* has been developed from 917 patients and validated on 931 patients from Christie Hospital near Manchester referred between 1983 – 89 and 1990-93 respectively. The latter dataset, the validation group, showed that the NPI and the new Cox model separated the patient profiles into prognostic groups with similar mean survival but with different risk group allocation, where NPI could be calculated (559 patients).

By cross-matching the two prognostic indices we are able to examine survival for patient groups within each matrix cell using Kaplan-Meier (KM) estimated survival curves in figure 1, in order to discover heterogeneity in estimated survival for any of the models prognostic groups. These differences in survival are an indication of the added value of cross-matching NPI with another survival model that uses additional variables and are providing supplementary information for prognosis.

This same idea of cross-tabulation can be extended to a scatter-plot of the prognostic indices, which allows the patient to be identified within this framework figure 2 and therefore identify how borderline a patient may be to adjacent prognostic groups or cells.

3. Individual Prognostic Predictions with Confidence Intervals

In addition to the detailed analysis of the group in which a particular patient belongs, there is interest in predictive inference of prognosis for that individual patient. The website http://www.adjuvantonline.com/ presents such information but without confidence intervals, so the uncertainties inherent in the prediction cannot be assessed. We present a method, using hazard predictions from a Partial Logistic Artificial Neural Network with Automatic Relevance Determination (PLANN-ARD) [8] and Monte Carlo methods, that give prognostic predictions with confidence intervals for individual patients.

The PLANN-ARD model provides a prediction of smooth estimates of the discrete time hazard. It is implemented as a direct extension of the Multi-Layer Perceptron (MLP) neural network applied as a discrete model of the hazard function. Using this MLP structure with time as an input we have

$$\frac{h_p(x_p,t_k)}{1-h_p(x_p,t_k)} = \exp\left(\sum_{h=1}^{N_h} w_h g\left(\sum_{i=1}^{N_i} w_{ih}x_{pi} + wt_k + b_h\right) + b\right) \quad (3)$$

Estimating the weights requires a likelihood term for the status of one patient at time t_k, by using an indicator label 0 if a patient is alive at time t_k and a label 1 for the event of interest. This generic non-linear model is called the Partial Likelihood Artificial Neural Network (PLANN) [9]. In contrast to a proportional hazards model [6], PLANN does not require proportionality of the hazards over time and predicts a smooth hazard function.

At time t_i the estimated summed weights to each output unit has a Gaussian distribution $N(a_i,\sigma_i^2)$ [10]. The estimated hazard is calculated by the sigmoidal activation:

$$h(t_i) = g(a_i) = \frac{1}{1+\exp(-a)} \quad (4)$$

Once the network weights are estimated, the survivorship is calculated from the estimated discrete time hazard by multiplying the conditionals for survival over successive time intervals treated as independent events, this gives:

$$S(t_k) = \prod_{l=1}^{k}\left(1-h(t_l)\right) \quad (5)$$

Estimating an individual prognosis for patient x we use Monte Carlo methods by taking a random sample $\tilde{a}_i$ from $N(a_i,\sigma_i^2)$, calculate $\tilde{h}_i = g(\tilde{a}_i)$ and finally estimate survival $\tilde{S}(t_k)$. Repeat these steps n times, enough to build up a distribution of survival estimates, as shown in figure 3. The personalised prognosis is the mean survival of the distribution with 95% confidence intervals determined by omitting the upper and lower 2.5% of the sample estimates.

The survival estimate can be presented as a simple colour coded green, amber and red bar representing probabilities of survival, with 95% confidence interval

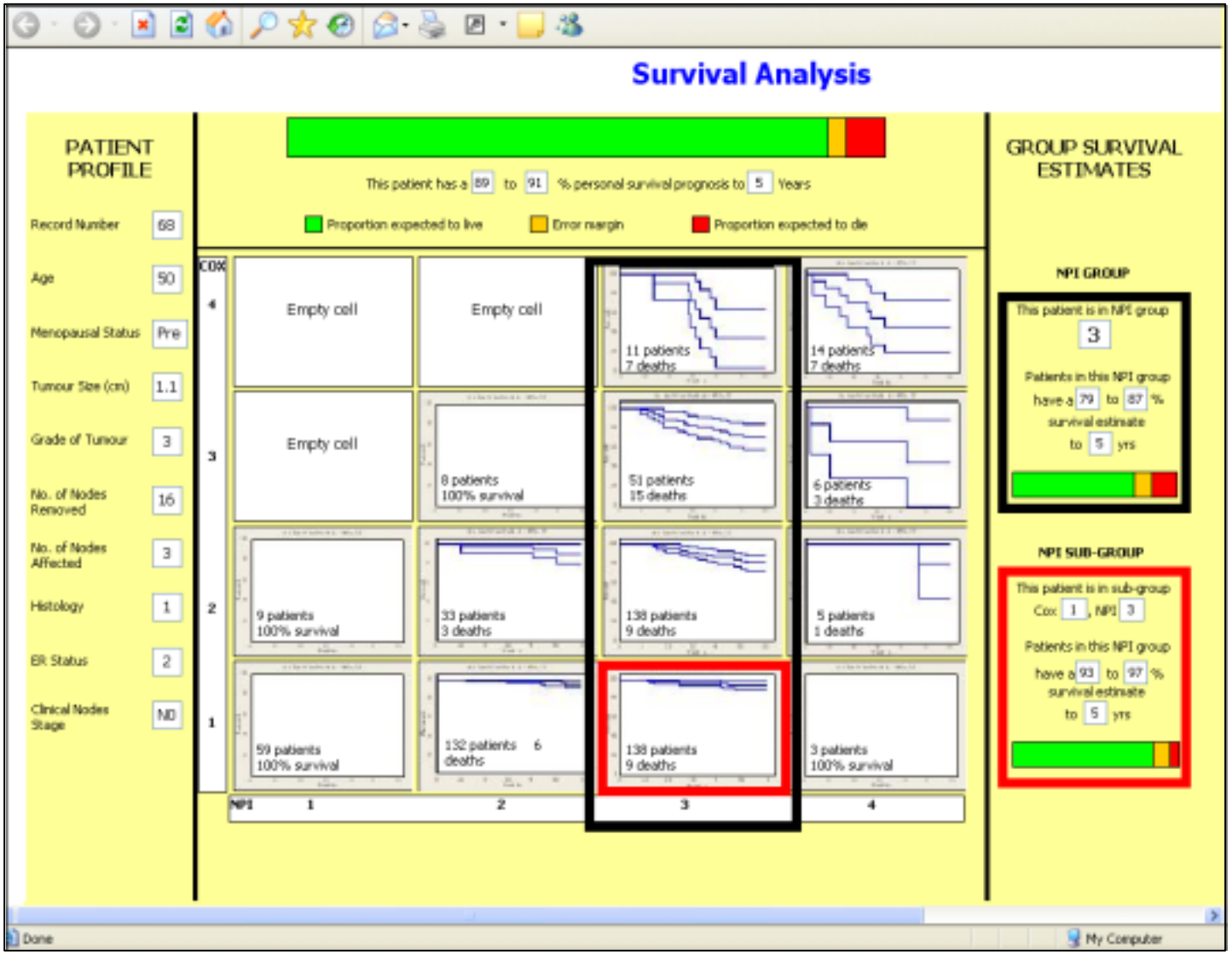


Figure 1. Proposed intelligent interface for survival. To the left is the patient profile containing the significant variables that affect both survival and treatment choice. Top Middle is the individual estimated survival bar indicating, from left to right; the proportion expected to be alive at 5 years, the degree of uncertainty in the prediction and the proportion expected not to survive. The right and centre graphic represent Kaplan-Meier survival estimates, highlighting the patient's NPI group and the sub-group within NPI from knowledge gained by cross-matching NPI with another prognostic model using additional variables, this cell also shows the Kaplan-Meier estimated survival curve for the time period of interest.

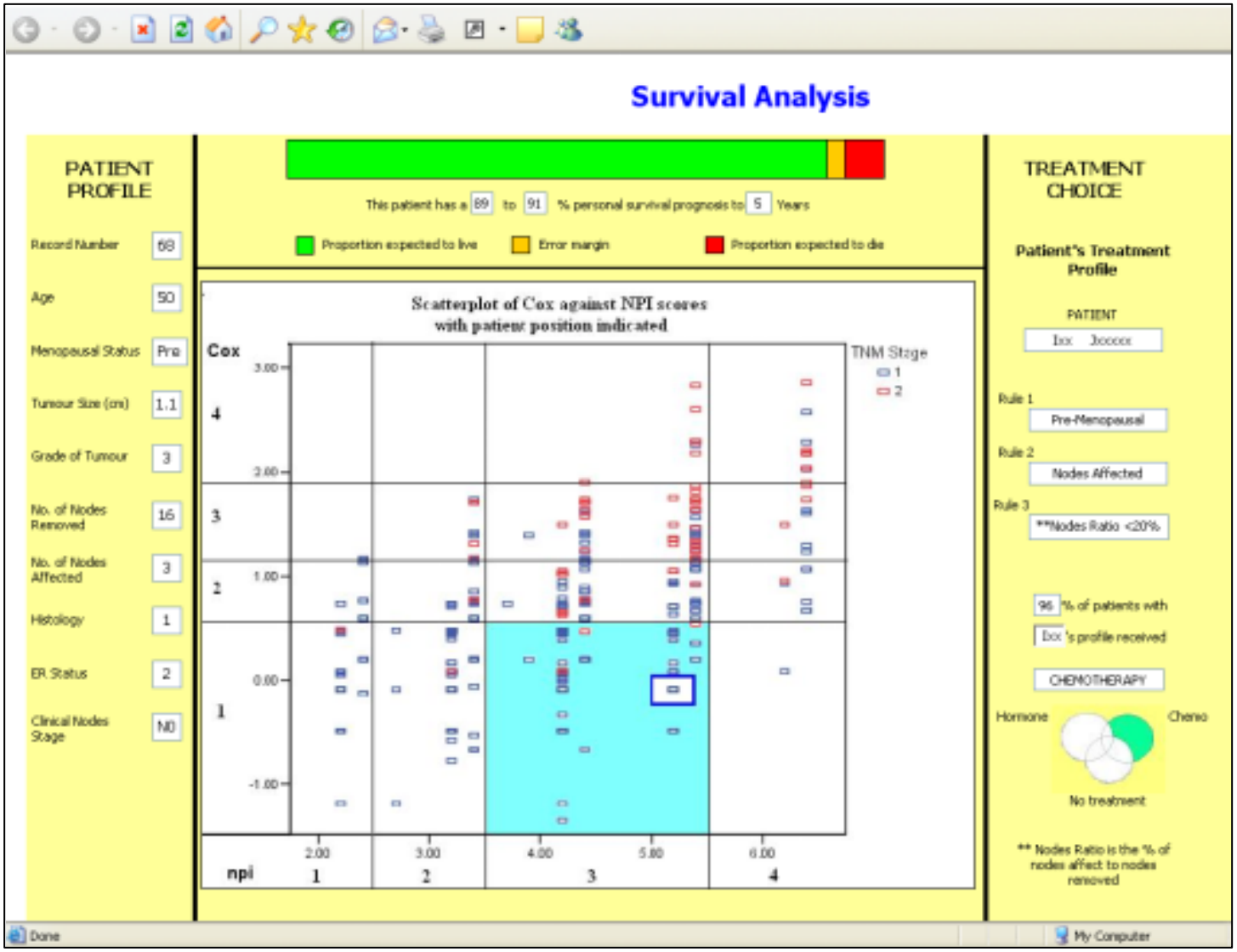


Figure 2. Proposed intelligent interface for treatment. To the left is the patient profile containing the significant variables that affect both survival and treatment choice. The centre graphic represents a scatterplot of the Cox and NPI scores, the patient position and cell within the cross-tabulation matrix is highlighted. To the right there are empirical rules describing a similar group of patients stating the proportion that received the recommended course of treatment

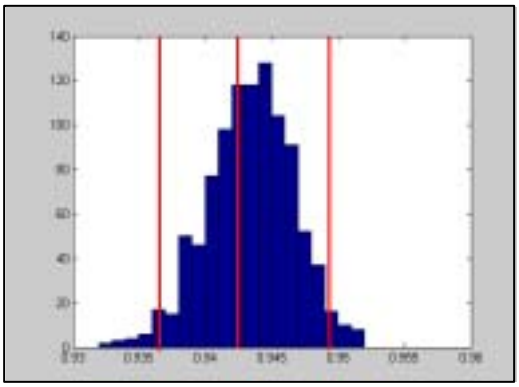


Figure 3. Distribution from 1000 iterations of estimated survival for an individual patient with mean survival and 95% confidence intervals indicated in red

and death respectively to a particular time period of interest, 5 years in this study figure 1.

4. Data-based Rules to Describe the Patient Allocations Made by the Analytical Risk Scores

In this paper the Orthogonal Search Rule Extraction (OSRE) algorithm [11] is used to extract rules for the treatment of the 559 patients in this study.

The OSRE algorithm finds conjunctive rules for classifications of data using a Multi-layer Perceptron (MLP), or any other smooth response surface, that has been trained to accurately predict the classifications of a dataset. A detailed account of the algorithm and the mathematical framework that underpins it can be found in chapters 3 & 4 of [12].

In essence OSRE searches for changes in response from an MLP, starting from each data point in turn in the data set and systematically searching in orthogonal directions. To demonstrate the algorithm we take a data set that has three variables and each variable has values ranging from 1 to 6. Figure 4 shows the data-space and a surface boundary that separates the in and out of class data. The arrows show the directions in which the algorithm searches for changes in the response of the surface. Notice that in the direction of the variable a1, there is no change in response from the surface.

The consequence of there being no change in response of the surface for a particular variable is that the variable does not feature in the set of conjunctive rules for this surface. Figure 5 shows the 'hyper-box' that the algorithm generates for the data-point represented in figure 4.

The rule generated from the 'hyper-box' is

$$(a1 \leq 6) \text{ AND } (a2 \leq 4) \text{ AND } (a3 \geq 3)$$

or more simply, as a1 takes all possible values,

$$(a2 \leq 4) \text{ AND } (a3 \geq 3).$$

This process is repeated for each data-point for which the surface predicts it to be in-class. A set of rules the size of the number of data predicted in-class is generated. The algorithm is enhanced with a refining method to reduce the number of explanatory rules conditional on maintaining sensitivity and specificity values above minimal acceptable thresholds [5].

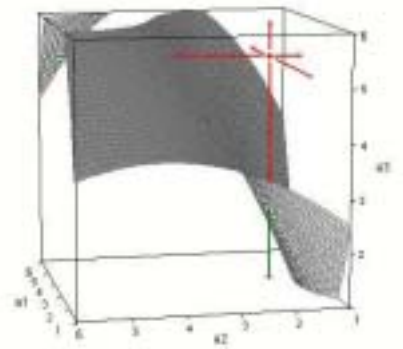


Figure 4. OSRE searching for changes in response in orthogonal directions.

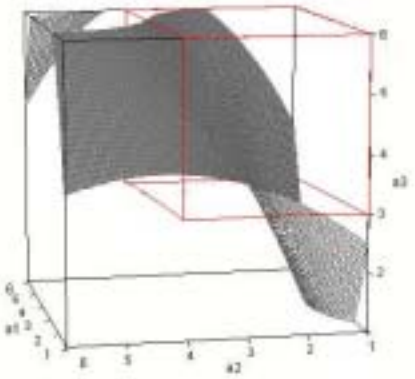


Figure 5. The OSRE algorithm generates a 'hyper-box' from which conjunctive rules are found.

5. Integrated Intelligent Interface for Breast Oncology

Combining all the elements described above enables us to present an integrated intelligent interface for clinicians. This is achieved with the cross-matching matrix where each column represents patients in prognostic risk groups for the current standard NPI model, the rows representing the risk groups for the new prognostic prediction. This can inform the clinician on a patient's survival outcome (figure 1) as well as giving NPI survival estimates with extended survival predictions for sub-groups within the cross-tabulation matrix. This allows the clinician to assess heterogeneity in survival within a prognostic risk group. Presenting a new model as an extension of NPI enables clinicians to relate to their own reasoning model, where the use of the current prognostic group allocation assists as an indicator for choice of therapy [13]. In addition the bar graphic above the cross-matching matrix presents the individual prognosis with 95% confidence intervals

Multi-Classification of Clinical Guidelines in Concept Hierarchies

Diego Sona[1], **Paolo Avesani**[1] **and Robert Moskovitch**[2]
[1] ITC/irst - Trento, Italy
{sona,avesani}@itc.it
[2] Ben Gurion University - Beer Sheva, Israel
robertmo@bgumail.bgu.ac.il

Abstract

Clinical practice guidelines (CPGs) are increasingly common in clinical medicine for prescribing a set of rules that a physician should follow. Recent interest is in accurate retrieval of CPGs at the point of care. Examples are the CPGs digital libraries National Guideline Clearinghouse (NGC) or Vaidurya (DeGeL), which are organized along predefined concept hierarchies, like MeSH and UMLS. In this case, both browsing and concept-based search can be applied. Mandatory step in enabling both ways to CPGs retrieval is manual classification of CPGs along the concepts hierarchy. This task is extremely time consuming. Supervised learning approaches, where a classifier is trained based on a meaningful set of labeled examples is not a satisfying solution, because usually too few or no CPGs are provided as training set for each class.

In this paper we present how to apply the *TaxSOM* model for multi-classification. *TaxSOM* is an unsupervised technique that supports the physician in the classification of CPGs along the concepts hierarchy, even when no labeled examples are available. This model exploits lexical and topological information on the hierarchy to elaborate a classification hypothesis for any given CPG. We argue that such a kind of unsupervised classification can support a physician to classify CPGs by recommending the most probable classes. An experimental evaluation on various concept hierarchies with hundreds of CPGs and categories provides the empirical evidence of the proposed technique.

1 Introduction

Clinical practice guidelines (CPGs) are an increasingly common and important format in clinical medicine for prescribing a set of rules and policies that a physician should follow. According to studies, clinical guidelines improve medical practice. They improve the quality (and possibly also the cost-efficiency) of care in an increasingly complex health care environment [Grimshaw and Russel, 1993]. It would be best if automated support could be offered to guideline-based care at the point of care.

To support tasks such as the run-time application of a guideline, it is often important to be able to quickly retrieve a set of guidelines most appropriate for a particular patient or task. Correctly classifying the guidelines, along as many semantic categories as relevant (e.g., therapy modes, disorder types, sighs and symptoms), supports easier and more accurate retrieval of the relevant guidelines using concept based search. This approach is implemented in *Vaidurya* – a concept based and context sensitive search engine for clinical guidelines [Moskovitch *et al.*, 2004], which is the search engine of the *Digital Electronic GuidelinE Library* (DeGeL). Electronic CPG repository, such as the National Guideline Clearinghouse (NGC) provide a hierarchical access to electronic CPGs in a free-text or semi-structured format (see <http://www.ngc.org>).

The construction of such concept hierarchies and the consequent classification of CPGs along the provided concepts is usually committed to physicians that in the following of this paper will be also referred to as *"taxonomy editors"*. This classification, however, is mostly manual and extremely time consuming. Thus, an automatic process where CPGs are classified automatically along the concepts hierarchy is crucial, while very challenging.

The main aim of this paper is to provide a tool that assists the domain expert (physician), who classifies the CPGs. The idea is that whenever the physician needs to classify a set of CPGs, the tool provides recommendations on the most probable classes for each CPG. In particular, the tool is specially suited to help the physician when concept hierarchy is built from scratch, and no examples of labeled CPGs are provided for each class. In this case there is not any premise for a successful training of any existing supervised classifier, therefore, recommendations can be given only using an unsupervised model. We refer this task to as the *bootstrapping* problem [McCallum and Nigam, 1999; Adami *et al.*, 2003a]. Then, once the physician is provided with the set of recommended classes for each CPG she can select the most appropriate.

The interesting part of this approach is that while the physician manufactures the concept hierarchy she also inserts some prior knowledge on the desired organization of data. Actually, each new concept added to the hierarchy is usually labeled by a few keywords describing the supposed semantic meaning of its content. Moreover, the concept is related to other concepts (more specific, more general, related to, etc.). This prior knowledge is exploited by the

proposed model in order to perform a preliminary classification of CPGs according to their contents and the desired organization within the hierarchy.

The evaluation of the proposed approach has been performed on a set of real data selected from the above mentioned NGC database. The promising results showed that the approach can be valuable in order to create and populate new electronic hierarchical repositories of CPGs.

In Section 2 some research works related to medical knowledge management are discussed. Section 3 gives a description of the addressed task with some references to works aiming at solving similar problems. Section 4 introduces the model used to test the proposed solution. Section 5 describes the experimental setup. Finally, Sections 6 and 7 discuss the results of experiments and draw some conclusions respectively.

2 Related Works

Traditional text retrieval systems use the "vector space model" in which terms are extracted from the document and represented by either their term frequency as a *bag-of-words* or their term presence/absence as a *set-of-words*. The limitation of this approach is that humans search using concepts instead of terms. In the medical domain, concept-based search refers to a text retrieval approach where the documents are mapped to concepts based on their contents. SAPHIRE system [Hersh and Greens, 1989], for example, uses an approach in which concepts used for indexing are automatically extracted from the document. Actually, within biomedical domains, documents and queries are often mapped into a large vocabulary such as MeSH (see <http://www.nlm.nih.gov/mesh>) or UMLS [Humphreys and Lindberg, 1993], which is one of the major resources offered by the National Library of Medicine. The concepts in these vocabularies are represented in a hierarchical structure. This approach is somewhat limited, since users aren't always familiar, while querying, with the concepts in these vocabularies. Moreover, several studies had shown that such implementation of concept-based search might actually decrease the retrieval performance [Hersh and Hickam, 1993], mainly because there are no good automatic concept extractors.

This hierarchical organization of documents, also allows browsing through the concepts using the hierarchical structure. Such a browsing method forces the user to navigate the conceptual hierarchical structure. Alternatively, in these directories, searches can be limited to a specific concept and its sub-concept contents. However, in the medical domain documents are usually classified by a multitude of concepts, often as many as a dozen or even tens of concepts.

An example of solution to this problem is provided by Vaidurya, a concept based and context sensitive search engine for clinical guidelines [Moskovitch *et al.*, 2004]. This engine implements a concept based search where the user has to choose few concepts and the logic relation between them. In his query the user defines a relevant subset of the collection, based on the conceptual indexing.

Recent results had shown that searching within a hierarchical concepts indexing improved full text retrieval, even at the first and second level of the hierarchy, especially when using conjunctive queries [Moskovitch and Shahar, 2004]. However, in order to implement an accurate concept based search manual classification should be applied by an expert, a very time consuming task. Thus, an automatic hierarchical classifier for clinical guidelines is crucial. At least it can help during the manual classification recommending the most probable concepts to be assigned to the documents.

3 Task Definition

A concept hierarchy (also referred to as taxonomy) is a hierarchy of categories (also referred to as classes) which are represented as nodes in a tree. Each node is described in terms of both linguistic keywords (also referred to as labels) that ideally denote the "semantic meaning" of the nodes, and relationships with other categories. The leaves of the tree represent specific concepts, while nodes near the root of the tree represent more general concepts. In our particular task, each node of the hierarchy can contain CPGs and, in general, each CPG can belong to more than one category.

Annotation of document to classes is a typical task in information retrieval. The goal here is to identify the set of categories that best describe the content of an unclassified CPG. A wide range of statistical and machine learning techniques have been applied to text categorization (see for example [Ceci and Malerba, 2003; Chakrabarti *et al.*, 1997; Cheng *et al.*, 2001; Doan *et al.*, 2003; Dumais and Chen, 2000; Joachims, 1998; Jordan and Jacobs, 1994; Koller and Sahami, 1997; Ruiz and Srinivasan, 2002; Sun and Lim, 2001; Wang *et al.*, 1999; Weigend *et al.*, 1999]). However, none of the above models can be used to solve the proposed task. Actually, these techniques are all based on having some initial pre-labeled documents, which are used to train a (semi)-supervised model. Moreover, Although many real world classification systems have complex hierarchical structure, few learning methods capitalize on this structure. Most of the approaches above ignore the hierarchical structure and treat each category or class separately, thus in effect 'flattening' the hierarchical structure. In the case this hierarchical structure is kept the models only classify on the leaves of the structure.

These problems are partially solved by the way we use the *TaxSOM* model [Adami *et al.*, 2003b]. The model uses the prior knowledge to drive a clustering process and, as a result, it organizes the CPGs on a given concept hierarchy without any need of supervision during training. Basically, the model bootstraps the given taxonomy with a preliminary classification of CPGs that afterward need to be reviewed by the taxonomy editor.

The basic idea of the *bootstrapping* process is to support and alleviate the manual labeling of a set of unlabeled examples, providing the user with an automatically determined preliminary hypothesis of classification. The idea is to exploit the linguistic and the relational information encoded within a taxonomy through an unsupervised learning model. The paper illustrates how *TaxSOM* can be used to learn the prior knowledge encoded within a concept hierarchy in order to perform this preliminary classification of

CPGs.

In particular, the task goal is to provide the user with a list of recommended classes for each CPG, i.e., the most probable k classes to which the CPG could belong.

4 Classification Models

A strategy to classify documents using prior knowledge is proposed by Yang [Yang, 1994]. Unlabeled documents are classified according to the lexical information associated to the categories. Specifically, a reference vector is built for each category, through the encoding of its labels. The documents are then associated to the category having the nearest reference vector (a standard prototype–based minimum error classifier). In the following, this simple class of keyword matching algorithms will be referred to as *baseline* categorization approach.

This classification method uses only lexical information, while topological information is neglected. To also use the hierarchical information we revised the *baseline* model according our scenario. Specifically, hierarchical knowledge was exploited building codebooks through the encoding of all labels in the current node and in all its ancestors, i.e., all labels of the nodes in the path from the root to the current node.

The above idea has been developed even more in the *TaxSOM* model [Adami *et al.*, 2003b]. Specifically, a *TaxSOM* is a collection of computational units connected so as to form a graph having the shape isomorphic to a given taxonomy. Such computational units, namely codebooks, are initialized as for *baseline*. Then an unsupervised training algorithm (similarly to Self Organizing Maps [Kohonen, 2001]) adapts these codebooks in order to take into account both the documents similarity and the constraints determined by the labels and the relationships. The basic idea is that once a *TaxSOM* has been properly trained the final configuration of the codebooks describes a clustered organization of documents that tailors the desired relationships between concepts.

The learning procedure of a *TaxSOM* is designed as an iterative process that can be divided into two main stages: a competitive step and a cooperative step. During *competitive* step the codebook most similar to the current input vector (a document) is chosen as the *winner* unit. In the *cooperative* stage all codebooks are moved closer to the input vector, with a learning rate proportional to the inverse of their topological distance from the winner unit. The iterations of the two steps are interleaved with an additional phase where the codebooks are constrained by the *a priori* lexical knowledge localized on the nodes.

5 Experimental Setup

We used the NGC CPGs collection to evaluate the suggested approach. The CPGs in the NGC hierarchy are classified along two hierarchical concept trees, Disorders and Therapies. Each concepts tree has roughly 1,000 unique concepts, in some regions the concepts trees are 10 levels deep, but the mean is 4 to 6 levels. There are 1136 CPGs, each CPG may have multiple classifications at different nodes by both concept trees and within the same tree. The classification is not necessary only on the leaves. CPGs have a mean of 10 classifications, while there exist CPGs classified by 90 concepts.

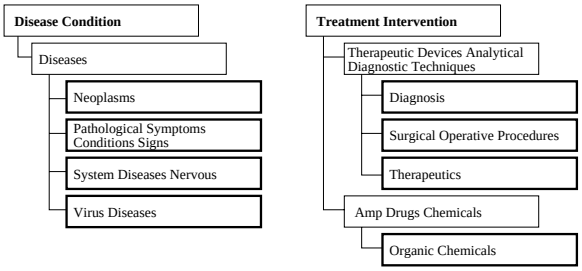


Figure 1: The eight selected taxonomies are sub-taxonomies of the two original concept hierarchies. Specifically, the eight leaves in the above two trees (dark bordered boxes).

To evaluate the model with a plurality of datasets, we decided to split down the two original dataset ("treatment intervention" and "disease condition") into eight smaller and different datasets (see Figure 1). These datasets were selected according to dimensional criteria decided in the beginning of our testing process – their depth (i.e., how far the leaves are from their root), the number of nodes and the number of CPGs. The variability of both topics and dimensions allows the evaluation of the model without biases due to any prior knowledge, such as topic vocabulary, dimension of taxonomy, number of classes for each CPG, etc.

Table 1 summarizes the statistics of the selected taxonomies. It can be seen that the depth of the hierarchies ranges from 5 to 9 layers, with few hundreds nodes. While the number of CPGs range from hundreds to thousand. More interestingly, many nodes are not represented by any CPG, therefore, supervised classifiers cannot be learnt on these datasets. The main characteristic of such datasets is that usually the leaves are not empty, while the interior nodes (i.e., nodes that appear as parent of other nodes) many times are empty (sometimes more that 50% of times).

Each taxonomy was preprocessed separately. The content of documents and the category labels were cleaned removing stop–words (articles, conjunctions, and prepositions) and reducing the vocabulary (i.e., the vector space representation) to 500 important keywords plus the labels of nodes. The important keywords were selected using the notion of Shannon Entropy[1]. Finally, CPG contents were encoded with a *set–of–words* representation (i.e., binary vectors).

As previously outlined, since to our knowledge there are not models devised to solve the proposed bootstrapping problem, we compared *TaxSOM* with the simple approach based on keyword matching refer to as *baseline*.

The model was tested on each taxonomy performing an hypothesis of classification for all CPGs, and the results were then compared with the original labeling. Actually, the addressed task requires the multi-classification of CPGs, therefore, given a CPG, both models generate a membership value for each class. These membership val-

[1] Shannon entropy is a standard information theoretic approach that can be used to measure the amount of information provided by the presence of a word in the dataset.

statistics	*taxonomies* diagnosis	neoplasms	organic chem.	pathol. sympt. cond. signs	surgical operat. proced.	system diseases nervous	therap.	virus
max tree depth	7	8	9	8	5	7	6	6
tot nodes	278	230	326	214	210	318	247	124
tot docs	1248	501	367	516	396	606	929	432
min docs/node	0	0	0	0	0	0	0	0
max docs/node	20	20	20	20	16	20	20	20
average docs/node	4.49	2.18	1.13	2.41	1.89	1.91	3.76	3.48
min w/doc	52	66	40	563	582	587	571	704
max w/doc	463	387	444	5132	5050	5726	6074	10599
average w/doc	218	223	232	1633	1378	1573	1707	261
docs on leaves	67%	63%	85%	62%	67%	66%	66%	61%
% of leaves	66%	55%	48%	62%	68%	56%	63%	52%
empty nodes	14%	23%	46%	21%	15%	25%	10%	30%
empty leaves	1%	3%	6%	4%	5%	9%	0%	0%
empty interiors	39%	46%	84%	50%	35%	46%	26%	64%

Table 1: Statistics of the selected concept hierarchies. The first group of rows describe the trees' dimension. The second group describes the datasets' dimension. The third group the documents' dimension. While the last two groups of rows describe respectively the distribution of CPGs in the hierarchies.

ues are then used to rank the classes, and this ranking is then used to select the best classes to recommend to the user. To evaluate the proposed two models we devised a specific measure – the *multi-classification k-coverage precision*. This measure allows a comparison of models rather than an objective evaluation.

The measure counts the percentage of CPGs "correctly" classified with respect to the total number of CPGs. The meaning of *k-coverage* is strongly related to the definition of "correct classification". In this case, a document is correctly classified when all the classes to which it belong are in the first k recommended classes. The idea is that the system provide the user with a set of probable classes with which to label a give CPG. If all the interesting classes are among the recommended k then the document is correctly classified.

For example, suppose we know that a CPG should be classified to three specific classes. If the model proposes all the three classes among the k recommended, then the document is considered correctly classified. If, on the contrary, the model fail to propose at least one of the three classes in the recommended k classes then the CPG is considered wrongly classified.

For example, a model having *k-coverage* equal to 60% for $k = 10$ means that for 60% of the documents all the corresponding correct classes appear in the first ten ranked classes.

6 Discussion of Results

The evaluation of the proposed model has been performed on all 8 smaller taxonomies and the two original taxonomies determining the *k-coverage* for all possible ks. In Figure 2 are depicted the graphs of the *k-coverage* for all ks for all the eight smaller taxonomies. It can be easily seen that for almost all reasonable ks the proposed *TaxSOM* model always outperform the *baseline* approach.

Actually, providing the best k ranked classes for each CPG (where k should be reasonable small to be explored by the physician) the probability of finding all the correct classes is higher for *TaxSOM* than for *baseline*. This means

taxonomies	*k-coverage* *baseline* $k = 10$	$k = 20$	$k = 10\%$	*TaxSOM* $k = 10$	$k = 20$	$k = 10\%$
diagnosis	11.3	23.3	30.9 (28)	32.9	46.8	55.9 (28)
neoplasms	38.2	46.2	48.1 (23)	47.2	63.7	67.5 (23)
organic chem.	60.7	71.0	79.3 (33)	73.1	80.0	81.4 (33)
pathol. sympt.	42.8	55.4	56.8 (21)	64.2	76.5	77.9 (21)
surgical op.	44.1	70.4	72.0 (21)	68.8	76.9	78.0 (21)
system dis.	26.1	38.9	50.0 (32)	53.5	71.2	79.6 (32)
therapeutics	23.9	39.2	40.9 (25)	52.5	69.0	73.4 (25)
virus diseases	27.5	48.9	30.5 (12)	58.0	72.5	59.5 (12)
disease cond.	8.0	14.0	55.3 (283)	21.6	34.7	80.0 (283)
treatment int.	3.9	5.9	38.3 (299)	11.2	20.0	57.0 (299)

Table 2: Results of *baseline* and *TaxSOM k-coverage* with three different values for k.

that the recommendations made by *TaxSOM* are "more correct" than those made by the *baseline* approach. Notice that the curves sometimes intersect with high values of k. In this case, however, the result is less interesting. In fact, the system is used to recommend the best classes, and the number of suggested classes should be as small as possible. Actually, in a real task, what we can expect from such a system is that for each CPG few classes are recommended as probable labeling classes for the given CPG.

In Table 2 are shown the results for three different situations: *(i)* a case where the system suggest a selection of 10 possible classes; *(ii)* a case where the system suggest a selection of 20 classes; *(iii)* a case where the system select the 10% most probable classes among all classes. For all the three cases and for all taxonomies it is always valuable to use *TaxSOM* driven by the prior knowledge encoded in the taxonomy than just using *baseline* which only uses the keywords that best represent the concepts.

In the table we also provided the results of the same type of analysis done for the original two NGC hierarchies. From these results it can be seen that the behavior of the models is (obviously) influenced by the absolute number of classes in the hierarchy. Nonetheless, *TaxSOM* is still better that the *baseline* approach. Moreover, looking at the results of the third case (i.e. $k = 10\%$) the model still give interesting results also for the two big hierarchies.

7 Conclusions and Future Work

In the paper we presented an approach for helping physicians to organize CPGs into hierarchies of concepts. The challenge was twofold: to avoid the need for labeled documents in advance and to exploit relational knowledge encoded by a taxonomy. Experimental evaluation on a collection of CPGs gave the empirical evidence of the potential benefit for physicians while using the proposed model.

References

[Adami *et al.*, 2003a] G. Adami, P. Avesani, and D. Sona. Bootstrapping for hierarchical document classification. In *Proc. of CIKM-03, 12th ACM Int. Conf. on Information and Knowledge Management*, pages 295–302. ACM Press, New York, US, 2003.

[Adami *et al.*, 2003b] G. Adami, P. Avesani, and D. Sona. Clustering documents in a web directory. In *Proc. of WIDM-03, 5th ACM Int. Workshop on Web Information and Data Management*, pages 66–73. ACM Press, New York, US, 2003.

[Ceci and Malerba, 2003] M. Ceci and D. Malerba. Hierarchical classification of html documents with webclassii. In *Proc. of the 25th European Conf. on Information Retrieval (ECIR'03)*, volume 2633 of *Lecture Notes in Computer Science*, pages 57–72, 2003.

[Chakrabarti *et al.*, 1997] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Using taxonomy, discriminants, and signatures for navigating in text databases. In M. Jarke, M.J. Carey, K.R. Dittrich, F.H. Lochovsky, P. Loucopoulos, and M. A. Jeusfeld, editors, *VLDB'97, Proc. of 23rd Int. Conf. on Very Large Data Bases*, pages 446–455. Morgan Kaufmann, 1997.

[Cheng *et al.*, 2001] C.H. Cheng, J. Tang, A.W.C. Fu, and I. King. Hierarchical classification of documents with error control. In *PAKDD 2001 - Proc. of 5th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, volume 2035 of *Lecture Notes in Computer Science*, pages 433–443, 2001.

[Doan *et al.*, 2003] H. Doan, P. Domingos, and A. Halevy. Learning to match the schemas of data sources: A multistrategy approach. *Machine Learning*, 50:279–301, 2003.

[Dumais and Chen, 2000] S. Dumais and H. Chen. Hierarchical classification of web document. In *Proc. of the 23rd ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR'00)*, 2000.

[Grimshaw and Russel, 1993] J.M. Grimshaw and I.T. Russel. Effect of clinical guidelines on medical practice: A systematic review of rigorous evaluations. *Lancet*, pages 1317–1322, 1993.

[Hersh and Greens, 1989] W.R. Hersh and R.A. Greens. Saphire - an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval and hierarchical relationships. *Computers and Biomedical Research*, 23:410–25, 1989.

[Hersh and Hickam, 1993] W.R. Hersh and D.H. Hickam. A comparison of two methods for indexing and retrieval from a full text medical database. *Medical Decision Making*, 13(3):220–26, 1993.

[Humphreys and Lindberg, 1993] B.L. Humphreys and D.A. Lindberg. The umls project: making the conceptual connection between users and the information they need. *Bull Med Libr Assoc*, 81(2):170–177, 1993.

[Joachims, 1998] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML '98)*, 1998.

[Jordan and Jacobs, 1994] M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.

[Kohonen, 2001] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Series in Information Sciences*. Springer, Berlin, 2001.

[Koller and Sahami, 1997] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In D.H. Fisher, editor, *ICML 1997, Proc of the 14th Int. Conf. on Machine Learning*, pages 170–178. Morgan Kaufmann, 1997.

[McCallum and Nigam, 1999] A. McCallum and K. Nigam. Text classification by bootstrapping with keywords, em and shrinkage. In *In ACL Workshop for Unsupervised Learning in NLP*, 1999.

[Moskovitch and Shahar, 2004] R. Moskovitch and Y. Shahar. Effective concept-search in hierarchical organized library. Technical Report ISE-TR-314/2004, Dept. of Information Systems Engineering, Ben Gurion University, 2004.

[Moskovitch *et al.*, 2004] R. Moskovitch, A. Hessing, and Y. Shahar. Vaidurya - a concept-based, context-sensitive search engine for clinical guidelines. In *Proc. of the joint conf. of AMIA04 and Medinfo-2004*, San Francisco, CA, US, 2004.

[Ruiz and Srinivasan, 2002] M.E. Ruiz and P. Srinivasan. Hierarchical text categorization using neural networks. *Information Retrieval*, 5(1):87–118, 2002.

[Sun and Lim, 2001] A. Sun and E.P. Lim. Hierarchical text classification and evaluation. In N. Cercone, T.Y. Lin, and X. Wu, editors, *ICDM 2001 - Proc. of the 2001 IEEE Int. Conf. on Data Mining*, pages 521–528. IEEE Computer Society, 2001.

[Wang *et al.*, 1999] K. Wang, S. Zhou, and S.C. Liew. Building hierarchical classifiers using class proximity. In *Proc. of the 25th VLDB Conference*, 1999.

[Weigend *et al.*, 1999] A.S. Weigend, E.D. Wiener, and J.O. Pedersen. Exploiting hierarchy in text categorization. *Information Retrieval*, pages 1(3) 193–216, 1999.

[Yang, 1994] Y. Yang. Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In *Proc. of the 17th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 13–22, 1994.

derived by using an Artificial Neural Network with Monte Carlo methods as described in section 3.

By replacing survival estimates with a scatter-plot of prognostic scores for each prognostic model (figure 2), we can examine whether a particular patient is borderline between cells in the matrix as the cross-tabulation matrix is placed over the cut-off points for prognostic group scores. In addition, a patient's TNM stage (a commonly used prognostic model) is highlighted by colour coding the data points, this shows the wide scatter of the TNM stage across the map, thus providing another level of insight to the clinician.

This information is also supported by empirically derived rules using the rule extraction method, OSRE, described in section 4, this informs the clinician about the treatment given to similar patient groups and presents the rules derived from the data for the treatment received by this group.

With all elements combined an intelligent interface is presented to the clinician, by expanding NPI into a matrix, maintaining their current knowledge of survival expectation and treatment allocation for patient groups while showing the difference additional information has on sub-groups of patients survival prognosis. It also informs on patient cases that may be borderline between prognostic groups. Additionally, it provides an individual prognosis of survival to 5 years with 95% confidence intervals and presents a Boolean expression of group characteristics for treatment derived from evidence in historical data.

6. Conclusion

An interface for breast oncology is proposed, which shows the value of additional covariates in discriminating patients by mortality risk. The interface starts from a currently used clinical index, NPI, and extends this to include a cross-matching matrix of grouped survival curves and the position a patient resides within the matrix, complemented with individual prognostic predictions qualified with predicted confidence intervals, additionally treatment allocation is explained by data-based rules. These data in combination add significantly to the discriminatory information currently available to clinicians about prognostic risk and allocation of adjuvant treatment.

This complex information is presented in a format designed to match the clinician's own reasoning. Further work is now required to evaluate the clinical acceptance of the proposed methodology.

7. Acknowledgement

This work was supported by EPSRC research grant and the European Network of Excellence BioPattern.

8. References

1. Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, Nicholson RI, Griffiths K, 'A prognostic index in primary breast cancer', Br. J. Cancer, 45: 3621, 1982.

2. Ravdin P.M., Siminoff L.A., Davis G.J. et al 'Computer Program to Assist in Making Decisions about Adjuvant Therapy for Women with Early Breast Cancer'. Journal of Clinical Oncology, 74 (4) 980-991; 2001.

3. Lisboa, P.J.G. 'A review of evidence of health benefit from artificial neural networks in medical intervention', Neural Networks, Invited Paper, 15(1), 9-37, 2002.

4. Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A.L., Sandercock, P., Spiegelhalter, D. and Tryer, P. 'Framework for design and evaluation of complex interventions to improve health'. *BMJ*, 321: 694-696, 2000. www.mrc.ac.uk/complex_packages.html

5. Etchells, T.A., Jarman, I.H., Lisboa, P.J.G., 'Empirically derived rules for adjuvant chemotherapy in breast cancer treatment' MEDSIP, Proc. 2nd International Conference, 345-351 , 2004,

6. Cox DR, 'Regression models and life tables', Journal of the Royal Statistical Society, B, 74: 187-220, 1972.

7. Collett D. 'Modelling Survival Data in Medical Research'. Chapman and Hall, 1994.

8. Lisboa, P.J.G., Wong, H., Harris, P. and Swindell, R. 'A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer'. Artificial Intelligence in Medicine. 28(1):1-25, May 2003.

9. Biganzoli E, Boracchi P, Mariani L, Marubini E, 'Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach', Statist. Med, 17, 1169-1186, 1998.

10. Bishop C.M. 'Neural network for pattern recognition'. Oxford; 1995: Clarendon Press.

11. Etchells, T.A. & Lisboa P.J.G. 'On rule extraction from smooth decision surfaces'. NWSMED/CIMED, Proc. 5th International Conference, pp 23-28, 2003.

12. Etchells, T.A., Rule Extraction from Neural Networks: A practical and efficient approach. Unpublished PhD thesis, 2003. http://www.cms.livjm.ac.uk/etchells/phd/etchells.pdf

13. http://www.nice.org.uk

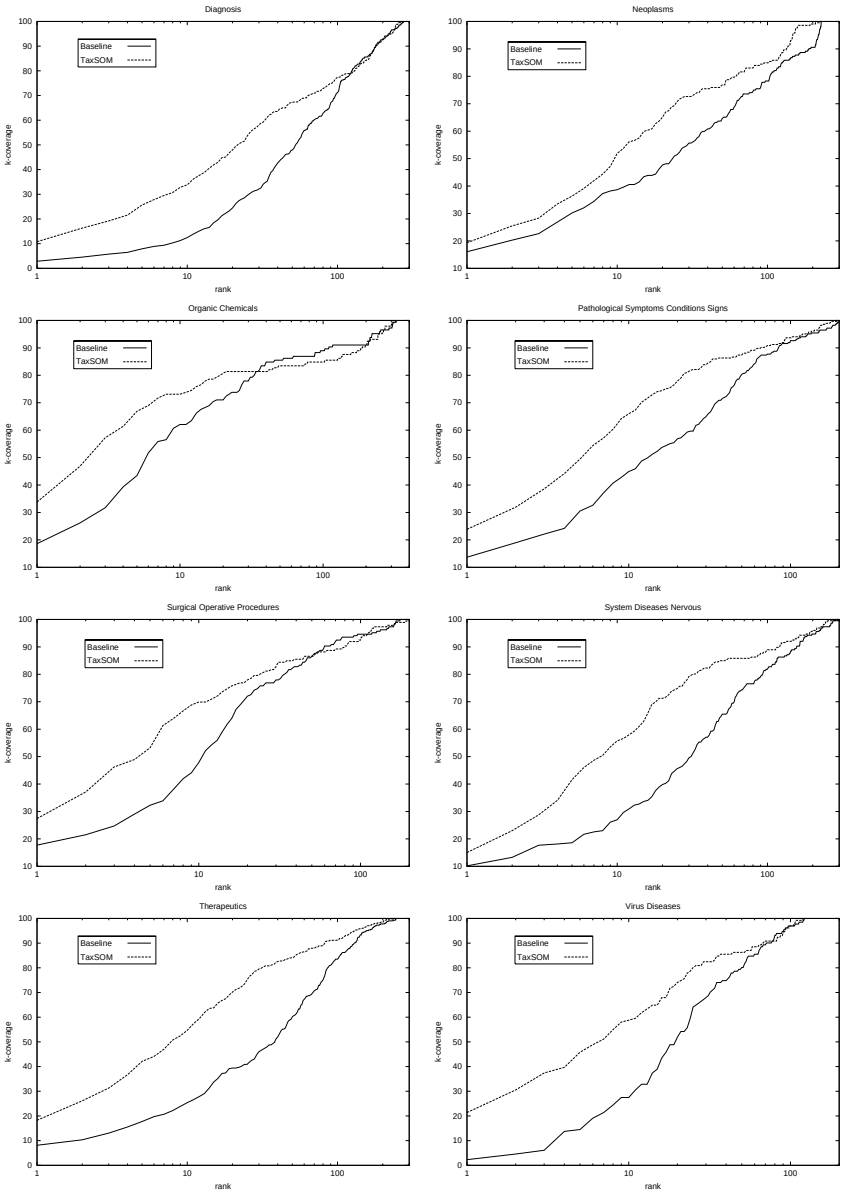


Figure 2: The graphs depict the *k-coverage precision* for all eight datasets.

Influence diagrams for medical decision problems: Some limitations and proposed solutions*

Manuel Luque and **Francisco Javier Diez**
Dept. Inteligencia Artificial. UNED
28040 Madrid. Spain
mluque@bec.uned.es, fjdiez@dia.uned.es

Carlos Disdier
Pulmonary Section
San Pedro Alcantara Hospital
1004 Caceres. Spain
cdisdier@separ.es

Abstract

When trying to solve two medical decision problems we have encountered several difficulties: how to represent and operate with decomposable utility functions, how to calibrate our human experts and explain them the "reasoning" of our influence diagrams, and how to deal with partially ordered decisions. This paper describes these difficulties and the solutions we have adopted.

1 Introduction

One of the medical problems we are currently working on is the mediastinal staging of non-small cell lung cancer. There are several tests available, such as computed tomography scan (CT scan), transbronchial needle aspiration (TBNA), mediastinoscopy (MED) and others, which have different characteristics of sensitivity, specificity, morbidity and mortality. The other problem on which we are working is the management of mild head injury.

Influence diagrams are a framework which serves as an effective modeling tool for decision problems. An influence diagram (ID) [3], consists of a directed acyclic graph having three kinds of nodes: decision (graphically represented by squares or rectangles), chance (circles or ovals), and utilities (diamonds). Each decision node represents to actions under the direct control of the decision maker. Each chance node represents a random variable. In medical IDs, utility nodes represent medical outcomes and costs (morbidity, mortality, economic cost...).

The quantitative information that defines an ID is given by assigning to each chance node X_i a probability distribution $p(X_i|pa(X_i))$, where $pa(X_i)$ represents the parents of the node X_i in the graph, and assigning to the utility node U a function $\psi(pa(U))$. The objective of the evaluation of an influence diagram is obtaining a policy for each decision, which prescribes a set of optimal actions for the decision maker. The policy for each decision is a function of the variables that are known when the decision is made.

*This research has been supported by the Spanish Ministry of Science and Technology under grant TIC-2001-2973-C05-04.

2 Limitations of influence diagrams for medical decision problems

This section describes the difficulties we have found when building those IDs and how we have extended Elvira to cope with them. Elvira[1] is a Java tool to construct probabilistic decision support systems. Elvira works with Bayesian networks (BN) and influence diagrams and it can operate with discrete and continuous variables. It has an easy Graphical User Interface (GUI) for constructing BNs and IDs.

2.1 Decomposable utilities

An essential component of an influence diagram is the utility function. In its original formulation [3], each ID had only one utility node, which entails several disadvantages. First, the human expert has to assess more parameters. Second, the bigger the utility function the more time and memory space is required for the computational evaluation of the ID. Third, policies tend to include more variables than when using decomposable utility functions.

In order to overcome these shortcomings, Tatman and Shachter [5] introduced a new kind of utility node, called super-value nodes, which represent a function of their parents' utilities, and proposed an algorithm for evaluating such generalized IDs.

The first limitation we encountered when building our medical IDs is that current software tools do not admit super-value nodes. At most, they accept several utility nodes under the assumption that the global utility is the sum of all of them. For this reason, we extended Elvira's GUI and format so that it could cope with super-value nodes. Figure 1 shows the current version of our ID for mediastinal staging. The three rectangles represent the decisions: one of them represents the decision of performing a TBNA or not, the second represents the decision about performing a mediastinoscopy, and the third represents the decision about the treatment, which can be thoracotomy, radiotherapy, chemotherapy, or palliative care. Rounded rectangles represent chance variables: one is the main diagnosis (N2-N3), three are tests, and the forth indicates whether the patient survives the mediastinoscopy. At the bottom, there are seven utility nodes; two of them are super-value nodes, which indicates a decomposition of the utility function.

[1] See `http://www.ia.uned.es/~elvira` and [1].

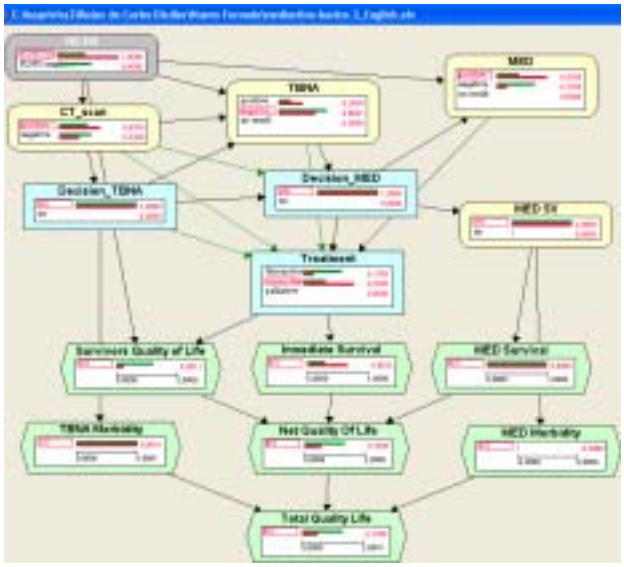


Figure 1: Explanation for a medical decision problem in Elvira

Furthermore, we realized that the algorithm of Tatman and Shachter was unsatisfactory because it is based on arc reversal, an operation that involves inefficient divisions of potentials and often introduces unnecessary variables in the resulting policies. For this reason, we developed a new algorithm which is in general more efficient and does not tend to introduce so many unnecessary variables in the policies [4].

2.2 Explanation in influence diagrams

One of the key factors for the acceptance of expert systems in real world domains and especially in medicine is the capability to explain their reasoning. For this reason we have extended Elvira's explanation facilities from Bayesian networks to IDs. In addition to showing the resulting policies by opening a window for each decision, Elvira can also display numerical and graphical information inside each node (see Figure 1): horizontal bars inside the nodes represent the probability that a chance variable takes on a certain value, the probability that the decision maker chooses a certain action for a decision,[2] or the expected utility for a utility node.

In Elvira it is possible to assign values to chance and decision variables in the same way as evidence is assigned to the corresponding nodes of a Bayesian network. It is also possible to show several "evidence cases", i.e., the probabilities and utilities for several subpopulations. For instance, Figure 1 shows two horizontal bars for each value of a chance or decision variable and for each utility node, thus comparing the situation in which we know the patient belongs to the N2-N3 positive group with the situation in which we have not any previous information. Please note that the node *Total Expected Utility* shows the global utility, while other utility nodes, such as *Survivors Quality of Life*, represent partial utilities.

[2]Policies are deterministic functions, but it makes sense to speak of the probability of an action because decisions are based on chance variables.

2.3 Order of decisions

The traditional definition of the influence diagrams assume that there is an order between decisions. However, in some medical problems the question is just which tests should be performed and in what order. For example, in the mediastinal staging problem, we may wonder what is the best order among the three tests, CT scan, MED and TBNA.

Several representations of decision problems have been proposed in order to let have a partial order between decisions (see [2] and references therein). The task of finding the best order is performed by the evaluation algorithm. The aim of these representations is representing and solving asymmetric decision problems. However, they tend to obscure the structure of the medical decision problems, because they are too general and do not consider the specific characteristics of the medical decision problems. This makes more difficult any kind of explanation of the reasoning and less efficient the evaluation of the decision problem.

For this reason we are currently exploring new representational schemes that will lead to more simple and intuitive influence diagrams, and in turn would require new algorithms for their evaluation.

3 Conclusion

In this paper we have discussed the shortcomings of influence diagrams for solving real-world problems in medicine, as well as the limitations of some of the current algorithms and software packages. We have also shown some of our current efforts aimed at having more efficient representation schemes, algorithms and software tools.

References

[1] The Elvira Consortium. Elvira: An environment for creating and using probabilistic graphical models. In *Proceedings of the First European Workshop on Probabilistic Graphical Models (PGM'02)*, pages 1–11, Cuenca, Spain, 2002.

[2] T. D. Nielsen F. V. Jensen and P. Shenoy. Sequential influence diagrams: A unified asymmetry framework. In *Proceedings of the Second European Workshop on Probabilistic Graphical Models*, pages 121–128. P. Lucas (eds.), 2004.

[3] R. A. Howard and J. E. Matheson. Influence diagrams. In R. A. Howard and J. E. Matheson, editors, *Readings on the Principles and Applications of Decision Analysis*, pages 719–762. Strategic Decisions Group, Menlo Park, CA, 1984.

[4] M. Luque and F. J. Diez. Variable elimination for influence diagrams with super-value nodes. In *Proceedings of the Second European Workshop on Probabilistic Graphical Models*, pages 145–152. P. Lucas (eds.), 2004.

[5] J. A. Tatman and R. D. Shachter. Dynamic programming and influence diagrams. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(2):365–379, 1990.