5^{th} International Workshop on Intelligent Data Analysis in Medicine and Pharmacology

(IDAMAP-2000)

A workshop at
the 14^{th} European Conference on Artificial Intelligence
(ECAI-2000)

Berlin, Germany, August 2000

Proceedings

Editors:

Nada Lavrač
Silvia Miksch
Branko Kavšek

Foreword

In all human activities, automatic data collection pushes towards the development of tools able to handle and analyze data in a computer-supported fashion. In the majority of the application areas, this task cannot be accomplished without using the available knowledge on the domain or on the data analysis process. This need becomes essential in biomedical applications, since medical decision-making needs to be supported by arguments based on basic medical and pharmacological knowledge.

Intelligent data analysis (IDA) methods support information extraction from data by potentially exploiting domain knowledge. Most common techniques include data mining, machine learning, temporal data abstraction, information visualization, case-based reasoning, statistical methods and combination thereof. To increase the chances of utilizing these methods within clinical practice, the intelligent data analysis process typically gains from interaction by medical experts in all its phases, from data gathering and cleaning to evaluation and exploitation of the results.

The objective of Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP) activities is to foster the development, adaptation or re-use of existing IDA methods to cope with real medical tasks. The ultimate goal of the people working on IDAMAP is the successful integration and employment of these methods in modern hospital and other information systems. The IDAMAP community is trying to increase the awareness and acceptance of IDA methods in medicine through papers that show their successful application.

The 16 (long and short) scientific papers included in the proceedings were selected after a detailed review by two members of the Program Committee.

History of the IDAMAP workshops

This is the fifth international workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2000) held as a one day workshop at the 14^{th} European Conference on Artificial Intelligence (ECAI-2000) in Berlin, Germany, August 2000.

`url: http://www.ifs.tuwien.ac.at/~silvia/idamap2000`

The former IDAMAP workshops were as follows:

- The **first** Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-96) at the European Conference on Artificial Intelligence, 1996 (ECAI-96), in Budapest, Hungary, August 1996.

  `url: http://www-ai.ijs.si/ailab/activities/idamap96.html`

- The **second** Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-97) at the International Joint Conference on Artificial Intelligence, 1997 (IJCAI-97), in Nagoya, Japan, August 1997.

  `url: http://www-ai.ijs.si/ailab/activities/idamap97.html`

- The **third** Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-98), at the European Conference on Artificial Intelligence, 1998 (ECAI-98), in Brighton, UK, August 1998.

  `url: http://aim.unipv.it/~ric/idamap98`

- The **fourth** Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-99), at the American Medical Informatics Association (AMIA) 1999 Annual Symposium, in Washington, DC, USA, November 1999.

  `url: http://www.ifs.tuwien.ac.at/~silvia/idamap99`

Acknowledgments

The IDAMAP-2000 workshop was organized under the umbrella of ECAI-2000, the 14^{th} European Conference on Artificial Intelligence. We are grateful to ECCAI and ECAI-2000 organizers for their support in the organization of this scientific event.

We wish to thank the researchers for submitting their papers to IDAMAP-2000 and the Program Committee members for their thorough reviews.

The work of the editors was supported by the Slovenian Ministry of Research and Technology, the EU funded project IST-1999-11495 Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprize, and the Institute of Software Technology (Vienna University of Technology).

Ljubljana — Nada Lavrač
Vienna — Silvia Miksch
June 2000

Editors of the IDAMAP-2000 Proceedings

- Nada Lavrač, J. Stefan Institute, Ljubljana, Slovenia
- Silvia Miksch, Vienna University of Technology, Vienna, Austria
- Branko Kavšek, J. Stefan Institute, Ljubljana, Slovenia

IDAMAP-2000 Program Chairs

- Nada Lavrač, J. Stefan Institute, Ljubljana, Slovenia
- Silvia Miksch, Vienna University of Technology, Vienna, Austria

IDAMAP-2000 Program Committee:

- Sarabjot Anand, University of Ulster, Newtownabbey, Northern Ireland
- Steen Andreassen, Aalborg University, Aalborg, Denmark
- Lars Asker, Stockholm University and Royal Institute of Technology, Stockholm, Sweden
- Riccardo Bellazzi, University of Pavia, Pavia, Italy
- Werner Horn, Austrian Research Institute for Artificial Intelligence, Vienna, Austria
- Elpida Keravnou, University of Cyprus, Nicosia, Cyprus
- Cristiana Larizza, University of Pavia, Pavia, Italy
- Nada Lavrač, J. Stefan Institute, Ljubljana, Slovenia
- Xiaohui Liu, Birkbeck College, University of London, London, UK
- Silvia Miksch, Vienna University of Technology, Vienna, Austria
- Christian Popow, Department of Pediatrics, University of Vienna, Austria
- Yuval Shahar, Stanford University, Stanford, USA
- Blaž Zupan, University of Ljubljana, Ljubljana, Slovenia

Contents

Qualitative and Fuzzy Reasoning for identifying non-linear physiological systems: an application to intracellular thiamine kinetics

R. Bellazzi[1] and **R. Guglielmann, L. Ironi** [2]

Abstract.
The meaningful description of the behavior of complex dynamic systems through mathematical models it is always related to the identification of a model parameters set. This optimization procedure, called dynamic system identification (SI) may be really problematic for those application domains, such as the medical/physiological one, of which either the available knowledge is incomplete or the observed data are poor in number and in quality. This paper deals with the application of an hybrid method, which builds a fuzzy system identifier upon a qualitative structural model, to solve identification problems of the intracellular kinetics of Thiamine (vitamin B_1). The model obtained is robust enough to be used as a simulator, and then to provide physiologists with a deeper understanding of the Thiamine metabolism in the cells.

1 Introduction

A structural model of the the dynamics of complex real-world systems is a set of mathematical equations that meaningfully describes the system behavior. The equations are related to the *physical structure* of the domain; therefore such model offers potential benefits to the deep comprehension of the system at study as well as to the performance of certain tasks. If we focus our attention on physiology and medicine, such models may allow for the calculation of physiological quantities that can not be directly measured and may also allow the physiologist to formulate hypotheses dealing with the physiological and biochemical structure of the system, help the clinician to formulate and test diagnostic hypotheses and finally to plan therapeutical treatments. Unfortunately, the formulation of such structural models may be hampered by the incompleteness of the available knowledge of the underlying nonlinear dynamics. Moreover, the identification of model parameters might be impossible, due to a complex experimental design or to a limited number of available data. In such cases, the system dynamics is often studied under the hypothesis that minimal perturbations affect the system, that is under the linearity assumption. Although the resulting model captures limited aspects of the system dynamics, it may give useful information; nevertheless, also the linear formulation may be prohibitive as identifiability problems may occur.

In theory, a valid alternative to structural modeling, although potentially less informative, could be represented by non-parametric black-box modeling approaches to SI [14, 16, 22]. But, in practice, such models, which learn the nonlinear dynamics of the system from input-output data, result to be very inefficient and not robust when the available experimental data are poor either in number or in quality. Such a situation is not rare in the fields of physiology and medicine.

Motivated by these considerations, we started a project which aims at the design and implementation of an efficient and robust method capable to make the most of both the available structural knowledge and the observed data. The method, that we call FS-QM, is domain-independent and results from the integration of qualitative models, namely QSIM [17] models, and fuzzy systems [4, 2]. As both frameworks have been introduced to cope with the complexity of real-world systems, their combination should benefit from the analytical power of the former one as well as from the approximation properties of the latter.

In outline, the method exploits the incomplete structural knowledge to build a QSIM model of the system dynamics, and then it infers, through simulation, all of its possible behaviors. The set of behaviors is mapped, in accordance with the a priori expert knowledge, into a fuzzy rule-base, where each rule may be seen as a measure of the possible transition from states to the next ones. The mathematical interpretation of such a rule-base properly defines and initializes a nonlinear functional approximator, which is then tuned to the experimental data.

The emphasis of this paper is rather on applicative aspects than on methodological issues. We discuss the identification problems which arise from modeling a real-world system in the physiological domain, the intracellular thiamine kinetics, and the solutions given by the application of our method [4, 2]. The comparison of our results with those obtained by means of a traditional application of fuzzy systems to SI [22] highlights the good performance of our method when applied to derive a simulator of the thiamine kinetics in the intestine cells. The significant improvement in terms of efficiency and robustness of FS-QM over traditional methods is due to the good initialization of both the structure of the fuzzy identifier and its parameters built by encoding the system dynamics captured by its qualitative behaviors [3].

For the sake of completeness, let us remark that the idea of exploiting QR techniques for SI is not new. Most of the work done addresses the problem of the automation of the traditional process of SI, that is the automation of both structural identification and the choice of the most appropriate numerical techniques for parameter estimation and their initialization [6, 5, 11, 7, 8, 12]. Another piece of work deals with a method for SI capable to deal with states of incomplete knowledge [15] in which both the candidate model space and the stream of observations are defined semi-quantitatively. What distinguishes this

[1] Dipartimento di Informatica e Sistemistica - Università di Pavia, Pavia, Italy, e-mail:ric@aim.unipv.it
[2] Istituto di Analisi Numerica - C.N.R., Pavia, Italy

piece of work from the other ones is its capability to deal with system characterized by both incomplete structural knowledge and poor stream of data.

2 Modeling problems in the physiological/medical domain

The application of mathematical modeling techniques to the study of a wide spectrum of metabolic and endocrine processes has been largely described in the literature [9]. A metabolic system may be essentially viewed as a system of chemical reactions and transport processes controlled by substances produced by the endocrine system. The description of the dynamics of such systems, even in the most simple cases, is a really complex task, and it has been made tractable by the compartmental modeling methodology [1, 13]. Within this framework, a system is decomposed into a finite set of subsystems, called *compartments*, and the compartments interact either with each others or with the environment [3] by exchanging material.

A compartment is fundamentally an idealized store of a substance, which may often be adequately assumed homogeneously distributed. The transfer of material through the system that occurs by physical transport or chemical reactions is represented as transfer from one compartment to another. The model equations are expressed by Ordinary Differential Equations (ODE) in terms of the state variables of the system, denoted by $x_i(t)$, that represent the concentration or amount of substance in the $i-$th compartment which exchanges matter with other compartments at time t. Then, the rate of change of each $x_i(t)$ is based on the mass balance law:

$$\dot{x}_i \;=\; f_{i0} \;+\; \sum_{\substack{j=1\\ j\neq i}}^{n} f_{ij}(x_j) \;-\; \sum_{\substack{j=1\\ j\neq i}}^{n} f_{ji}(x_i) \;-\; f_{0i}(x_i) \qquad (1)$$

where $\dot{x}_i$ denotes the time derivative of x_i; f_{ij} denotes the rate of mass transfer into the $i-th$ compartment from the $j-th$ compartment. In general, the transfer of material depends on the quantity or concentration of material in the source compartment and may also be dependent on the quantity or concentration in some other compartments, that is:

$$f_{ij} \;=\; f_{ij}(x_j; x_l, x_m, ...) \qquad (2)$$

where x_j denotes the state variable of the source compartment, whereas $x_l,\ x_m, ..$ indicate the variables controlling f_{ij}.

The mathematical model of a compartmental structure then consists of a set of ODE's which are fully defined when the functional relations (2) are explicitly stated. Mostly, given the complexity of the processes dealt with, such relations are naturally nonlinear, and their definition may very often be intractable due to the incompleteness of the available knowledge. However, for systems intrinsically nonlinear, a linearity assumption ($f_{ij}(x_j) = k_{ij}x_j$) may be reasonably adopted when the observed dynamics is obtained in response to a small-signal perturbation around the system steady-state condition produced by the administration of a tracer material.

The next step in the system identification process deals with the estimation of the unknown parameters from data. Also in the linear case, this step may be critical if the *a priori* identifiability condition is not satisfied, that is, if from the ideal data that the experiment would generate it is not possible to determine uniquely the theoretical estimate of the unknown parameters. However, as real data are not noise-free, theoretical identifiability does not guarantee that the estimation results are accurate enough to identify a good model of the system dynamics, i.e. *a posteriori identifiability*. A model can be considered valid, and then give useful information if the identifiability conditions are satisfied. Methods for testing both a priori and a posteriori identifiability are discussed in the literature [10, 18].

[3] indicated as compartment 0

3 The intracellular thiamine kinetics: Identification problems and solutions

Thiamine (Th), also known as vitamin B_1, is one of the basic micronutrients present in food and essential for health. In particular, Th is contained in dried yeast, meat, nuts, legumes and potatoes. Within the cells, Th participates in the carbohydrate metabolism, in the central and peripheral nerve cell function and in the myocardial function. Deficiency of Th causes beriberi with peripheral neurologic, cerebral and cardiovascular manifestations [21].

More in detail, after its absorption in the intestinal mucosa, Th is released into plasma for the distribution to the other tissues, either in its original chemical form (Th) or in a mono-phosphorilated one (ThMP). Th is transported through the cell membrane by means of an enzyme-mediated mechanism, and is then directly transformed into a higher energy compound, Thiamine Piro-Phosphate (ThPP); ThPP is dephosphorylated into ThMP, and it is in equilibrium with Thiamine Tri-Phosphate (ThTP). ThPP is the active element that participates in the carbohydrate metabolism. The chemical transformations occurring within the cells are described in Fig. 1.

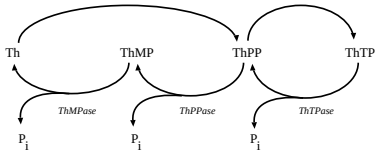


Figure 1. The chemical pathway of Th within cells. Th transforms into ThPP; ThPP transforms into ThMP, that is transformed back into Th. ThPP also transforms in a reversible way into ThTP.

3.1 Identification of the structural model

Since early 80's several studies have been carried out to quantitatively assess the Th metabolism in the cells [19, 20]. All these studies were performed on rats, and had the basic goal to quantitatively define the normal and pathological conditions underlying Th chemical transformations and cellular uptake and release. Since the Th metabolism is intrinsically nonlinear, the first exploratory approach to its quantitative characterization consists in its analysis around the steady state conditions. Therefore, from an experimental viewpoint, all these studies were based on tracer experiments, in which a small amount of labeled (radio-active) Th was injected in plasma or in peritoneum; the specific activity (radioactivity per gram) of labeled Th was subsequently measured in plasma and in the cells. From a modeling viewpoint, a linear compartmental model has been used to study the Th kinetics in several organ tissues, with particular reference to the nervous ones. Let us observe that the ThTP form can be neglected in the model. As a matter of fact, the fast chemical pathway between ThPP and ThTP and the relatively low concentration of ThTP allows us to consider ThTP in equilibrium with ThPP. Then, the model, whose structure is shown in Fig. 2, is described by the following ODE's:

$$\dot{x}_1 = k_{14}x_4 - (k_{01} + k_{21})x_1 \quad (3)$$
$$\dot{x}_2 = k_{21}x_1 - k_{32}x_2 \quad (4)$$
$$\dot{x}_3 = k_{32}x_2 + k_{35}x_5 - (k_{03} + k_{13})x_3 \quad (5)$$
$$\dot{x}_4 = k_{40}u_1 - k_{14}x_4 \quad (6)$$
$$\dot{x}_5 = k_{50}u_2 - k_{35}x_5 \quad (7)$$

where x_1 is the intracellular Th, x_2 is the intracellular ThPP, x_3 is the intracellular ThMP, x_4 is the quantity of Th in the cell membrane while x_5 is the quantity of ThMP in the cell membrane; u_1 is the plasmatic Th and u_2 is the plasmatic ThMP. Finally, the parameters k_{ij} are the transfer coefficients to be estimated from data. As a matter of fact, compartments denoted by 4 and 5 are fictitious as they do not correspond to any chemical form of Th, but they are just used to model the absorption process of Th in the cells. The model (3-7) proved to satisfy a priori identifiability conditions when a bolus injection in plasma is delivered.

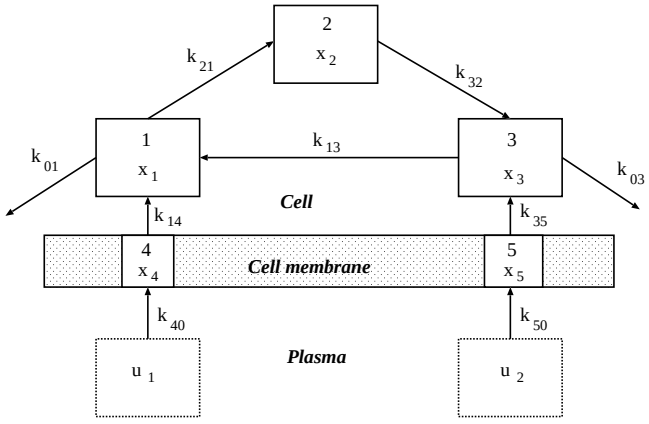


Figure 2. The compartmental model of the Th metabolism within cells. The inputs of the system are the quantity in plasma of Th and ThMP (measured as the concentrations u_1 and u_2, respectively). The flows between the quantity x_1 (Th), x_2 (ThPP) and x_3 reflects the chemical pathway described Figure 1

The same model and the same experimental setting were applied to study the intestine tissue metabolism in normal subjects and in subjects suffering from diabetes (one of the main disfunctions of carbohydrate metabolism), both treated and non treated rats. In this case, the main purpose of the study was to quantitatively evaluate the differences in the transfer constants and turnover rates in the three different classes of subjects; on the basis of this evaluation it would also be possible to understand if insulin treatment is able to re-establish quasi-normal conditions in Th metabolism. Unfortunately, the compartmental model identification was unsuccessful, even using different optimization techniques, ranging from standard nonlinear estimation procedures to Bayesian ones. This results in an a posteriori unidentifiability of the model.

The main reason of this failure may be explained by the intrinsical problems related to the experimental setting: as mentioned before, the intracellular labeled Th is measured after a bolus in plasma. However, the physiological Th pathway in the intestine tissue presents a Th *absorption* way directly from the intestinal mucosa and a subsequent *release* into the plasma tissue. On the contrary, it is completely unknown how the Th quantity is physiologically absorbed by intestine cells from plasma, and also how such absorption is regulated. Therefore, the linearity assumption for the transport process from plasma into cells results to be completely inadequate.

This problem hampers the use of compartmental modeling techniques for analyzing the data, and, at a first glance, the use of the data themselves. This is particularly dramatic for this kind of experiments: at each sampling time four rats are sacrificed, and a single measurement is derived as the mean of the four subjects. The efforts and costs of the experimental setting motivate the exploitation of other techniques for data modeling.

3.2 The need for a novel approach

An alternative solution to structural identification is to resort to the so-called "non-parametric" modeling methods. This term is somehow misleading, since the models are always characterized by a set of equations and parameters; however, such parameters do not have a precise physical meaning, and this gives the reason for the "non-parametric" wording. The non-parametric methods aim at reproducing the functional relationships between the observable variables only on the basis of the available data, without requiring knowledge on the physiological system at hand. In our case, due to the complexity of the problem, a natural choice is to exploit nonlinear dynamic discrete models, known as Nonlinear AutoRegressive models with eXogenous inputs (NARX). In such a framework the system dynamics of an output variable y is described by the input-output equation[4]:

$$y_{k+1} = f(y_k, \ldots, y_{k-l}, \underline{u}_k, \ldots, \underline{u}_{k-h}) \quad (8)$$

where $\underline{u} \in \Re^{n-1}$ and $y \in \Re$ are discrete-time sequences, l and h are known delays, and the function $f(\cdot)$ is in general unknown. Assuming $l = h = 0$, equation (8) may be written as follows:

$$y_{k+1} = f(\underline{x}_k), \quad \text{where} \quad \underline{x}_k = \{y_k, \underline{u}_k\}$$

In this context, methods recently proposed to find a function approximator of f are Feed-forward Neural Networks, Radial Basis Functions, Wavelet Functions, Fuzzy Systems. However, to build f with the desired accuracy only from the observations, all these approximation schemes usually require sufficiently large data sets .

As far as our application is concerned, such schemes cannot be applied, since no more than 17 measurements are available for each Th chemical form. Moreover, the identification problem is a nonlinear one: the treatment of nonlinear problems is not straightforward and demands some prior information to properly state a reasonable initial guess on the parameter values, and then to get convergence to their optimal estimate. The methods mentioned above, except Fuzzy Systems (FS) are not capable to embed prior knowledge, and therefore the initial values of the parameters are randomly chosen.

In this setting, the adoption of a non-parametric model able to exploit also the available structural knowledge seems the natural solution for effectively coping with the problems mentioned above. As a matter of fact, FS's are able to embed the a priori knowledge of the domain under the form of inferential linguistic information, called Fuzzy Rules (FR), but in practice, the information in the linguistic form about a complex system is often poor or unavailable, and then the function f is usually inferred only from the data.

This paper deals with the application of an hybrid method [3], based on the integration of QSIM models [17] and FS's [22], called

[4] Without loss of generality we consider here nonlinear Multiple Input-Single Output systems

FS-QM, which builds a fuzzy identifier upon the available a priori knowledge. The idea underlying our method is simple: the set of behaviors $\{B_1, .., B_m\}$ generated by the simulation of a QSIM model of the system at study is mapped into M FR's which, as a whole, capture the structural and behavioral knowledge of the system dynamics. As a matter of fact, such mapping is possible whenever the available knowledge allows us to define a bijective mapping between the quantity-space Q_L, in the QSIM representation, and the fuzzy-quantity space Q_F, whose elements are fuzzy sets. In outline, the main steps of the method are sketched in Fig. 3.

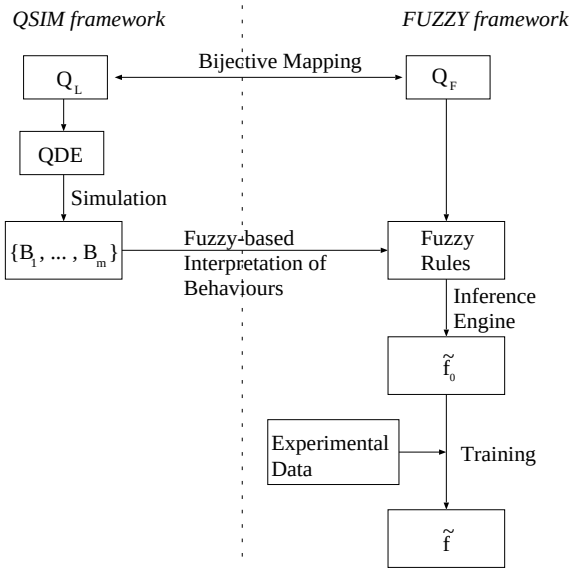


Figure 3. Main step of FS-QM.

The mathematical interpretation of the generated rules, through suitable fuzzy operators, such as the *singleton fuzzifier*, the *product inference rule*, the *center average defuzzifier*, and the characterization of fuzzy sets by *Gaussian membership functions*, allows us to initialize the approximator $\tilde{f}_0$ of f:

$$\tilde{f}_0(\underline{x}) = \frac{\sum_{j=1}^{M} \hat{y}^j [\prod_{i=1}^{n} exp(-(\frac{x_i - \hat{x}_i^j}{\sigma^j_i})^2)]}{\sum_{j=1}^{M} [\prod_{i=1}^{n} exp(-(\frac{x_i - \hat{x}_i^j}{\sigma_i^j})^2)]} \quad (9)$$

where: $\{\hat{x}_i^j\}$ and $\{\sigma_i^j\}$ are the parameters characterizing the Gaussian membership function which is related to the input variable x_i and appears in the j-th rule, $\hat{y}^j$ is the point where the membership function of the output, or equivalently of the consequent, in the j-th rule reaches its maximum value. Such an expression allows us to interpret the nonlinear function approximation problem with a FS as the process of tuning on a set of data the vector of parameters $\underline{\theta} = \{\underline{\hat{y}}, \underline{\hat{x}}, \underline{\sigma}\}$, initialized by the vector $\underline{\theta}_0$ in equation (9). The approximator derived in equation (9) is known to possess the universal approximation property, i.e. the capability of approximating any continuous function with an arbitrary degree of accuracy [22].

4 A non-parametric model of Thiamine kinetics

Although a complete knowledge on the mechanism of Th transport in the intestine cells from plasma is not known, the overall structure of the model in Fig. 2 still remains valid: the fluxes and the compartments in plasma and in the cells reflect the available information on the system. On the contrary, the number of compartments that model the membrane and the functional relationships describing the cellular absorption are not completely known. Therefore, we can ignore the compartments 4 and 5, and consequently the equations (6-7), and directly model the plasmatic Th absorption process.

As data sets for all the state variables are available, a non-parametric model of the overall system can be obtained (i) by splitting it into three decoupled subsystems, related to the three Th chemical forms (Th, ThPP and ThMP) obtained in response to the tracer input signals, namely plasmatic Th (u_1) and ThMP (u_2), and (ii) by formulating a NARX model for each of them.

Such models can be written as follows:

$$x_{1_{k+1}} = f_1(x_{1_k}, x_{3_k}, u_{1_k}) \quad (10)$$
$$x_{2_{k+1}} = f_2(x_{2_k}, x_{1_k}) \quad (11)$$
$$x_{3_{k+1}} = f_3(x_{3_k}, x_{2_k}, u_{2_k}) \quad (12)$$

The first step consists in the identification of $f_1,\ f_2,\ f_3$ in normal subjects. Our final goal deals with the construction of a *simulator* of the overall nonlinear intracellular Th kinetics. Such a simulator will allow us to understand the discrepancies in the Th metabolism between the different classes of subjects, namely normal, diabetic either treated or not, by comparing the results obtained by the simulator against the actual data.

4.1 Construction of the fuzzy identifiers

The construction of each f_i proceeds as sketched in Fig. 3, and starts with the construction of the QSIM models of each decoupled subsystem. Each model is described by a single Qualitative Differential Equation (QDE).

1 - *Th subsystem*: The Th dynamics is described by the QDE:

$$\dot{x}_1 = S^+(u_1) + M^+(x_3) - M^+(x_1) \quad (13)$$

where:
– S^+ and M^+ have the usual QSIM meaning, i.e. saturation and monotonicity (respectively);
– $S^+(u_1)$ models the nonlinear absorption process which governs the transfer of Th from plasma. The saturable functional relation is justified by the limited quantity of the mediating enzyme in the time unit;
– $M^+(x_3)$ models the chemical reaction of ThMP into Th. Let us observe that x_3 is modeled as a triangular shaped function: this modeling assumption is based on the knowledge of the tracer qualitative behavior in the cells.
– $M^+(x_1)$ models the chemical transformation of Th into ThPP.

2 - *ThPP subsystem*: The dynamics of ThPP is modeled by:

$$\dot{x}_2 = M^+(x_1) - M^+(x_2) \quad (14)$$

where:
– $M^+(x_2)$ models the chemical transformation of ThPP into ThMP. x_1 is analogously modeled as x_3, and $M^+(x_1)$ has the same meaning as above.

3 - *ThMP subsystem*: The equation modelling the dynamics of ThMP is:

$$\dot{x}_3 = S^+(u_2) + M^+(x_2) - M^+(x_3) \quad (15)$$

The functional constraints are analogously defined as in the other subsystems.

The input-output variables in (10-12) assume values in R^+, and their qualitative representations in both QSIM and FS frameworks are defined by their respective Q_L's and Q_F's. Table 1 summarizes the Q_L's and Q_F's of each x_i and u_i, and highlights the one-to-one correspondence between each Q_L and the respective Q_F. Let us observe that the elements of Q_F are represented in the linguistic form as well as through the values of the parameters which characterize the related membership functions. In our context, such parameters are the mean values ($\hat{x}$) and standard deviations (σ) and have been derived on the basis of the available physiological knowledge.

[ht]

Variables	Q_L	Q_F		
			$\hat{x}$ (nCi/g)	σ (nCi/g)
x_1	0	Low	0	13
	$(0\ Th^*)$	Medium	30	13
	Th^*	High	60	13
x_2	0	Low	5	30
	$(0\ ThPP^*)$	Medium	80	30
	$ThPP^*$	High	165	35
x_3	0	Low	0	22
	$(0\ ThMP^*)$	Medium	50	20
	$ThMP^*$	High	130	44
u_1	0	Low	20	400
	$(0\ \ U1S)$	Medium	1000	400
	$U1S$	High	2000	400
	$(U1S\ \ inf)$	Very High	3000	400
u_2	0	Low	70	140
	$(0\ \ U2S)$	Medium	330	70
	$U2S$	High	470	50
	$(U2S\ \ inf)$	Very High	600	60

Table 1. Mapping between Q_L and Q_F related to each variable. The last two columns report, respectively, the values of $\hat{x}$ and σ.

Since the data used for SI come from tracer experiments, each subsystem is simulated starting from $x_i(0) = 0,\ i = 1, 2, 3$. For the same reason, among all of the generated behaviors we consider only those that reach the system quiescent state. The translation of the generated Quiescent Qualitative Behaviors (QQB) into fuzzy rules is preceded by their analysis with the aim of (i) aggregating those behaviors that do not present any differences with respect to the variables of interest, (ii) filtering those behaviors which are inconsistent with physiological constraints not explicitly embedded in the model. The remaining Admissible Behaviors (AQB) are automatically mapped into FR's.

Table 2 summarizes, for each model, the number of QQB's, of AQB's, and of the generated IF-THEN rules.

Let us remark that the set of AQB's does not include spurious behaviors, which, on the other hand, would have been easily filtered on the basis of the a priori knowledge of the admissible experimental profiles. Generally, the absence of any spurious behavior in the AQB set is not guaranteed: in such a case, a reduction in FS-QM efficiency might be caused.

[ht]

Subsystem	# QQB's	# AQB's	# FR's
1	20	2	11
2	6	6	9
3	42	7	12

Table 2. Results of the qualitative simulation of the 3 models, in terms of the number of the generated QQB's, and of the AQB's. The number of the generated IF-THEN rules from the translation of the AQB's into the fuzzy framework is also reported.

The mathematical interpretation of each set of rules, in accordance with the choices underlying equation (9), allows us to derive a good initialization of each approximator $\tilde{f}_{i_0}$, and then the system is described by:

$$\begin{aligned}
\tilde{f}_{1_0}(x_1, x_3, u_1) &= \frac{\sum_{j=1}^{11} \hat{X}_1^j [e^{-(\frac{x_1-\hat{x}_1^j}{\sigma_1^j})^2} e^{-(\frac{x_3-\hat{x}_3^j}{\sigma_3^j})^2} e^{-(\frac{u_1-\hat{u}_1^j}{\sigma_{u_1}^j})^2}]}{\sum_{j=1}^{11} [e^{-(\frac{x_1-\hat{x}_1^j}{\sigma_1^j})^2} e^{-(\frac{x_3-\hat{x}_3^j}{\sigma_3^j})^2} e^{-(\frac{u_1-\hat{u}_1^j}{\sigma_{u_1}^j})^2}]} \\
\tilde{f}_{2_0}(x_2, x_1) &= \frac{\sum_{j=1}^{9} \hat{X}_2^j [e^{-(\frac{x_2-\hat{x}_2^j}{\sigma_2^j})^2} e^{-(\frac{x_1-\hat{x}_1^j}{\sigma_3^j})^2}]}{\sum_{j=1}^{9} [e^{-(\frac{x_2-\hat{x}_2^j}{\sigma_2^j})^2} e^{-(\frac{x_1-\hat{x}_1^j}{\sigma_3^j})^2}]} \quad (16) \\
\tilde{f}_{3_0}(x_3, x_2, u_2) &= \frac{\sum_{j=1}^{12} \hat{X}_3^j [e^{-(\frac{x_3-\hat{x}_3^j}{\sigma_3^j})^2} e^{-(\frac{x_2-\hat{x}_2^j}{\sigma_2^j})^2} e^{-(\frac{u_2-\hat{u}_2^j}{\sigma_{u_2}^j})^2}]}{\sum_{j=1}^{12} [e^{-(\frac{x_3-\hat{x}_3^j}{\sigma_3^j})^2} e^{-(\frac{x_2-\hat{x}_2^j}{\sigma_2^j})^2} e^{-(\frac{u_2-\hat{u}_2^j}{\sigma_{u_2}^j})^2}]}
\end{aligned}$$

$\hat{X}_i^j$ denotes the mean value of the membership function which belongs to Q_F of x_i, and appears in the consequent part of the j-th rule. The vector of parameters in each approximator, initialized in accordance with the values in Table 1, provides a good initial guess for the optimization procedure for parameter estimation from data.

5 Results

In order to make significant the comparison of the performance of our method with a data-driven approach , we look at each equation in (16-18) as a three-layer feedforward neural network, and exploit the Back Propagation algorithm (BP) for parameter estimation. As data-driven approach, we consider fuzzy-neural identifiers (FS-DD) whose structures are dimensionally fixed equal to the instantiated values of M in (16-18) but built from the numerical evidence. Let us observe that, since we exploit information derived from the qualitative simulation to fix the dimension, the performance of FS-DD is here improved with respect to its traditional application where also its structural dimension has to be derived from the data.

The application of our method to identify the system for simulation purposes follows a three-steps scheme:

1. *identification*: for each $\tilde{f}_{i_0}$, the values of parameters are tuned on a set of real data by using the BP algorithm in order to get an estimate $\underline{\tilde{\theta}}$ of $\underline{\theta}$, starting from the initial guess $\underline{\theta}_0$ provided as explained above;

2. *validation*: the accuracy of $\tilde{f}_i$, derived at step 1, is tested in accordance with a *parallel scheme* [5] on a new data set;
3. *simulation*: the accuracy of all of the three $\tilde{f}_i$ as a whole model is tested on a new data set in accordance with a *parallel scheme* where only the current inputs to the overall system (u_1, u_2) are measured data whereas the current output and input to each subsystem are simulated values.

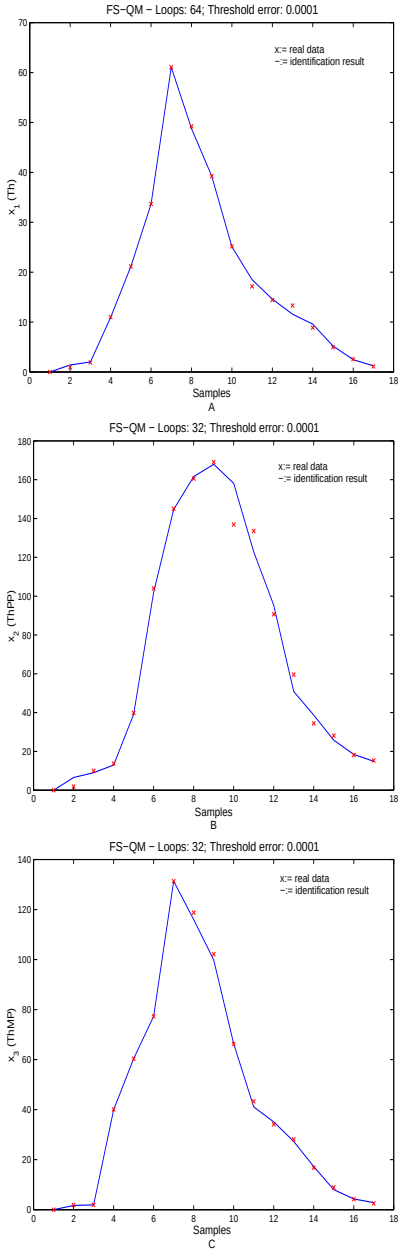


Figure 4. FS-QM identification results of x_1 (A), x_2 (B), x_3 (C) obtained with a threshold error equal to 0.0001.

[5] In a parallel scheme, the next value of the output variable is calculated given the current measurements of the input variables and the simulated value of the current output: $\tilde{y}_{k+1} = \tilde{f}(\tilde{y}_k, \underline{u}_k)$

Identification. Figure 4 shows the results we obtained with the application of our method in the identification phase of each x_i with a threshold error equal to 0.0001 by using a data set observed in normal subjects. Although the problem is ill-posed due to the small number of data, FS-QM performs quite well: this can be explained by the goodness of the initialization of both the identifier structures and the guesses of parameters. On the contrary, FS-DD, initialized by exploiting only the data, does not converge to the solution but it gets trapped in a local minimum. Let us fix our attention on the results of FS-DD identification of x_1: Fig. 5 highlights that, although the number of BP loops is highly increased from 350 (Fig. 5A) to 200000 (Fig. 5B), the training error remains constant. Moreover, we can observe a perfect fit on the first 11 data. Such a fit does not derive from identification but is rather related to the construction of the requested 11 rules.

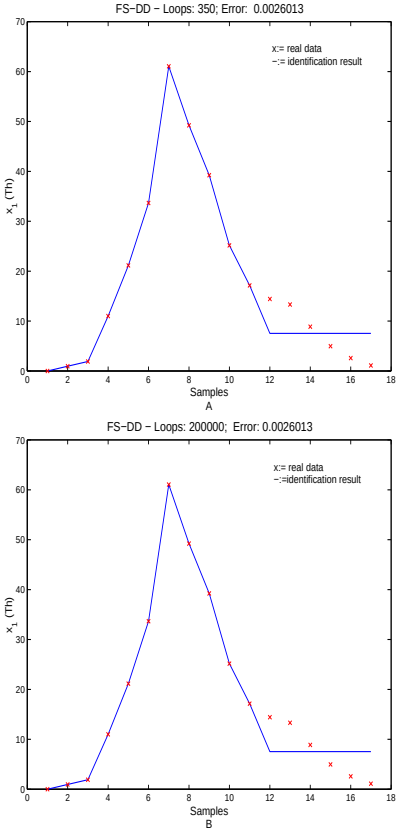


Figure 5. FS-DD identification results of x_1: (A) - number of BP loops equal to 350; (B) - number of BP loops equal to 200000. The training errors remain the same.

Validation and simulation. Each identified $\tilde{f}_i$ has been validated on a new set of data collected in an independent experiment, still on normal subjects. The results confirm the robustness of our approach to deal with NARX approximation schemes (see Table 3).

Our final goal is the construction of a simulator of the overall system dynamics that is capable to reproduce the system behavior in response to any input signals, at least in the range of the experimental settings previously defined. Such simulator is defined through the

[h]

subsystem	1	2	3
validation MSE	0.0065	0.0215	0.0018

Table 3. Validation errors calculated on data coming from an independent experiment on normal subjects. MSE stands for Mean Squared Error.

equations:

$$\begin{aligned} \tilde{x}_{1_{k+1}} &= \tilde{f}_1(\tilde{x}_{1_k}, \tilde{x}_{3_k}, u_{1_k}) \\ \tilde{x}_{2_{k+1}} &= \tilde{f}_2(\tilde{x}_{2_k}, \tilde{x}_{1_k}) \\ \tilde{x}_{3_{k+1}} &= \tilde{f}_3(\tilde{x}_{3_k}, \tilde{x}_{2_k}, u_{2_k}) \end{aligned} \tag{17}$$

where $\tilde{x}_{i_0} = x_{i_0}$ and $u_{i_k}, \forall k$, are the input data to the system. Our simulation results on the new data set (Fig. 6) clearly show the validity of FS-QM as an alternative methodology to identify nonlinear systems.

Remark. Clearly, within this approach the possibility of identifying parameters with a precise meaning is lost, but the reliable simulator at our disposal makes possible to enrich the knowledge of Th kinetics and to provide diagnostic and therapeutic information to physiologist: in particular, it will be possible to fit the main goal of the study, that is the understanding of insulin action on the Th metabolism in the cell. An indirect evaluation of the effects of diabetes on Th metabolism may be obtained by comparing the profiles simulated by (17) against the data of pathological subjects either treated or not. From a preliminary analysis of the results obtained, we can reasonably affirm that ThPP exhibits the same behavior both in normal and treated subjects.

6 Discussion

Mathematical modeling is often used in biomedical sciences to obtain a quantitative description of the (patho-)physiology underlying a physical system. Compartmental models represent a powerful class of such approaches: they are able to describe the mechanisms of release and uptake of a certain physiological substrate by contemporaneously expressing the system dynamics through a set of ODE's and quantifying the fluxes of substrate between compartments through a set of parameters. When the available data do not allow to identify the model parameters, due to measurement errors, inaccurate sampling time or, more simply, to an inadequacy of some model assumptions, the model itself is revised or discarded. An alternative solution is to resort to non-parametric modeling, that describes the dynamics of the system at hand relying on very general nonlinear functions, moulded by the available data. In this context, the structural assumptions made by compartmental models are relaxed, and only a descriptive quantitative knowledge may be derived.

Unfortunately, it may happen that also non-parametric approaches are likely to fail: as a matter of fact, since a posteriori unidentifiability may be also due to the lack of a sufficient number (or of a sufficient quality) of data, the search for a robust non-parametric description turns out to be infeasible in most cases.

In this paper we have described the successful application of a novel Intelligent Data Analysis methodology that aims at filling the gap between the parametric (compartmental) and non-parametric modeling. Thanks to the application of QR techniques, the structural assumptions on the relationships between the problem variables are

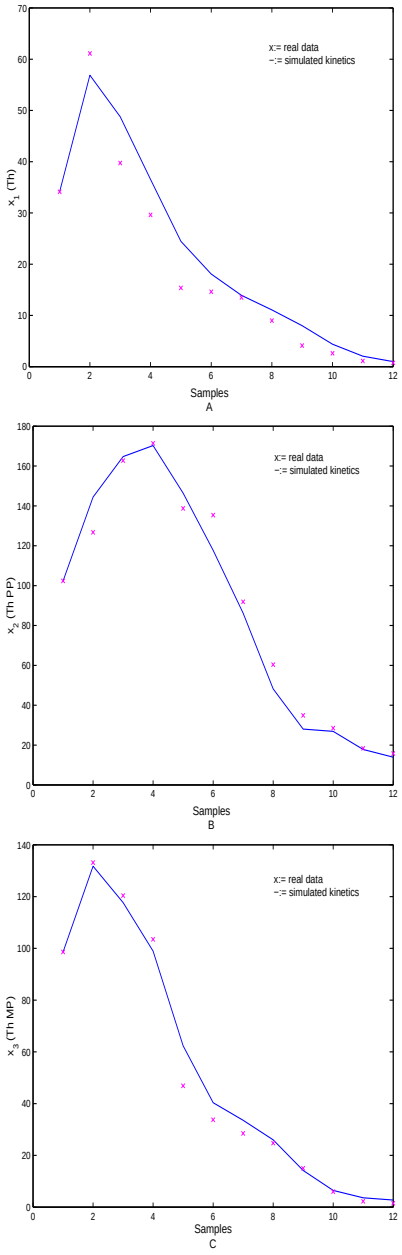


Figure 6. FS-QM simulation results of x_1 (A), x_2 (B), x_3 (C).

retained; moreover, thanks to the application of FS's, such assumptions are translated into a non-parametric model, whose parameters are properly initialized on the basis of a priori knowledge. Finally, the approximator of the system dynamics derived is robust enough for the purposes of the study.

From the application viewpoint, our proposed approach enabled us to draw physiologically sound conclusions from a set of data, that revealed to be unexploitable by classical compartmental modeling.

In conclusion, the results presented in this paper confirm our belief in the potential usefulness of our methodology for several classes of domains, among which medicine represents a prominent field: the presence of structural knowledge and the availability of costly data set, poor in number and in quality, motivate the development of approaches able to combine qualitative and quantitative information. The marriage of QR and fuzzy-based methods allows us to smooth down the distinction between mathematical models identification and Data Mining approaches, moving towards new approaches able to intelligently analyze the available data. In our future work, our aim will be to better systematize FS-QM in order to allow for its broader application in different areas.

7 Acknowledgement

We would especially like to thank A. Nauti and C. Patrini who provided us with the experimental data and physiological knowledge.

REFERENCES

[1] G.L. Atkins. *Multicompartmental Models in Biological Systems.* Chapman and Hall, London, 1974.

[2] R. Bellazzi, R. Guglielmann, and L. Ironi. A qualitative-fuzzy framework for nonlinear black-box system identification. In T. Dean, editor, *Proc. Sixteenth International Joint Conference on Artificial Intelligence (IJCAI 99)*, volume 2, pages 1041–1046, Stockholm, 1999. Morgan Kaufmann, San Francisco.

[3] R. Bellazzi, R. Guglielmann, and L. Ironi. How to improve fuzzy-neural system modeling by means of qualitative simulation. *IEEE Trans. on Neural Networks*, 11(1), 2000.

[4] R. Bellazzi, L. Ironi, R. Guglielmann, and M. Stefanelli. Qualitative models and fuzzy systems: an integrated approach for learning from data. *Artificial Intelligence in Medicine*, 14:5–28, 1998.

[5] E. Bradley, A. O'Gallagher, and J. Rogers. Global solutions for nonlinear systems using qualitative reasoning. In L. Ironi, editor, *Proc. 11th International Workshop on Qualitative Reasoning*, pages 31–40, Cortona, 1997. Istituto di Analisi Numerica - C.N.R., Pavia.

[6] E. Bradley and R. Stolle. Automatic construction of accurate models of physical systems. *Annals of Mathematics and Artificial Intelligence*, 17:1–28, 1996.

[7] A. C. Capelo, L. Ironi, and S. Tentoni. The need for qualitative reasoning in automated modeling: a case study. In *Proc. 10th International Workshop on Qualitative Reasoning*, pages 32–39, Stanford Sierra Camp, 1996.

[8] A.C. Capelo, L. Ironi, and S. Tentoni. Automated mathematical modeling from experimental data: an application to material science. *IEEE Trans. SMC*, 28(3):356–370, 1998.

[9] E.R. Carson, C. Cobelli, and L. Finkenstein. *The Mathematical Modeling of Metabolic and Endocrine Systems.* Wiley, New York, 1983.

[10] C. Cobelli and J.J. DiStefano III. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. *Am. J. Physiol.*, 239:R7–R24, 1980.

[11] M. Easley and E. Bradley. Generalized physical networks for automated model building. In T. Dean, editor, *Proc. Sixteenth International Joint Conference on Artificial Intelligence (IJCAI 99)*, volume 2, pages 1047–1052, Stockholm, 1999. Morgan Kaufmann, San Francisco.

[12] L. Ironi and S. Tentoni. An integrated quantitative-qualitative approach to automated modeling of visco-elastic materials from experimental data. In R. Teti, editor, *Proc. ICME 98 - CIRP International Seminar on Intelligent Computation in Manufacturing Engineering, Capri, 1-3 July 1998*, pages 381–388. CUES-Salerno & RES Communication-Naples, 1998.

[13] J. A. Jacquez. *Compartmental Analysis in Biology and Medicine.* Ann Arbor, Michigan, 1972.

[14] J. Jang. Anfis: Adaptive network based fuzzy inference system. *IEEE Trans. on Systems, Man and Cybernetics*, 23:665–685, 1993.

[15] H. Kay, B. Rinner, and B. Kuipers. Semi-quantitative system identification. *Technical Report*, TR AI99-279, 1999.

[16] T. Khannah. *Foundations of neural networks.* Addison-Wesley, Reading, MA, 1990.

[17] B. J. Kuipers. *Qualitative Reasoning: modeling and simulation with incomplete knowledge.* MIT Press, Cambridge MA, 1994.

[18] L. Ljung. *System Identification - Theory for the User*. Prentice-Hall, Englewood Cliffs, 1987.

[19] G. Rindi, C. Patrini, V. Comincioli, and C. Reggiani. Thiamine content and turnover rates of some rat nervous region, using labeled thiamine as a tracer. *Brain Res.*, 181:369–380, 1980.

[20] G. Rindi, C. Reggiani, C. Patrini, G. Gastaldi, and U. Laforenza. Effect on ethanol on the in vivo kinetics of thiamine phosphorilation and dephosphorilation in different organs-ii. *Acute effects Alcohol and Alcoholism*, 27:505–522, 1992.

[21] *The Merck Manual.* Merck Sharp and Dohme Research Laboratories, 1987.

[22] L.X. Wang. *Adaptive Fuzzy Systems and Control: design and stability analysis.* Englewood Cliff, NJ:Prentice–Hall, University of California at Berkeley, 1994.

Using Neural Nets for Decision Support in Prescription and Outcome Prediction in Anticoagulation Drug Therapy

S. Byrne[1], **P. Cunningham**[2], **A. Barry**[3], **I. Graham**[4], **T. Delaney**[3] and **O.I. Corrigan**[1].

Abstract. In this paper we consider the use of Artificial Neural Networks (ANNs) in decision support in anticoagulation drug therapy. In this problem domain ANNs can be used to learn the prescribing behaviour of expert physicians or alternatively to learn the outcomes associated with such decisions. Both these possibilities are evaluated and we show how, through the prediction of outcomes that the prescribing practice of physicians may be improved – the ANN can learn to predict the outcome resulting from prescribing levels better than expert physicians. In evaluating the use of ANNs on these regression problems the possibility of *ensembling* these nets to improve performance is also considered. We show that ensembles of networks do not improve over individual networks, presumably because of the lack of diversity in the outputs of the ANNs.

1. INTRODUCTION

Artificial Neural Networks (ANNs) have the ability to discern complex and latent patterns in the data presented to them. This attribute of ANN makes them powerful tools for modelling and predictive purposes. Consequently, neural network technology is finding increasing uses in medicine as medical decision support systems. Maclin *et al*. reported the use of ANN to aid in the diagnosis of renal and hepatic carcinomas [1], [2]. Several other clinical applications have also been reported e.g. in the treatment of cardiac diseases [3] and to predict patient gentamicin and cyclosporin peak and trough blood concentrations. [4], [5].

In this paper we present some results of a study on the use of ANNs in anticoagulant drug therapy. In such a situation there are various possible rôles for Machine Learning (ML) in decision support. The ML system may learn the decision-making behaviour of an expert as shown by the arrow at A in Figure 1. In a financial loan approval system this would involve learning some of the skills of the underwriter; in our system this involves learning the prescribing patterns of the clinician. In this scenario the learning system may achieve the competence of the domain expert but it will never surpass the competence of the human expert. Thus, for this scenario to be successful it requires skilled domain experts from which to learn.

[1] Department of Pharmaceutics, Trinity College, Dublin 2.
[2] Department of Computer Science, Trinity College, Dublin 2.
[3] Pharmacy Department, The Adelaide and Meath Hospitals incorporating the National Children's Hospital, Tallaght, Dublin 24.
[4] Cardiology Department, The Adelaide and Meath Hospitals incorporating the National Children's Hospital, Tallaght, Dublin 24.

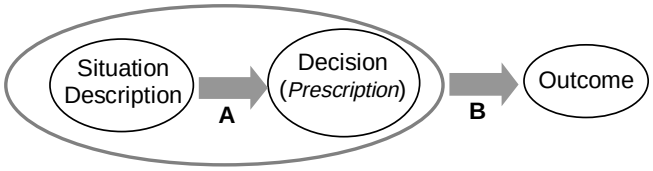


Figure 1. Rôles for Machine Learning systems in decision support.

In many ways, the more exciting prospect is for the ML system to learn likely outcomes associated with specific scenarios (option B in Figure 1). In the loan approval scenario this would involve learning the kinds of loans that turn into bad debts; or in our scenario, the medical outcomes associated with prescribing decisions. In this paper we evaluate both of these approaches in the anticoagulant therapy domain.

Anticoagulant drug therapy inhibits or delays coagulation of the blood. The decision to anticoagulate a patient is made on an individual basis, taking into account both the potential benefits and risks. Warfarin has a very narrow therapeutic index and therefore, for safe rational anticoagulant therapy, an understanding of the interrelationships, which influence an individual's response to the drug, is essential. Several studies have shown that satisfactory clinical control of patients on anticoagulant therapy was maintained in approximately half of the patients at any given time [6], [7]. In order to improve the quality of patient care, we consider that there is a role for the use of ANNs to help in the prediction of warfarin dosage adjustment. We present results on the use of single ANNs and ensembles of ANNs in learning appropriate prescribing levels and also results on predicting outcomes associated with such decisions. The ensembling of ANNs is a *committee of experts* idea where the aggregated results of several networks is better than the results of a single network. The methods and results are presented in the next sections and the paper concludes with an assessment of the results and a discussion on issues of bias in the predictions of the ANNs.

2. METHOD

The Adelaide and Meath Hospitals in Dublin, incorporating the National Children's Hospital, Tallaght was identified in previous work as the first hospital in the Rep. of Ireland with a pharmacist involved in it's anticoagulant clinic [8]. Ethics committee approval for the study was obtained and a researcher attended the outpatient department of the Tallaght Hospital twice weekly when the anticoagulant clinic was in operation between October 1998 and July 1999.

163 patients were approached and asked to take part in the study. 105 patients gave written consent and by doing so, agreed to undergo a structured interview with the researcher every time they attended the clinic and to have additional liver function tests (LFTs) carried out during the study period.

After an initial literature review 25 parameters were measured, these were age, weight, height, gender, marital status, reason for anticoagulation, target INR (International Normalised Ratio) range, measured INR, previous and subsequent dose of warfarin, duration of therapy, number of hours since last warfarin dose, number of adverse drug reactions, concurrent prescribed and over the counter medication, LFT results, smoking habits, alcohol consumption and compliance to prescribed medication.

As outlined in the introduction, work to date has concentrated on training neural networks to learn a patient's dose of warfarin prescribed by clinicians and also on predicting a patient's INR reading given a particular dose. The INR reading is an international standardised method of reporting a patient's prothrombin time (the time it takes for the patient's blood to clot). Initial work using an in-house neural network package concentrated on developing optimum networks for the above output parameters i.e. the number of hidden units, the learning rates and momentum variables were adjusted until the error of the networks were minimised. The training algorithm used on each occasion was the backpropagation algorithm.

3. RESULTS AND DISCUSSION

Two patients decided to drop out of the study, leaving a total of 103 patients in the study population. Data was collected from the 103 patients on 656 occasions, mean=6.5 (S.D.±2.7). This corresponds to 49.89 patient treatment years (PTY) of anticoagulation therapy. The population had a mean age of 57.2yrs, mean height of 171.5cm and a mean weight of 80.1Kg. The main indication for anticoagulation was atrial fibrillation 40.8%(n=42). Overall, patients spent 21.56 PTY outside their desired INR ranges and for 14.44 (67.0%) of these PTY, the patients were under anticoagulated. This difficulty in control indicated a rôle for ML based decision support systems.

3.1. Learning prescribing practice

The data collected was firstly used to train a neural network with a patient's dose as the desired output variable – the objective being to learn the practice of expert clinicians in this area. The input layer to the network had 23 nodes and the output layer had 1 node. The 'reason for anticoagulation' was omitted from the input layer, as this information was covered by the 'target INR' input parameter. The network had one hidden layer and the number of nodes within this layer was varied, along with the learning and momentum rates, to produce a model that best fitted validation data. The neural network structure that gave the best results had 1 node in the hidden layer and learning and momentum rates equal to 0.1. Using 10-fold cross validation on the 656 data samples available we obtained an estimation of generalisation error of 6.6%. That is, the network should be able to prescribe doses to within 6.6% of that of an expert clinician.

In addition an ensemble of 5 neural nets were built and trained on different subsets of training data and were then tested on 101 data points held back from the training process. The output from the ensemble was produced by simply averaging the outputs from the individual nets in the ensemble for each test point. This produced an average error of 6.24% on the 101 test data points. (This figure is not directly comparable to the 6.6% figure, which is derived from cross validation.) In a regression task such as this the error reduction due to the ensemble is expressed as:

$$E = \overline{E} - \overline{A} \qquad (1)$$

where E is the overall error of the ensemble, $\overline{E}$ is the average generalisation error of the ensemble components and $\overline{A}$ is the ensemble diversity (i.e. variance in individual predictions). In this situation there is little scope for the ensemble to reduce the error because it is already small and more importantly there is little diversity (ambiguity) among the ensemble members (see Cunningham & Carney, 2000 for more details on this) [9]

Subsequently it was decided to design and train another neural network with a patient's dose as the desired output variable, but in this case the LFT data was to be omitted from the input variables (LFT data would not normally be available to the clinicians). Therefore the number of input parameters was reduced to 17 and similarly the neural network structure that gave the best error measurements had 1 node in the hidden layer and learning and momentum rates of 0.1. A 10-fold cross validation test was performed to produce an estimate of error and a figure of 6.1% was obtained. This result is surprising because it indicates that LFT data does not inform the prescribing process. Again, an identical ensemble exercise as above was conducted and produced an error of 4.3% on the 101 test samples. Since this scenario with the LFT data excluded is the more realistic the results are shown in more detail in Figure 2. It can be seen from this graph that the predictions across all 101 cases are good and the average error of 4.3% does not hide any individual large errors.

If we compare the average error of this ensemble with its component nets we see that it represents a very slight improvement. This improvement is shown in Figure 3 where the range of errors for the component nets is also shown. Clearly the advantage of the ensemble is not material in reducing the error; it is significantly better than the worst component nets but not better than the best net in the ensemble.

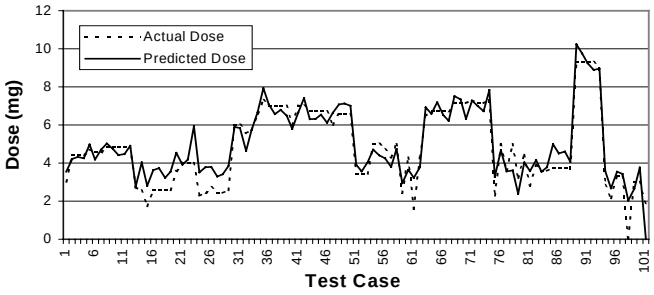


Figure 2. Results of an Ensemble of 5 ANNs in determining dose for 101 test cases.

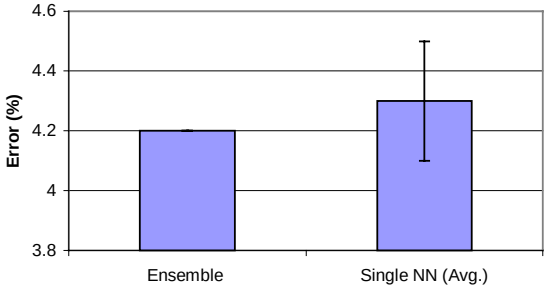


Figure 3. Comparison of error of ensemble of 5 ANNs with that of the component networks.

3.2. Predicting Outcomes

In order to illustrate ANNs flexibility it was decided to design and train a network using a patient's INR measurement as the desired output variable. The input layer had 22 nodes and the output layer had 1 node i.e. the 'measured INR' was the desired output variable and the 'subsequent warfarin dosage' was omitted from the exercise. The neural network structure that gave the minimum error values had 4 nodes in the hidden layer and learning and momentum rates equal to 0.1. This more elaborate network structure suggests that this process is more complex than the prescribing process described above. 10-fold cross validation indicated a generalisation error of 8.6% and again an ensemble of 5 networks did not produce an appreciable error reduction on its component networks (8.8% for the ensemble and 8.9% on average for the component networks). Given the higher error here it might be expected that the ensemble would have a greater benefit however it appears that there is little diversity in the component networks. This in turn suggests that the amount of training data available covers the problem domain well (see Cunningham, Carney & Jacob, 2000 for related discussion on this issue) [10].

It is important to assess the quality of this prediction of INR outcome because it offers the potential to improve the current prescribing practice. With this objective the target INR that the physician had in mind was also recorded in the data gathering process. For the given dose this is the INR that the physician expects to achieve, so we can compare this directly to the INR prediction produced by the ANN.

Figure 4 below shows the mean physicians' INR predictions and the neural networks mean INR predictions compared to the actual INR values for a random sample of 101 meetings with the study population. The physicians' INR prediction was on average 1.05 INR units away from the actual INR value and the neural network was 0.75 INR units away from the actual value. Using a paired sample t-test, one can conclude that the network's prediction was significantly closer to the actual patient's INR measurement than that of the physician ($p<0.00014$). This offers the prospect of developing a tool that will predict the outcome for a given dose more accurately than at present. This would help the physician to refine his/her prescribing to better target the desired outcome.

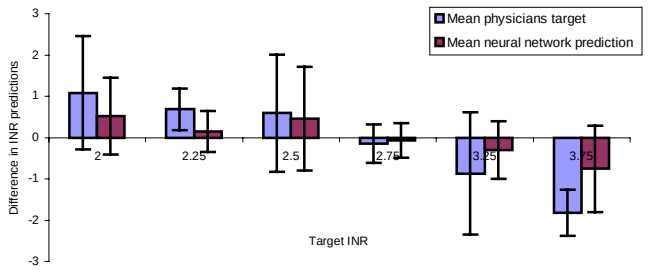


Figure 4. Plot of differences between the physician's INR prediction and the desired INR value and the neural networks INR prediction and the desired INR value.

In Figure 4 the results are divided into intervals according to target INR outcome. The predictions for physician and ANN are most accurate for INR values around 2.75. Above that the ANN and physician underestimate and below that there are overestimations. If we look at the actual distributions of INR in the training data we can see the reason for this (Figure 5). The distribution of INR values in the training data is skewed with a large number of samples around 2.25 to 2.75. It seems that ANNs trained with this *a priori* distribution are biased toward these values. INR values around 2 are poorly represented in the training data and these outcomes in the test data are overestimated towards the mean of the training data. In turn high INR values are also poorly represented in the training data and in turn test cases in this region are underestimated.

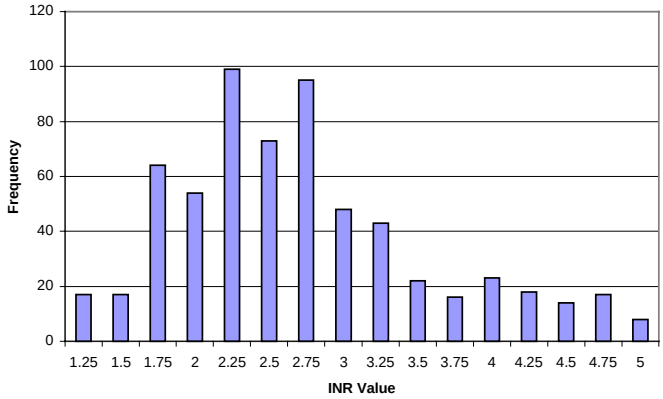


Figure 5. Distribution of INR values in training data.

This is a manifestation of a general problem of skewed distributions in machine learning. This problem is well documented in classification problems (Provost & Fawcett, 2000) [11] where it is well known that classifiers are biased away from minority classes. What we see here is a regression situation where the ANN is biased toward commonly occurring values.

4. CONCLUSIONS

The evaluation presented here shows that ANNs can be used to learn the prescribing patterns of expert physicians in anticoagulant drug therapy to within an accuracy of 6%. Further, ANNs can be used to predict the outcome of prescribing decisions. The evaluation shows that these predictions of outcome are twice as accurate as those of physicians. This offers the possibility of improving the performance of expert physicians by using these accurate predictions of outcome to guide and refine prescribing decisions.

We also evaluated the use of ensembles of ANNs to improve on the error of an individual net on these tasks. No significant improvement was found presumably because of the lack of diversity amongst individual nets. We speculate that this suggests that the 656 training samples used in this study provide good coverage of the problem domain. A good model that does not overfit to training data can be built with just one ANN.

It emerged in the evaluation that the accuracy on predicting the outcome (INR level) was not uniform across the distribution of outcomes. It appears that this is due to the skewed distribution of INR outcomes. An analysis of the errors across the distribution of INR values shows that there is a bias towards the dominant values. Correcting this bias warrants further investigation and is a promising area for future research.

ACKNOWLEDGEMENTS

We would like to thank the phlebotomy and laboratory departments of the Adelaide and Meath Hospitals, incorporating the National Children's Hospital, Tallaght for their assistance in taking and

analysing blood samples. We would also like to thank Antigen Pharmaceuticals Ltd., Enterprise Ireland and the College of Pharmacy Practice for financial support.

REFERENCE

1. Maclin P.S., Dempsey J., Brooks J. and Rand J., Using neural networks to diagnose cancer, J Med Syst 1991, **15** (1), 11-19
2. Maclin P.S., Dempsey J., Using an artificial neural network to diagnose hepatic masses. J Med Syst 1992, **16** (5), 215-25
3. Furlong J.W., Dupuy M.E., and Heinsimer J.A., Neural network analysis of serial cardiac enzymes data. A clinical application of artificial machine intelligence, Am J Clin Pathol, 1991, **96** (1), 134-41
4. Brier M.E., Neural network gentamicin peak and trough predictions from prior dosing data, Clin Pharm and Ther, 1995, **57** (2), 157
5. Brier M.E., Neural network predictions of cyclosporin blood concentrations, Clin Pharm and Ther, 1995, **57** (2), 157
6. Duxbury B. McD., Therapeutic control of anticoagulant treatment, BMJ, 1982, **284**, 702-704
7. Radley A.S., Hall J., Farrow M. and Carey P.J., Evaluation of Anticoagulant Control in a Pharmacist Operated Anticoagulant Clinic, J Clin Pathol 1995,**48**,545-547
8. Byrne S., Corrigan O.I., An overview of pharmaceutical manpower and services in hospitals situated in the republic of Ireland during 1998, IPJ, 1999, **77** (7), 220-25
9. Cunningham, P., Carney, J., Diversity versus Quality in Classification Ensembles based on Feature Selection, *accepted for presentation at the 11th European Conference on Machine Learning (ECML 2000)*, Barcelona, Spain, May 2000.
10. Cunningham, P., Carney, J., Jacob, S., Stability Problems with Artificial Neural Networks and the Ensemble Solution**,** to appear in *AI in Medicine*, 2000, Vol. **20**.
11. Provost, F., T. Fawcett, Robust Classification for Imprecise Environments, to appear in *Machine Learning*, 2000 (available at http://www.stern.nyu.edu/~fprovost/).

Inconsistency tests for patient records in a coronary heart disease database

Dragan Gamberger [1], **Nada Lavrač** [2], **Goran Krstačić** [3], and **Tomislav Šmuc** [4]

Abstract. The work presents the results of inconsistency detection experiments on the data records of an atherosclerotic coronary heart disease database collected in the regular medical practice. Medical expert evaluation of some preliminary inductive learning results have demonstrated that explicit detection of outliers can be useful for maintaining the data quality of medical records and that it might be a key for the improvement of medical decisions and their reliability in the regular medical practice. With the intention of on-line detection of possible data inconsistences, sets of confirmation rules have been developed for the database and their test results are reported in this work.

1 Introduction

The motivation for the research presented in this work stems from the fact that modern medical decision processes are generally based on patient data from many different sources which are typically collected and archived by a multiterminal or distributed computer systems. Such organization enables prompt and high quality decisions of medical doctors supported by abundance of available data [2] but it also enables that errors of different sources, caused by the work of many different people and/or instrumentation can directly enter patient records. Detection of data inconsistencies at the global level of every patient record can help in tracing systematic as well as spurious errors in the data acquisition process and in this way it can be an important part for insuring high quality and reliability of data used for medical decision making [5]. On the other side, the existence of data inconsistencies do not need to be the sign of data errors but may be the consequence of some atypical medical case. Attracting attention of medical doctors to such patient records may be interesting both from the point of view of medical science as well as for avoiding everyday medical practice routine errors.

Our interest in inconsistency testing is the result of many inductive learning experiments performed on a database of atherosclerotic coronary heart disease (ACHD) patients prepared at the Institute for Prevention of Cardiovascular Disease and Rehabilitation, Zagreb Croatia. In the data preparation phase, the saturation filter [4] was used to detect and eliminate outliers from the database. This is a necessary step in the knowledge discovery process which enables the induction of globally relevant rules. Medical expert evaluation has demonstrated that the detection of outliers is a very interesting result by itself, which suggested the idea of using noise detection algorithms developed for data preprocessing in inductive machine learning as a tool for data cleaning of patient records. In this work two different approaches to the problem of inconsistency testing are presented in Section 2. This is followed by the presentation of the medical domain used in experiments and the results of the experiments in Sections 3 and 4, respectively. The algorithms used for outlier detection and rule construction are out of the scope of this work since their description can be found in [3, 4].

[1] Rudjer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia, E-mail Dragan.Gamberger@irb.hr

[2] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia, Nada.Lavrac@ijs.si

[3] Institute for Prevention of Cardiovascular Disease and Rehabilitation, Draškovićeva 13, 10000 Zagreb, Croatia

[4] Rudjer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia, E-mail Tomislav.Smuc@irb.hr

2 Inconsistency tests

Machine learning approaches to inconsistency testing in patient records can be either based on outlier (noise) detection algorithms or on a set of rules that are supposed to be true for the data in patient records. The later approach can be used also for on-line inconsistency testing but it requires the construction of rules with specific properties. In both cases, testing is based on supervised machine learning algorithms which require that patient records are grouped in two or more classes. The classes can be defined either by domain experts or by values of one or more descriptors available in the patient record. Correlation between defined classes and data contained in patient records is the main mechanism used in inconsistency detection. Appropriate class assignment is one of the main problems of machine learning approaches to inconsistency testing and it is specially analyzed in Section 4.

2.1 Explicit outlier detection

The first approach, in the work called *explicit outlier detection* can be without changes used on very different patient records. It is actually a noise detection algorithm for data of two classes, described in detail in [4], used in the data preprocessing phase (data cleaning) of inductive learning algorithms. It works on the set of records, trying to identify significant differences among positively and negatively classified records. Complexity of the least complex hypothesis correct for all available examples (true for all positive and false for all negative examples) is estimated without constructing any concrete hypothesis. Those records which are difficult for correct classification and which, by their elimination enable direct reduction of the complexity of the least complex hypothesis, are detected as outliers. The approach is appropriate for off-line data analysis. Its main drawback is its time complexity as well as the fact that for multiclass problems the algorithm must be repeated for every reasonable definition of class positive and class negative records.

Descriptor	Abbreviation	Characteristics
Anamnestic data		
sex	SEX	1-man 2-woman
age	AGE	continuous (years)
height	H	continuous (m)
weight	W	continuous (kg)
body mass index	BMI	continuous ($kg\ m^{-2}$)
family anamnesis	F.A.	1-negative 2-positive
present smoking	P.S.	1-negative 2-positive 3-very positive
diabetes mellitus	D.M.	1-negative 2-pos. medicament therapy 3-pos. insulin therapy
hypertension	HYP	1-negative 2-positive 3-very positive
stress	STR	1-negative 2-positive 3-very positive
Laboratory tests		
total cholesterol	T.CH.	continuous ($mmol\ L^{-1}$)
trygliceride	TR	continuous ($mmol\ L^{-1}$)
high density lipoprotein	HDL/CH	continuous ($mmol\ L^{-1}$)
low density lipoprotein	LDL/CH	continuous ($mmol\ L^{-1}$)
uric acid	U.A.	continuous ($\mu mol\ L^{-1}$)
fibrinogen	FIB	continuous ($g\ L^{-1}$)

Table 1. The names and the characteristics of 16 anamnestic and laboratory testing descriptors.

2.2 Rule-based outlier detection

The second approach, called *rule-based outlier detection* is more appropriate for on-line inconsistency testing. It works with data of one patient record only and the consequence is its simplicity and high execution speed. The approach is actually a set of logical tests that must be satisfied by every patient record. If one or more of the tests is not satisfied, the record is detected as an outlier. The logical tests are defined by the set of rules that hold for the patient records in the domain. The main drawback of the approach is that the set of rules must be developed specially for the tested type of records. Moreover, the used rules must be highly reliable rules with a very small number of mispredictions, leading to false outlier alarms. Such rules can be constructed by domain experts but also by inductive learning algorithms. Because of the required rule reliability, the concept of *confirmation rules* seems appropriate for this task [3]. In this concept, separate rules are constructed for the positive and negative class cases. The confirmation rules for the positive class must be true for many positive cases and for no negative case. If a negative case is detected true for any confirmation rule developed for the positive class, it is a reliable sign that the case is an outlier. In the same way, confirmation rules constructed for the negative class can be used for outlier detection of positive patient records. An additional advantage of the approach is that the user can have the information about the rule which caused the alarm what can be useful in the error detection process.

3 Data Set

For this work, a database representing typical medical practice in atherosclerotic coronary heart disease ACHD diagnosis has been prepared. The data describe patients who entered the Institute for Prevention Cardiovascular Disease and Rehabilitation, Zagreb Croatia, in a few months period. The set of descriptors represents all potentially interesting and typically available information about patients. The descriptor set includes anamnestic data (10 items), laboratory test results (6), the resting ECG data (5), the exercise test data (5), echocardiogram results (2), vectorcardiogram results (2), and long term continuous ECG recording data (3). It makes altogether 33 descriptors. Only patients with complete data have been included into the data set, resulting in the data set with 238 patients in total. The descriptors are cited in Tables 1 and 2.

Descriptor	Abbreviation	Characteristics
ECG at rest		
heart rate	HR	continuous (beats min^{-1})
ST segment depression	ECGst	1-negative 2-positive 1mm 3-positive $\geq$ 2mm
serious arrhythmias	ECGrhyt	1-negative 2-positive
conduction disorders	ECGcd	1-negative 2-positive
left ventricular hypertrophy	ECGhlv	1-negative 2-positive
Exercise ECG		
ST segment depression	ExECGst	continuous (mm)
serious arrhythmias	ExECGrhyt	1-negative 2-positive
conduction disorders	ExECGcd	1-negative 2-positive
hypertensive reaction	ExECGhyp	1-negative 2-positive
New York Heart Ass. functional class	ExECGNYHA	class I - IV
Echocardiography		
left ventr.		
internal diameter	EchoLVID$_d$	continuous (mm)
left ventr. ejection fraction according to Simpson	EchoLVEF	continuous (%)
Vectorcardiography Q		
transmural MI	VCG Q	1-negative 2-positive
left ventricular hypertrophy	VCGhlv	1-negative 2-positive
Long term continuous ECG		
serious arrhythmias	HOLrhyt	1-negative 2-positive
conduction disorders	HOLcd	1-negative 2-positive
ST segment depression	HOLst	continuous (mm)

Table 2. The names and the characteristics of 17 non-invasive diagnostic descriptors.

The classification of all patients was performed by the cardiologist and it reflects generally accepted medical knowledge. The classification is mostly based on the results of the most important tests. These are: exercise testing, long term ECG recording and echocardiography. In exercise testing ST segment depression or elevation, serious cardiac arrhythmias, and conducting disturbances are the important parameters. Additionally, the NYHA classification [6] and clinically significant metabolic equivalents (METs) are used [1]. Similar parameters can be found in long term ECG recording, except MET and NYHA classification. Dyastolic internal diameter of left ventricular with parasternal short axis view and left ventricular ejection fraction (according to Simpson) are determined from the echocardiogram.

For this research the cardiologist classified patients into 5 groups whose main features are summarized below:

Group I Healthy patients without verified ACHD but with possible

present cardiovascular risk factors.

Group II-V These are patients with previous myocardial infarction. They were classified by the results of non-invasive cardiovascular tests and their condition after some coronary angioplastic or cardiosurgery treatment. They are all under medicament treatment.

Group II Patients with normal results of exercise testing, long term recording and echocardiogram.

Group III Patients with ST segment depression 1.00 mm in exercise testing and during long term ECG recording, left ventricular ejection fraction higher than 55%, METs 10.

Group IV Patients with ST segment depression equal or higher than 2.00 mm in exercise testing and during long term ECG recording, left ventricular ejection fraction less than 55% (40-54%), left ventricular internal diameter more than 6.0 cm, NYHA II-III, METs 5–10.

Group V Patients having ST segment elevation or depression > 3.00mm, left ventricular ejection fraction less or equal to 30%, left ventricular inner diameter greater than 6.5 cm, NYHA III-IV, METs<5.

In experiments with expert defined classes the groups III-V represented the positive class while groups I-II were in the negative class.

4 Experimental results and medical evaluation

The domain of 238 patient records consists of two sets: the dataset of 150 patients collected earlier, has been used for preliminary experiments and rule development, while the set of remaining 88 records collected later, has been used for test purposes only. The first set will be in the rest of the paper called the main set and the second one the test set.

4.1 Explicit outlier detection results

The set of experiments started with explicit outlier detection for the main set and for the positive and negative classes as defined in Section 3 by medical doctors. There have been only two detected outliers (patient records number 28 and 52) and both of them are very interesting cases. The first one is actually an older patient after a serious cardiosurgical treatment who was in spite of non-optimal laboratory tests intentionally put into Group II (patients with normal results of exercise testing, long term recording and echocardiogram). The second patient was also in Group II but after its detection as an outlier, medical doctor agreed that Group III would be much more appropriate for the patient. In the analysis it was detected that the main reason for its inclusion in Group II were good exercise testing results. But the results were misleading because the patient was so weak that he could sustain only 2 minutes (instead of 7 - 9 minutes) of exercise. It means that an actually important outlier had been detected, that its detection helped in finding more appropriate diagnostic group for the patient, and that the medical doctor has found out that exercise testing results are reliable data only if the tests could be and have been performed completely and correctly.

The same explicit approach to outlier detection was also applied on the test set which resulted in the detection of patient records 174, 214, 227, and 230. More reliable results should be expected if the outlier detection algorithm is applied on the target dataset consisting of both main and test set. The result of this experiment was the detection of the same four detected records from the test set and in total five records from the main set. Besides examples 28 and 52, that were detected also in the first experiment, the set of detected records from the main set included also cases number 1, 43, and 98. Table 3 summarizes the results of the first three experiments with explicit outlier detection. The results show a weakness of the explicit approach to noise detection demonstrated by the fact that different outliers have been detected for the main set depending on the target set.

Target set	Detected outliers	
	In main set	In test set
main	28 52	- - -
test	- - -	174 214 227 230
main + test	1 28 43 52 98	174 214 227 230

Table 3. Results of explicit outlier detection for different sets of target patient records.

Medical evaluation showed that cases 1 and 43, patients from Group III, are not expert recognized outliers. These cases can be accepted as false alarms of the explicit outlier detection approach. A completely different situation is the patient number 98 from Group III, classified as a serious case but with practically normal results of laboratory tests. A medical doctor accepted the patient as a special case and was satisfied that machine learning methods recognized it as an outlier. Due to the patient's medical history, the case classification remained unchanged. Two out of four cases detected in the test set are border line cases (cases number 174 and 227) and the remaining two are real medical outliers: one is a difficult coronary patient from Group V with diagnosed cardiopathia dilatativa therefore, not an ACHD patient (number 214), while the other one is an atypical ACHD patient whose disease could be detected only by echocardiography (number 230).

4.2 Rule-based outlier detection results

With the intention to show the application of a rule-based outlier detection approach, the main set was used to induce confirmation rules, explained in Section 2.2, for both classes. The rules for the positive class were $ExECGst > 0.45mm$ and $HOLst > 0.65mm$. Each of these two rules is true for about 95% of the positive class cases in the main set and false (except for explicitly detected outliers) for all negative class cases in this set. These two rules can be used as constraints that should not be satisfied by negative class cases. The rules detected the following negative class outliers in the test set: 165, 174, 185, and 227. There was only one confirmation rule for the negative class consisting of two conditions $ExECGst \leq 0.45mm \wedge HOLst \leq 0.65mm$. The rule is true for about 98% of negative class cases and false for all positive class cases in the main set. The rule detected positive patient records 214 and 230 as outliers within the test set. The results are presented in the first row of Table 4. Comparing results obtained by explicit and rule-based outlier detection it can be noted that the later approach selected the same records as the former approach but that it also detected two more records: 165 and 185. The result demonstrates the applicability of the rule-based method for inconsistency testing. The method is interesting because its application is much simpler for all other future patient records. Medical evaluation of the cases 165 and 185 showed that they are actually not false alarms. One of them is a border line case who was intentionally put in Group II because of its medical history (case number 185) while in the other case the medical doctor accepted the suggestion and he has changed the patient group classification (case 165).

Classes defined by	Outlier detection	
	Explicit	Rule-based
medical doctors	174 214 227 230	165 174 185 214 227 230
$ExECGst > 0.45mm$	165 174 185 195 197 199 202 227 232 237	165 174 185 195 197 199 202 227 232 237
$HOLst > 0.65mm$	165 185 195 197 199 202 214 230 232 237	165 185 195 197 232

Table 4. Comparison of the outliers detected by explicit and rule-based detection approaches for differently defined patient record classes.

4.3 Results obtained by descriptor-based classifiers

In all previous experiments inconsistency testing was based on classes defined by medical doctors. Although the results are very reasonable for both methods and they agree in most detected records, this way of inconsistency testing is appropriate only for domains in which doctor classification exists. Existence of such classification assumes that there exists a dependency between the determined class and record data what practically ensures the quality of the inconsistency tests. In a general case, when there is no expert classification, one or more descriptors from the patient record should be used as a classification parameter. In this situation it is essential to select classifiers for which it can be assumed that their dependency with other data in the record exists. In the ACHD domain the induced rules demonstrated strong dependencies between expert classes and ST segment depression during exercise and long term continuous ECG monitoring. This is the main reason for selecting these data as appropriate classifiers when expert classification does not exist. The limit values between positive and negative classes are based on induced rules in previous experiments. Table 4 in its second and third row includes results obtained for positive classes defined by conditions $ExECGst > 0.45mm$ and $HOLst > 0.65mm$ respectively. In the left column are records detected as outliers in the test set by explicit approach while in the right one are detected by the rule-based method. In the explicit approach, the target sets included both the main and the test sets. Rules for the rule-based detection have been constructed always from the main set so that its own outliers have been previously excluded from it. The method resulted in successful detection of the most outliers detected based on expert classification. It should be noted that among outliers occur also some completely new cases, like numbers 195, 197, 232, and 237. Their medical evaluation demonstrated that in three out of four cases the problem was that the patient exercise testing was incomplete and the obtained results were misleading. In the fourth case (number 195) the patient had an asymptomatic (silent) ischaemic heart disease known by its differences between exercise and long term continuous measurements. It must be noted that detection of these four outliers was medically completely justified. Analysis of medical classifications in all four cases showed that medical expert reasoning also successfully detected the problems and that all cases were in appropriate groups in spite of data inconsistencies.

4.4 Subconcept discovery using outlier detection

The rules induced in the previous experiments show that both approaches for outlier detection use dependencies of a small number of data from the record. This practically means that only inconsistencies in a relative small part of the record can be detected. The problem can be solved by using the suggested methods iteratively based on same patient classifiers but with different descriptor subsets. The results of experiments in the ACHD domain with different descriptor subsets were not consistent and their reasonable medical evaluation was difficult. The cause of the problem can be the insufficient dependency among less important data in the records.

Some experiments with descriptor subsets led to very interesting results, typically detecting small patient subsets with special properties. In one of them all potential outliers in the positive class were characterized by positive reaction to drug therapy (i.e. their risk factor parameters were inside normal limits) as compared to a majority of ill patients that did not react positively to the drug therapy. Detection of this subset is an example of subconcept discovery which might be interesting for medical research purposes. In this way, our approach could be used in the same way as a subgroup discovery system.

5 Conclusions

This paper introduces two approaches to outlier (inconsistency) detection in medical datasets: explicit outlier detection and rule-based outlier detection. Explicit outlier detection is applicable in off-line analysis since it operates on the dataset level and involves data cleaning algorithms usually used in machine learning preprocessing. Rule-based outlier detection relies on rules induced upon previously collected data in the same dataset. It is applicable for on-line detection of inconsistencies in future records. We have applied both approaches on the ACHD patient dataset. Experiments were performed with expert classified records as well as with different descriptor-based classifications. The results indicated the sensitivity of both approaches for inconsistency detection. Although detected outliers differed from one experiment to another, most outliers were confirmed as special cases by subsequent medical expert evaluation. In order to detect possible inconsistencies in less important descriptors, experiments with different descriptor subgroups have been performed. Medical evaluation in some of these experiments recognized outliers with similar characteristics representing interesting domain subconcepts. The results could be important both with respect to everyday medical practice as well as for future medical research.

REFERENCES

[1] ACP/ACC/AHA Task force on Exercise Testing (1990). *Journal of American College Cardiology* **16**: 1061-1065.

[2] Diamond, G.A.,& Forester, J.S. (1979). Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. *New England Journal of Medicine* **300**:1350.

[3] Gamberger, D., Lavrač, N., and Grošelj C. (1999) Diagnostic rules of increased reliablity for critical medical applications. In *Proc. of Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making (AIMDM'99)*, pp.361–365.

[4] Gamberger, D., Lavrač, N., and Grošelj C. (1999) Experiments with noise filtering in a medical domain. In *Proc. of International Conference of Machine Learning (ICML'99)*, pp. 143–151.

[5] Grošelj, C., Kukar, M., Fetich, J. & Kononenko, I. (1997). Machine learning improves the accuracy of coronary artery disease diagnostic methods. *Computers in Cardiology* **24**: 57–60.

[6] Wayne A.R., Schlant R.C., Fuster V., In HURST'S: The Heart, Arteries and Veins. McGrawc Hill, NY, 1127-1263.

A System for Monitoring Nosocomial Infections

E. Lamma[1], M. Manservigi[2], P. Mello[1], F. Riguzzi[3], R. Serra[4], S. Storari[1]

Abstract. In this work, we describe a project, jointly started by DEIS University of Bologna and Dianoema S.p.A., in order to build a system which is able to monitor nosocomial infections. To this purpose, the system computes various statistics that are based on the count of patient infections over a period of time. The precise count of patient infections needs a precise definition of bacterial strains. In order to find bacterial strains, clustering has been applied on the microbiological data collected along two years in an Italian hospital.

1 MICROBIOLOGICAL DATA ANALYSIS

A very important problem that arises in hospitals is the monitoring and detection of nosocomial infections. A hospital-acquired or nosocomial infection is a disease that develops after the admission to the hospital, and is the consequence of a treatment, not necessarily a surgical one, or work by the hospital staff. Usually, a disease is considered a nosocomial infection if it develops 72 hours after the admission to the hospital. In Italy, this problem is very serious: actually almost the 15% of patients admitted to hospitals develop a nosocomial infection. In order to monitor nosocomial infections, the results of microbiological analyses must be carefully collected and analysed.

In Italy, a great number of hospitals manages analysis results by means of a software system named Italab C/S, developed by Dianoema S.p.A. Italab C/S is a Laboratory Information System based on a Client/Server architecture, which manages all the activities of the various analysis laboratories of the hospital. Italab C/S stores all the information concerning patients, the analysis requests, and the analysis results. In particular, for bacterial infections data includes:

- information about the patient: sex, age, hospital unit where the patient has been admitted;
- the kind of material (specimen) to be analysed (e.g., blood, urine, saliva, pus, etc.) and its origin (the body part where the specimen was collected);
- the date when the specimen was collected (often substituted with the analysis request date);
- for every different bacterium identified, its species and its antibiogram.

For each isolated bacterium, the antibiogram represents its resistance to a series of antibiotics. The set of antibiotics used to test bacterial resistance can be defined by the user, and the antibiogram is a vector of couples (antibiotics, resistance), where four types of resistance are possibly recorded: R when resistant, I when intermediate, S when sensitive, and null when unknown.

The antibiogram is not uniquely identified given the bacterium species but it can vary significantly for bacteria of the same species. This is due to the fact that bacteria of the same species may have evolved differently and have developed different resistances to antibiotics. Bacteria with similar antibiograms are grouped into "strains".

From these data, infections are now monitored by means of a Italab C/S module called "Epidemiological Observatory" that periodically generates reports on the number of infections detected in the hospital. These reports are configurable and show the number of found infections with respect to other data such as specimen characteristics (material and origin) and patient characteristics (hospital unit, sex, age, etc.). Examples of such reports are:

- for every species, for every material and for every origin, show the number of infections found;
- for every antibiotics and for every species, show the number of found bacteria that are resistant (sensitive or intermediate) to the antibiotics.

In order to count the number of infections, the "Epidemiological Observatory" analyses the data regarding the positive culture results of a particular time period (3 or 6 months). Every identified bacterium compared with the other bacteria found on the same patient in the previous N days (usually N is 30). The bacterium is counted as an infection provided that:

1. its species is different from that of the others;
2. its strain is different from that of bacteria of the same species previously found on the patient.

This is because, in case the strain is the same, the new bacterium is considered as a mutation of the previous one rather than a new infection.

In order to detect when two bacteria belong to the same strain, Italab C/S uses a very simple difference function for computing the percentage of antibiotics in the antibiogram having different values for the two bacteria. If this percentage is below a user defined threshold (usually 30%), then they belong to the same strain.

However, this approach for detecting when two bacteria belong

[1] D.E.I.S., Università di Bologna, Italy, e-mail: pmello@deis.unibo.it
[2] DIANOEMA S.p.A. Via Carracci 93, 40100 Bologna, Italy
[3] Dipartimento di Ingegneria, Università di Ferrara, Italy
e-mail: {mpiccardi, friguzzi}@ing.unife.it
[4] Ospedale Molinette (San Giovanni Battista), Corso Bramante 88/90, 10134 Torino, Italy

to the same strain is quite rough: it is not universally accepted by microbiologist and does not seem to work in all possible situations (different hospitals, different units within a hospital).

In order to improve the accuracy of the system in recognising strain membership, we defined, helped by microbiologists, a new strain membership criteria.

The first step consists in identifying all existing strains in a target hospital. In some cases, strain descriptions can be provided by the microbiologist, in other cases this is not possible and clustering is applied to all the antibiograms found in the past for every bacterium species. Each cluster found is considered as a strain and its description is stored by the system.

A new bacterium is considered as a new infection provided that no bacterium of the same species and strain is found in the same patient in the previous N days. The new bacteria is classified as belonging to a strain by using a membership function that depends on the strain description used.

In order to find bacterial strain, the clustering algorithm is executed on data regarding the positive cultures (only bacterial specie and relative antibiogram) of a large period of time (ex. 12 months) that have been found at the hospital where the system will be installed.

Applying clustering to find bacterial strain is useful also because it can be useful for giving the microbiologist new insights about the hospital population of bacteria and their resistance to antibiotics.

In order to test this approach for strain identification, we have performed a number of prototypical clustering experiments on data from various bacterial species. In this experimental phase we have used Intelligent Miner by IBM [3] for its free availability to academic institutions and its powerful graphical interface. However, clustering in final system will be performed by special purpose code.

2 THE DEMOGRAPHIC CLUSTERING ALGORITHM

The demographic clustering algorithm that is enclosed in Intelligent Miner [1] builds the clusters by comparing each record with all clusters previously created and by assigning the record to the cluster that maximizes a similarity score. New clusters can be created throughout this process.

The similarity score of two records is based on a voting principle, called Condorset [1]. The distance is computed by comparing the values of each field, assigning it a vote and then summing up the votes over all the fields. For categorical attributes, the votes are computed in this way: if the two records have the same value for the attribute, it gets a vote of +1, otherwise it gets a vote of –1. For numerical attributes, a tolerance interval is established and the vote is now continuous and varies from -1 to 1: -1 indicates values far apart, 1 indicates identical values and 0 indicates that the values are separated exactly by the tolerance interval. The overall score is computed as the sum of the score for each attribute.

In order to assign a record to a cluster, its similarity score with all the clusters is computed. To this purpose, the distribution of values of each field for the records in the cluster is calculated and recorded. The similarity between a record and a cluster is then computed by comparing the field values of the record with the value distribution of the cluster. In this way, it is not necessary to compare the record with each record in the cluster.

The algorithm assigns the record to the cluster with the highest similarity score. In case the score is negative for all clusters, then the record is a candidate for forming a new cluster. In this way, the number of clusters does not have to be known in advance but can be found during the computation.

Table 1. Modal values of the resistance for each cluster.

Cluster ®	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Dimension ®	339	1266	65	276	2	2	5	3	2
AMIKACINA	S	R	R	S	S	R	S	S	R
AMOXI_A_ CLAVULANIC	S	R	S	R	S	R	S	R	R
AMOXICILLINA	R	R	R	R	R	R	R	R	R
CEFAZOLINA	S	R	S	R	S	R	S	R	S
CEFOTAXIME	S	R	S	R	S	R	S	R	R
CEFUROXIME_ PARENTE	S	R	S	R	S	R	S	R	S
CIPROFLOXACINA	S	R	R	S	R	R	R	S	I
CLINDAMICINA	S	R	R	S	S	S	R	S	S
COTRIMOXAZOLO	S	R	R	S	R	S	S	R	R
DOXICICLINA	S	S	S	S	S	S	S	R	S
ERITROMICINA	S	R	R	S	R	R	R	S	R
GENTAMICINA	S	R	R	S	I	R	S	I	R
IMIPENEM	S	R	S	R	S	R	S	R	S
MEZLOCILLINA	R	R	R	R	-	R	S	-	-
NETILMICINA	S	R	R	S	S	R	S	S	R
OFLOXACINA	S	R	R	S	R	R	R	S	R
OXACILLINA	S	R	S	R	S	S	S	R	S
PEFLOXACINA	S	R	R	S	R	R	R	S	R
PENICILLINA_G	R	R	R	R	R	R	R	R	R
RIFAMPICINA	S	S	S	S	R	R	R	S	S
TEICOPLANINA	S	S	S	S	S	S	S	S	S
TIAMFENICOLO	S	S	S	S	S	S	S	R	S
VANCOMICINA	S	S	S	S	S	S	S	S	S
Resistance level	14,7	69,4	44,8	44,2	20,8	50,9	32,7	51,4	47,9

This process is repeated a fixed number of times ("phases") and clusters are updated until either the maximum number of phases is reached or the maximum number of clusters is achieved or the clusters centres do not change significantly as measured by a user-determined margin.

3 RESULTS

We have considered all the bacteria belonging to the species Staphilococcus Epidermidis. The dataset contains 1961 records having the attributes described in section 1. They have been collected from the 5th of March 1997 to the 20th of November 1999 at Le Molinette Hospital in Turin, Italy.

As in the PTAH system [2], an additional feature was computed for each record: the level of resistance, that represents the percentage of antibiotics for which the bacterium was resistant over the total number of antibiotics whose resistance was known (R, S, I).

In this experiment, we have set the maximum number of phases to 3 and we have found 9 clusters with a global Condorset value of 0.843. The clustering has been performed by considering only the record fields relative to antibiotics resistance. Table 1 shows the modal values of antibiotics reaction in the 9 clusters. The second row shows the number of elements of the cluster and the last the average resistance level of the cluster.

Cluster 1 is the biggest and is the one with the highest level of resistance (average of 69.4 %).

Figure 1 shows the resistance level to antibiotics of cluster 1: in each pie-chart, the internal pie is referred to the cluster, while the external ring is referred to the overall dataset. From figure 1 we can observe that in cluster 1 the percentage of resistant bacteria is higher for all antibiotics with respect to the complete dataset except for Doxiciclina for which the percentage of sensitive bacteria is higher. Cluster 0 has the same behaviour with R substituted for S: Doxiciclina is the only antibiotics for which the percentage of resistant bacteria is higher.

Clusters 3 and 2 are characterised by values of the resistance level that are intermediate between those of cluster 1 and 0. Clusters 4, 5, 6, 7 and 8 contain few elements and this means that some antibiograms are significantly different from all the others.

On the basis of these results, some comments can be made. We expected that the majority of bacteria from the same species had similar behaviour and that, more rarely, we could find "abnormal" bacteria that had become more resistant. On the contrary, by clustering the Staphylococcus Epidermidis bacteria, we have found that the majority of bacteria is highly resistant and that rarer cases are characterised by a higher sensitivity to antibiotics. This is probably due to the nature of this bacterium. In fact, another clustering experiment performed over *Escherichia Coli* bacteria has shown that bigger clusters have a lower resistance level and smaller cluster have a higher resistance.

Clustering of the antibiograms was performed as well in the PTAH system [1]. In PTAH, the clusters are hierarchically organised: low level clusters are grouped into higher level cluster and so on, up to the root cluster that contains all the data. The hierarchy enables the user to study the clusters at different levels of granularity. In this way it is possible to discover the different types of resistance vectors and to evaluate their frequency.

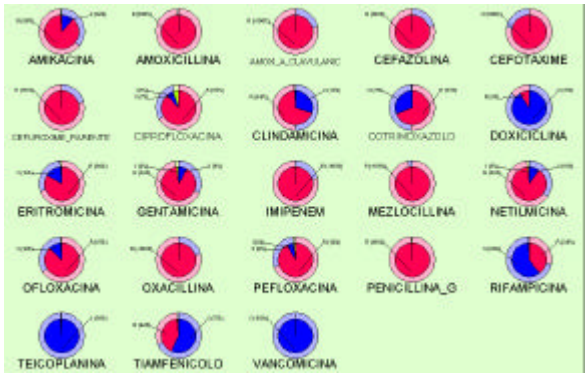


Figure 1. resistance to antibiotics in cluster 1.

We owe to PTAH a number of inspiring ideas, first of all the introduction of the resistance level variable for a bacterium that is very useful for providing an indication of the dangerousness of bacteria, and also the clustering of bacteria. However, we do not use hierarchical clustering as PTAH does: this is due to the fact that the results here presented are obtained from a first study. In the future we plan to adopt as well a hierarchical clustering algorithm because we think that the results will probably be easier to be interpreted by a medical doctor.

ACKNOWLEDGEMENTS

We are grateful to Dr. Furlini (S.Orsola Malpighi Hospital, Bologna) and Dr. Andollina (Officine Ortopediche Rizzoli, Bologna) for helpful discussions. This work has been partially supported by DIANOEMA S.p.A., Bologna, Italy and by the MURST project ``Intelligent Agents: Interaction and Knowledge Acquisition".

4 BIBLIOGRAPHY

[1] Cabena, Hadjinian, Stadler, Verhees, Zanasi, "Discovering Data Mining – from concept to implementation", Prentice Hall – IBM

[2] M. Bohanec, M. Rems, S. Slavec, B. Urh, "PTAH: A system for supporting nosocomial infection therapy", in N. Lavrac, E. Keravnou, B. Zupan (eds) "Intelligent Data Analysis in Medicine and Pharmacology", Kluwer Academic Publishers, 1997

[3] Intelligent Miner, http://www.software.ibm.com/data/iminer/fordata

Informal identification of outliers in medical data

Jorma Laurikkala[1], **Martti Juhola**[1] and **Erna Kentala**[2]

Abstract. Informal box plot identification of outliers in real-world medical data was studied. Box plots were used to detect univariate outliers directly whereas the box plotted Mahalanobis distances identified multivariate outliers. Vertigo and female urinary incontinence data were used in the tests. The removal of outliers increased the descriptive classification accuracy of discriminant analysis functions and nearest neighbour method, while the predictive ability of these methods reduced somewhat. Outliers were also evaluated subjectively by expert physicians, who found most of the multivariate outliers to truly be outliers in their area. The experts sometimes disagreed with the method on univariate outliers. This happened, for example, in heterogeneous diagnostic groups where also extreme values are natural. The informal method may be used for straightforward identification of suspicious data or as a tool to collect abnormal cases for an in-depth analysis.

1 INTRODUCTION

There are many definitions for outliers which differ in words [1-3]. We use the one of Barnett and Lewis [1 pp. 4], who defined an outlier in a set of data to be "*an observation (or subsets of observations) which appears to be inconsistent with the remainder of that set of data"*. This type of observations are often a problem in statistical analysis where outliers may especially lower the model fit. To illustrate, consider fitting a linear regression equation to data shown in Figure 1 [1 pp. 261-263]. The regression line runs nicely through the scatter plot of ages and salaries of electrical engineers (*N*=55, United Kingdom, 1974), but the most extreme observation M has a surprisingly strong impact on the analysis. Dropping M and refitting the regression line elevates the goodness of fit statistics R^2 from 0.452 to 0.526.

There are various origins of outliers. Human error often produces unintentional outliers. Data entry may be incorrect and missing value codes are sometimes used as real data. Outliers are frequently generated as the result of the natural variation of population or process one cannot control. These outliers are from the intended population, but their values are unusual in comparison with the normal values. It is also possible to have an outlier that is not a member of population due to a sampling error [1,3].

Machine learning researchers often use the concept of noise rather than that of outliers. Noise is defined as mislabeled examples (class noise) or errors in the values of attributes (attribute noise) [4 pp. 92]. Outlier is, therefore, a broader concept which includes errors, as well as discordant data produced by the natural variation of population [5 pp. 175]. Examples with class noise are outliers produced by sampling error [6], while attribute noise may or may not show in the data as outliers.

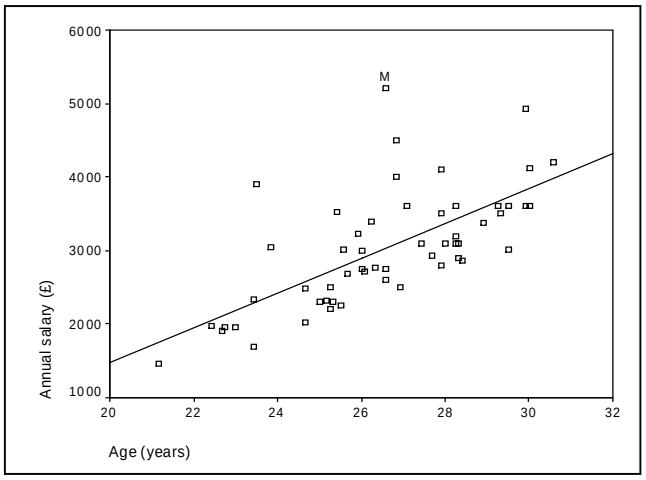


Figure 1. Ages and salaries of electrical engineers (UK, 1974)

Outlier identification (and consequent removal or accommodation) is a part of the data screening process which should be done routinely before statistical analyses [2,3]. The simplest and the most researched case is the identification of univariate outliers, where the distribution of a single variable is examined [1]. Extreme data values are obvious outlier candidates. When the distribution is symmetric, we suspect that candidate outliers are the extremes of the left or right tail. Correspondingly, the identified outliers are referred to as the lower and upper univariate outliers. In a skewed distribution, the suspect outliers are likely to be the extremes of the longer tail (see Figure 2). Multivariate outlier detection is more difficult, because the multivariate distribution has no tails [1,7]. Multivariate outliers, such as engineer M, are sometimes also univariate outliers. However, multivariate outliers are not necessarily univariate outliers, because unusual combinations of normal values may cause the case to be a multivariate outlier.

Filtering examples before the analysis seems to be a less studied area in the machine learning [8]. Brodley *et al*. eliminated outliers (mislabeled examples) from the training data by using ensemble filters, before passing the data to the final learning algorithm [6]. Wilson filtered the examples misclassified by nearest neighbour classifier (k = 3) to another nearest neighbour classifier (k = 1) [9]. Majority of the machine learning methods deals with irrelevant examples within the algorithm itself (embedded ap-

[1] Department of Computer and Information Sciences, FIN-33014 University of Tampere, Finland, email: jpl@cs.uta.fi
[2] Department of Otorhinolaryngology, FIN-00029 Helsinki University Central Hospital, Helsinki, Finland

proach) [e.g. 5] or apply some suitable method as a sub-routine during the learning (wrapper approach) [8].

We research machine learning methods, such as genetic algorithms and decision trees, in the context of descriptive and predictive analysis of medical data [10-12]. These methods seem to be quite robust and, therefore, they perform well with the data containing missing values and outliers [10-12]. Identification of outliers has recently begun to interest us for two reasons. Firstly, we consider balancing the imbalanced class distribution [e.g. 13] by reducing the largest classes before analysis. Outliers of the major classes seem to be worthwhile candidates for removal. In this line of work, the outliers are treated as poor data which may be removed without further analysis. Secondly, during the enlargement of the vertigo data [11,12], we noticed that the outliers may give us some additional insight of data. Outliers are not outright dropped from data, instead they are presented to expert physicians for further consideration.

In this paper, we identify both univariate and multivariate outliers with box plots [14] which are an informal method for outlier detection. The functioning of the informal method was tested in this work with two medical data sets. The results were evaluated objectively by performing discriminant analysis and nearest neighbour classification for the reduced data. Subjective evaluation was done by the expert physicians who studied the outliers manually.

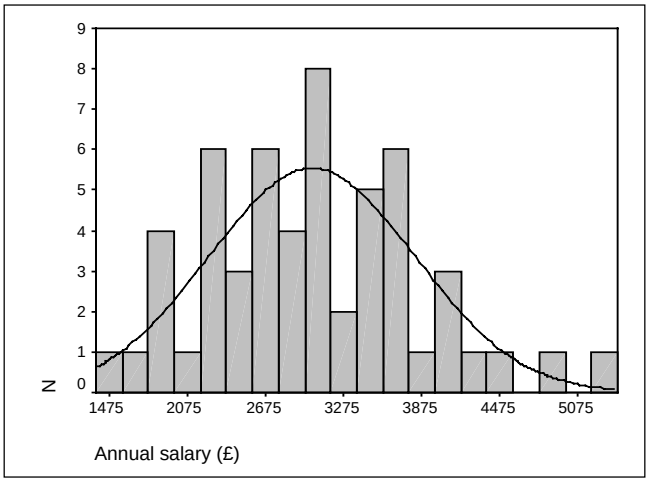


Figure 2. A histogram for annual salary

2 METHODS

Test of discordancy, formal or informal, is needed to declare extreme values as outliers. Formal testing requires a test statistic, which usually assumes some well-behaving distribution, on the basis of which the extremes are possibly declared outliers. Most of the test statistics, for example many Dixon-type tests, are designed to identify a single univariate outlier or an outlier pair using a normal distribution [1]. Unfortunately, the medical data sets we study are problematic in the statistical point of view. The data sets may be mixed, i.e. they contain both quantitative and qualitative variables, and the distribution of the continuous variables is frequently skewed or non-normal. Application of various test statistics would require identification of the distributions, transformations and possibly estimation of distribution parameters. For large data sets this process would be very difficult and tedious. Considering the practical aims of our research, we decided to test discordancy informally using regular box plots [14].

2.1 Box plot outlier identification

The box plot is a well-known simple display of the five-number summary (lower extreme, lower quartile, median, upper quartile, upper extreme) [14]. Box plots are most suitable for exploring both symmetric and skewed quantitative data, but they can also identify infrequent values from categorical data. Unlike in the quick box plot, the extremes of the box plots are not the smallest and largest data values, but the most extreme data values that are not extreme enough to be considered outliers [14]. Figure 3 shows a box plot for the salaries of the electrical engineers discussed earlier.

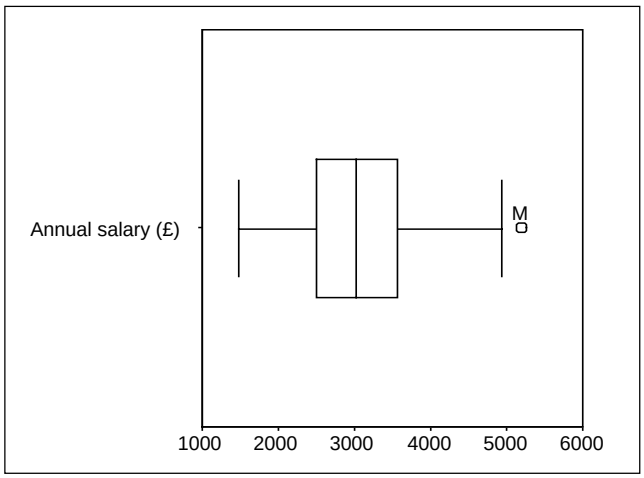


Figure 3. A box plot for annual salary

The thresholds for lower and upper outliers are defined as follows: lower threshold = lower quartile - step and upper threshold = upper quartile + step. Step is 1.5 times the interquartile range (upper quartile - lower quartile) which contains 50% of the data. Value x is a lower outlier, if x < lower threshold and an upper outlier, if x > upper threshold. Box plot identifies engineer M's salary as an upper univariate outlier (see Figure 3).

2.2 Univariate outliers

Univariate outliers were indentified for each variable within classes. Unfortunately, when there is a large number of univariate outliers, removal of all the identified outliers may cause a large portion of data to be excluded. For this reason, we decided to rank the non-multivariate outliers according to the frequencies of univariate outlier values in these examples and to discard (or trim) examples with the highest frequencies. Since often a small fraction of the extreme values are trimmed [1,7,15], we decided to trim 10% of the worst examples within each class.

2.3 Multivariate outliers

Methods for identifying univariate outliers are based on unarguable order of data values. For example, in the box plot method salaries are sorted in ascending order and, on the basis of the order, extremes, quartiles and outliers can be found. There is no unambiguous total ordering for N multivariate observations, but different sub-orderings have been suggested [1,16], of which the reduced sub-ordering is the most often used in the outlier study [1].

Reduced sub-ordering is established in two phases [1,16]. Firstly, a set of scalars $R = \{ r_i \}$ (i=1,...,N) is produced by transforming each multivariate observation $\mathbf{x}_i$ into a scalar r_i. Then, R is sorted to produce the actual ordering of the multivariate data. The transformation is often done with a distance metric [16] and, therefore, the extremes are those multivariate observations associated with the largest values in R.

We used in this study sub-ordering based on the generalised distance metric [1,7,16]

$$r^2_i = (\mathbf{x}_i - \mathbf{x}_0)' \, \mathbf{\Gamma}^{-1} (\mathbf{x}_i - \mathbf{x}_0), \quad (1)$$

where $\mathbf{x}_0$ indicates the location of the data set and $\mathbf{\Gamma}^{-1}$ weights variables inversely to their scatter. Different choices of these parameters result in different distance metrics. Fox example, when $\mathbf{\Gamma}$ is the identity matrix $\mathbf{I}$, (1) defines the Euclidean distance of $\mathbf{x}_i$ to the location of the data set.

We chose to use Mahalanobis distance [17,18] in the multivariate outlier identification. Mahalanobis distance is obtained from (1) by selecting $\mathbf{\Gamma}$ to be the population covariance matrix $\mathbf{\Sigma}$. As usual, the population mean $\boldsymbol{\mu}$ was used as the location parameter [1,3,7]. Often the population values are unknown and they are estimated with sample mean vector $\mathbf{m}$ and sample covariance matrix $\mathbf{S}$

$$r^2_i = (\mathbf{x}_i - \mathbf{m})' \, \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{m}). \quad (2)$$

Mahalanobis distance incorporates the dependencies between the attributes. This property is essential in multivariate outlier identification, where the goal is to detect unusual value combinations. Many distance metrics, including Euclidean distance, utilise only location information and are, therefore, unsuitable for this task. Another advantage of Mahalanobis distance is that the unit of variable has no influence on the distance, because each variable is standardised to mean of zero and variance of one [17,18].

Gamma-type probability plots are useful for informal outlier detection with generalised distances [1 pp. 274-275,7]. These graphical displays are produced by plotting the ordered reduced univariate measures r_i against the quantiles of a gamma distribution. If the multivariate observations are from a normal distribution, then the reduced measures follow approximately the gamma distribution. As a result, the points should cluster around a straight line and points that lie clearly off the linear relationship are considered to be outliers.

However, we applied again box plots, because we cannot assume that multivariate observations come from a normal distribution. Also, the gamma probability plots require a human, who must evaluate whether the anomalous points are really outliers. Box plots use objective rules for outlier identification.

3 MATERIALS

Outliers were searched from two medical data sets. The female urinary incontinence data (see Table 1) was collected retrospectively in the Department of Obstetrics and Gynaecology of Kuopio University Hospital, Finland [10]. The examples are described with 16 variables of which 7 are binary and 9 quantitative. Two variables (uroflowmetry and cystoscopy) were dropped from the analysis, because they had extremely high missing value rates.

The vertigo data (see Table 2) was collected in the vestibular unit of the Helsinki University Central Hospital, Finland. The patients, referred to the vestibular laboratory, filled out a questionnaire concerning their symptoms, earlier diseases, accidents, use of medicine, tobacco and alcohol [19]. The information was stored in the patient database of the expert system ONE [19]. The diagnoses were confirmed by an experienced specialist in the field of otoneurology. In this study, we focused on the six largest patient groups with vertigo and used the 38 most important variables of all the 170 available variables [11,12]. The subset of variables consisted of 16 quantitative variables, 10 ordinal variables and 12 nominal variables, 11 of which were dichotomous.

The missing values were replaced in both data sets with modes (nominal variables), medians (ordinal variables) and means within diagnostic classes. The imputed values of discrete variables were rounded.

4 EXPERIMENTAL SETUP

Multivariate and univariate outliers were identified separately with box plots by each diagnostic class, as usual [3]. Nearest neighbour classification (k = 1) with the heterogeneous value difference metric [20] and discriminant analysis were used for the objective evaluation. These methods were selected, because they are classical methods for classification in the areas of machine learning and statistics, respectively.

Four versions of both the medical data sets were evaluated: the original data A_0 ($|A_0| = N$) and three reduced data sets A_1, A_2 and A_3. Reduced sets were created by excluding 1) the multivariate outliers ($|A_1| = N - N_m$), 2) the multivariate outliers and 10% of the non-multivariate outlier examples with the most univariate outlier values ($|A_2| = N - N_m - N_u$) and 3) a random sample from the original data. The random removal was a baseline method where the number of removed examples was the same as the number of outliers excluded from the original data to produce the data set A_2.

Outlier removal was studied from two viewpoints. Firstly, we considered the descriptive analysis where statistical or machine learning methods are used to model the data. The goal is to find a reasonably simple model which fits the data well. This was studied by classifying data sets A_i with nearest neighbour method and discriminant analysis. Secondly, we studied the predictive analysis in which the aim is to build a model for accurate classification of new data which contains outliers. To assess the predictive ability of the models, the original data sets A_0 were classified with the nearest neighbour method, using the reduced data sets A_1, A_2 and A_3 as classifiers, and with the discriminant functions obtained from the reduced data sets.

The effect of removing the outliers was measured with the classification accuracy of the nearest neighbour classifiers and discriminant analysis functions. Classification accuracy ACC in per cents is $ACC = 100\% \, N_c / N$, where N_c is the number of cor-

rectly classified examples and N is the number of all examples in the data set. Nearest neighbour classification was performed with 10-fold cross-validation which was repeated 3 times, while discriminant analysis was not cross-validated. Therefore, the accuracies of nearest neighbour method were more realistic than those of the discriminant analysis.

5 RESULTS

Tables 1 and 2 show the frequencies of the multivariate outliers (N_m) and examples with the most univariate outliers (N_u) identified from the female urinary incontinence and vertigo data sets, respectively. The tables also show the sizes of diagnostic groups and, for the comparison with the number of outliers, the absolute frequencies corresponding to the 10% portion of each diagnostic group ($N_{10\%}$). The outlier frequencies behaved as expected. The largest classes had the highest number of outliers and the overall number of outliers was reasonably small in both the data sets.

Table 1. Frequencies of the original female urinary incontinence data (N) and its outliers by the diagnostic classes (N_m = multivariate outliers, N_u = examples with the most univariate outlier values).

Diagnosis	Original data		Outliers		
	N	$N_{10\%}$	N_m	N_u	Sum
Stress	323	32	29	20	49
Mixed	140	14	9	9	18
Sensory urge	33	3	2	2	4
Motor urge	15	2	0	1	1
Normal	18	2	0	1	1
Sum	529	53	40	33	73

Table 2. Frequencies of the original vertigo data (N) and its outliers by the diagnostic classes (N_m = multivariate outliers, N_u = examples with the most univariate outlier values).

Diagnosis	Original data		Outliers		
	N	$N_{10\%}$	N_m	N_u	Sum
Vestibular schwannoma	128	13	5	9	14
Benign positional vertigo	59	6	2	4	6
Meniere's disease	243	24	13	14	27
Sudden deafness	21	2	1	1	2
Traumatic vertigo	53	5	1	4	5
Vestibular neuritis	60	6	0	4	4
Sum	564	56	22	36	58

The descriptive accuracies of nearest neighbour method and the discriminant analysis functions are reported in Table 3. There was a clear improvement in the classification ability of both methods in the female urinary incontinence data. Also, removal of a randomly selected sample produced less accurate results, than excluding the identified outliers. The removal of outliers helped the classification of the vertigo data only slightly.

Table 3. Descriptive accuracies of the nearest neighbour method (NN) and discriminant analysis functions (DA) in different versions of the data.

Version of data	Accuracy (%)			
	Incontinence		Vertigo	
	NN	DA	NN	DA
Original (A_0)	83.5	84.3	90.9	94.3
No multivariate outliers (A_1)	85.9	86.7	92.7	94.6
No multi- and univarite outliers (A_2)	87.6	89.5	91.2	94.3
Random sample removed (A_3)	82.1	84.2	90.9	95.1

Table 4 shows the prediction accuracies of the two methods. Reduced data set A_2, from which the multivariate outliers and the examples with the most univariate outlier values were removed, was the worst nearest neighbour classifier. Also, discriminant functions produced from this data classified the original data with the lowest prediction accuracy.

Table 4. Prediction accuracies of the nearest neighbour method (NN) and discriminant analysis functions (DA) in the classification of original data.

Version of data	Accuracy (%)			
	Incontinence		Vertigo	
	NN	DA	NN	DA
No multivariate outliers (A_1)	96.2	84.3	98.9	94.0
No multi- and univarite outliers (A_2)	93.6	83.4	97.9	92.6
Random sample removed (A_3)	98.3	83.9	98.6	94.9

6 DISCUSSION

Outlier identification was studied with informal box plot method using as the test material two real-world data sets. There were two motivations for the identification. Firstly, outliers can be considered to be suspicious data whose removal before applying inductive machine learning methods is reasonable. Especially, dropping of the outliers of the largest classes balances the class distribution and, consequently, makes the classification of the members of the smaller classes easier [13]. Secondly, the knowledge carried by the outliers may be valuable for the domain experts, who may gain additional insight into the data by examining them.

The suitability of the method for the straightforward data reduction was studied objectively with discriminant analysis and nearest neighbour method which are well-known classification methods in statistics and machine learning. Removal of the identified outliers from the female urinary incontinence data improved clearly the classification ability of discriminant analysis functions and nearest neighbour method (see Table 3). The descriptive accuracy of the discriminant functions improved from 84.3% to 89.5% when both multivariate outliers and examples with the largest numbers of univariate outliers (14% of the data) were removed from the original data set. The prediction accuracy of the nearest neighbour classifier raised from 83.5% to 87.6%. However, the improvement was only marginal in the vertigo data set.

The most probable explanation for the differences is the characteristics of the data sets. All the female urinary incontinence data was collected retrospectively from the patient records [10], while the main body of vertigo data was obtained carefully in prospective fashion [19]. In addition, the vertiginous patients used in this study were selected to meet the definitions of the six diagnostic classes [19]. As a result, there were not much improvement left in the classification of this data. The earlier machine learning experiments [11,12] lend additional support for this conclusion. The best results were obtained with the decision trees [12] whose classification accuracy ranged from 94% (Meniere's disease) to 100% (benign positional vertigo and vestibular neuritis).

The prediction accuracy behaved oppositely to descriptive accuracy. Exclusion of outliers lowered the prediction accuracies. Again, the effect was stronger in the female urinary incontinence data than in the vertigo data. Nearest neighbour classification suffered from the outlier exclusion clearly in the incontinence data. This result suggest that the removed data were indeed outliers. Nearest neighbour classification should be more difficult, when

data having unusual examples is classified with cleaned data from which these examples have been removed. In addition, Quinlan [4 pp. 96] has shown experimentally that, for higher levels of attribute noise, a decision tree built from cleaned data is inferior to a tree built from the noisy data, when the data to be classified has same noise level. Discriminant analysis is in comparison with nearest neighbour method a highly advanced tool which produces more robust classifiers. For this reason, the prediction accuracies of the discriminant functions remained close to the descriptive accuracies of the original data set as in [15].

The identified univariate and multivariate outliers were presented to the expert physicians, who evaluated the suspect data to decide whether it was truly outlying. In their opinion, most of the multivariate outliers where abnormal cases. For example, in one case the post voiding residuals and urgency score were abnormally high for a stress-incontinent woman. However, closer examination revealed that the diagnosis was correct, because she had to drink excessively due to bowel problems. The experts sometimes disagreed with the box plot method on univariate outliers. The most frequent reason was the natural variation in diagnostic parameters between patients. The upper or lower outliers were extreme, but yet reasonable values for a parameter in a particular diagnostic class. This happened, especially in heterogeneous diagnostic groups, where also extreme values are natural. Vestibular schwannoma and Meniere's disease are examples of the heterogeneous diagnostic groups [19]. Both diseases worsen during the time and the extreme values come often from the patients who have had these diseases for a long time [19].

The experimental results suggest that box plots can be used for data reduction, but the benefit obtained of excluding outliers is data set dependent. Outlier removal helps the descriptive analysis, but the predictive analysis, i.e. classification of unseen cases, may suffer. Therefore, the usefulness of the method depends also the final goal of the analysis. The behaviour of predictive analysis needs to be studied further, because machine learning applications are often used to classify new data. The subjective evaluation by the experts gave controversial, but sound, results. There were real abnormalities in the multivariate outliers and the disagreement on the univariate outliers resulted from the method, which does not utilise any prior knowledge in the outlier identification.

The major limitation of this work is the use of the informal box plot method. Discordancy test statistics identify outliers on the basis of the solid theory. Unfortunately, these methods make many assumptions which should be met in order the tests to be applicable. Therefore, one can also argue that a well-known and widely used informal method may be a more appropriate choice for large practical applications, where these assumption are not usually fulfilled. However, the future work should also address the formal test statistics. There is also two limitation in the multivariate outlier identification with the Mahalanobis distance. Firstly, the Mahalanobis distance works best with quantitative normally distributed data [18]. Secondly, missing values must be treated before distance computation. Possibly some type of heterogeneous distance function [20] could address these problems.

REFERENCES

[1] V. Barnett and T. Lewis, *Outliers in Statistical Data*, John Wiley & Sons, Norwich, 2nd edn., 1987.

[2] R.J. Beckman and R.D Cook, 'Outliers', *Technometrics*, **25**, 119-149, (1983).

[3] B.G. Tabachnick and L.S. Fidell, *Using Multivariate Statistics*, HarperCollins, New York, 1996.

[4] J.R. Quinlan, 'Induction of decision trees', *Machine Learning*, **1**, 81-106, (1986).

[5] G.H. John, *Robust Decision Trees: Removing Outliers from Databases*, 174-179, Proceedings of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1995.

[6] C.E. Brodley and M.A. Friedl, *Identifying and Eliminating Mislabeled Training Instances*, 799-805, Proceedings of the Thirteenth National Conference on Artificial Intelligence, AAAI Press, Menlo Park, 1996.

[7] R. Gnanadesikan and J.R. Kettenring, 'Robust estimates, residuals, and outlier detection with multiresponse data', *Biometrics*, **28**, 81-124, (1972).

[8] A.L. Blum and P. Langley, 'Selection of relevant features and examples in machine learning', *Artificial Intelligence*, **97**, 245-271, (1997).

[9] D.L. Wilson, 'Asymptotic properties of nearest neighbor rules using edited data', *IEEE Transactions on Systems, Man, and Cybernetics*, **2**, 408-421, (1972).

[10] J. Laurikkala and M. Juhola, 'A genetic-based machine learning system to discover the diagnostic rules for female urinary incontinence', *Computer Programs in Biomedicine*, **55**, 217-228, (1998).

[11] E. Kentala, J. Laurikkala, I. Pyykkö, and M. Juhola, 'Discovering diagnostic rules from a neurotologic database with genetic algorithms', *Annals of Otology Rhinology & Laryngology*, **108**, 948-954, (1999).

[12] E. Kentala, K. Viikki, I. Pyykkö, and M. Juhola, 'Production of diagnostic rules from a neurotologic database with decision trees', *Annals of Otology Rhinology & Laryngology*, **109**, 170-176, (2000).

[13] M. Kubat and S. Matwin, *Addressing the Curse of Imbalanced Training Sets: One-sided Selection*, 179-186, Proceedings of the Fourteenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, 1997.

[14] A.F. Siegel, *Statistics and Data Analysis: An Introduction*, Wiley, New York, 1988.

[15] E. Krusinska and J. Liebhart, 'The influence of outliers on discrimination of chronic obturative lung disease', *Methods of Information in Medicine*, **27**, 167-176, (1988).

[16] V. Barnett, 'The ordering of multivariate data (with Discussion)', *Journal of the Royal Statistical Society A*, **139**, 318-354, (1976).

[17] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, New Jersey, 1988.

[18] J. Boberg, *Cluster Analysis: A Mathematical Approach with Applications to Protein Structures*, Academic dissertation, Turku Centre for Computer Science, Turku, Finland, 1999.

[19] E. Kentala, *A Neurotological Expert System for Vertigo and Characteristics of Six Otologic Diseases Involving Vertigo*, MD thesis, University of Helsinki, Helsinki, Finland, 1996.

[20] D.R. Wilson and T.R. Martinez, 'Improved heterogeneous distance functions', *Journal of Artificial Intelligence Research*, **6**, 1-34, (1997).

Enhancement of Learning by Declarative Expert-based Models

Peter Lucas
Department of Computing Science, University of Aberdeen
Aberdeen, AB24 3UE, Scotland, UK
E-mail: plucas@csd.abdn.ac.uk

Abstract. A major part of the knowledge in the medical field concerns diseases that are uncommon or even rare. Doctors, however, may face severe difficulties in handling such disorders, as they may not have sufficient experience with the management of these disorders in patients. Here there seems to be a clear role for medical decision-support systems. Unfortunately, the uncommon nature of these disorders renders it almost impossible to collect data from a sufficiently large number of patients as required for the development of models that faithfully reflect the subtleties of the domain. Often, one therefore resorts to the development of naive Bayesian models. However, under these unfavourable circumstances, it may still be feasible to design detailed model of the problem domain in collaboration with expert physicians. The advantage of such models, e.g. structured Bayesian-network models, is that they are often suitable for handling more than one task, e.g. not only predicting prognosis but also treatment selection. This raises the question whether such expert-based Bayesian models could incorporate enough structural background knowledge to improve on naive Bayesian models. In this paper, we discuss the development of several different Bayesian-network models of non-Hodgkin lymphoma of the stomach, which is an example of an uncommon disease. It is shown that a declarative, structured model, based on expert knowledge, can indeed outperform naive Bayesian models when supplied with probabilistic information based on data. The handling of missing values, and the checking of the stochastic independence structure are also discussed in the paper, as these are also important issues when dealing with small datasets.

Keywords & Phrases: Bayesian networks, machine learning, background knowledge, medical decision support.

1 Introduction

There is a great deal of experience in the medical field in analysing medical data of diseases with high prevalence, like breast cancer, lung cancer, and myocardial infarction. Most of the evidence underlying current medical practice is based on such analyses. The advantage linked with the frequent occurrence of those disorders is that it is practically feasible to collect data of large numbers of patients, thus making it possible to draw conclusions that are statistically significant. Typically, datasets collected in the context of such studies may include many thousands of patient records. Such datasets are quite attractive for evaluating particular machine-learning techniques, and, in fact, have been used for this purpose by many researchers.

However, for more than 90% of medical disorders, the picture is quite different: these disorders do only occur occasionally or rarely, and, as a consequence, even clinical research datasets may only include data of a hundred to a few hundred patient cases. Developing decision-support systems that assist clinicians in handling patients with these disorders is thus scientifically challenging, because there may not be sufficient data available. Furthermore, systems covering such disorders would be practically useful, as many doctors will lack knowledge and experience to deal with patients affected by those disorders effectively. Confronted with this situation when building a decision-support system, there is a clear place for using medical expert knowledge, as background knowledge, to compensate for lack of data.

Machine-learning literature in medicine not only tends to focus on problems for which much data is available, but in addition it focuses on models capable of performing single tasks, such as classifying patients in particular disease or prognostic categories to assist clinicians with the tasks of diagnosis or treatment selection. However, medical management is more complicated than that, and cannot be captured in terms of single, simple tasks. As a consequence there appears to be a mismatch between common task-specific computer-based models and the complexity of the field of medicine. We believe that it might be worthwhile to consider instead the development of declarative medical models, that can be used to explore different problems, and be reused for different tasks. One of the nice aspects of declarative models is that they can also be employed to look at particular problems from different angles, just by varying the supplied evidence and the questions posed to the model. Admittingly, developing such models will be more challenging, both in terms of required number of variables and probabilistic information, but the extra effort may be out-weighted by their greater potential. Yet, it is presently unclear what the potential and limitations of such declarative models are with respect to capturing knowledge from small datasets.

In this paper, we will try to find answers to a number questions related to the issues mentioned above, based on our experience in developing declarative prognostic models of non-Hodgkin lymphoma of the stomach. Non-Hodgkin lymphoma

of the stomach is a typical example of an uncommon, although not extremely rare disease. Here we focus on investigating the consequences of adopting different assumptions underlying Bayesian-network technology, in particular with respect to structure, number of variables included in a model and dealing with missing values. The following issues are addressed:

- Does a declarative model enhance the learning of knowledge?
- Is it worthwhile to handle missing values explicitly?
- Can the structure of a Bayesian network be learnt, either completely or partially, from the data of a small dataset?

The remainder of this paper is organised as follows. In the next two sections, the development of different Bayesian-network models of non-Hodgkin of the stomach is discussed. Sections 4 pays attention to their evaluation, whereas in Section 4 some results of checking the structure of a Bayesian network are presented. The paper is rounded-off by a comparison to related work and with a discussion of what has been achieved.

2 Preliminaries

A *Bayesian network* $\mathcal{B}$ is defined as a pair $\mathcal{B} = (G, \Pr)$, where G is a directed, acyclic graph $G = (V(G), A(G))$, with a set of vertices $V(G) = \{V_1, \ldots, V_n\}$, representing a set of stochastic variables $\mathcal{V}$, and a set of arcs $A(G) \subseteq V(G) \times V(G)$, representing conditional and unconditional stochastic independencies among the variables, modelled by absence of arcs among vertices [5, 6, 9]. On the variables $\mathcal{V}$ is defined a joint probability distribution $\Pr(V_1, \ldots, V_n)$, for which the following decomposition property holds:

$$\Pr(V_1, \ldots, V_n) = \prod_{i=1}^{n} \Pr(V_i \mid \pi(V_i))$$

where $\pi(V_i)$ denotes the conjunction of variables corresponding to the parents of V_i, for $i = 1, \ldots, n$. In the following, variables will be denoted by upper-case letters, e.g. V, whereas a variable V which takes on a value v, i.e. $V = v$, will be abbreviated to v. When it is not necessary to refer to specific values of variables, we will usually just refer to a variable, which thus stands for any value of the variable.

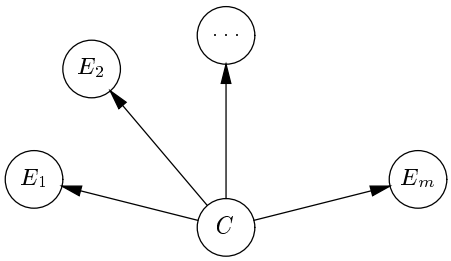


Figure 1. Independent-form Bayesian network.

The Bayesian-network models conforming to the topology shown in Figure 1 have been particularly popular in the statistical and machine-learning communities [8]. It corresponds to the situation where a distinction is made between evidence variables E_i and a class variable C, with the evidence variables assumed to be conditionally independent given the class variable. In the following such a network will be called an *independent-form Bayesian network*, by way of analogy with the special form of Bayes' rule, called its independent form, for which the same assumptions hold. This form of Bayes' rule is also known as the *naive Bayes' rule*. The independent form of Bayes' rule is used to compute the a posteriori probability of a class value c_k given the evidence $\mathcal{E}$ [6]:

$$\begin{aligned} \Pr(c_k \mid \mathcal{E}) &= \frac{\Pr(\mathcal{E} \mid c_k)\Pr(c_k)}{\Pr(\mathcal{E})} \\ &= \frac{\prod_{e \in \mathcal{E}} \Pr(e \mid c_k)\Pr(c_k)}{\sum_{j=1}^{q} \prod_{e \in \mathcal{E}} \Pr(e \mid c_j)\Pr(c_j)} \end{aligned}$$

where class variable C has q mutually exclusive values, and $\Pr(\mathcal{E}) > 0$. Note that

$$\Pr(\mathcal{E} \mid c_k) = \prod_{e \in \mathcal{E}} \Pr(e \mid c_k)$$

holds, because of the assumption that the evidence variables E_i are conditionally independent given the class variable C. Furthermore,

$$\Pr(\mathcal{E}) = \sum_{j=1}^{q} \Pr(\mathcal{E} \mid c_j)\Pr(c_j)$$

using marginalisation and conditioning.

Although an independent-form Bayesian network may ignore important probabilistic dependence information, it has the virtue that assessment of the required probabilities $\Pr(E_j|C)$ and $\Pr(C)$ is rather straightforward, and can be carried out with a relatively small dataset. Determination of the a posteriori probabilities $\Pr(C|\mathcal{E})$, where $\mathcal{E} \subseteq \{E_1, \ldots, E_m\}$, is computationally speaking a trivial task under the mentioned assumptions. Furthermore, it is quite straightforward to handle missing values with the independent form. For example, M. Ramoni and P. Sebastiani have developed an interval-based method, and implementation of it as the ROC system (Robust Bayesian Classifier), that is capable of dealing with missing values in a mathematically sound way [13]. These features of the independent form of Bayes' rule probably explain why it is again increasingly popular, having fallen into disgrace two decades ago.

An independent-form Bayesian network is especially suited for the classification of cases based on a set of known features $\mathcal{E}$. It is not meant for providing a description of the knowledge in a particular problem domain. In most cases, only those variables considered relevant for the classification task are included. Since relationships among features are also omitted, the result is relatively naive from the perspective of domain modelling, hence its nickname 'naive' Bayesian model. On the other hand, building a Bayesian network that models the problem domain more accurately would almost certainly require the elicitation of domain knowledge from human experts. The development of such expert-based models can be quite time-consuming. Hence, developing such models would only be worthwhile when collecting expert knowledge would counterbalance the lack of data in small datasets. The role of encoded human expertise in machine learning is a very relevant issue.

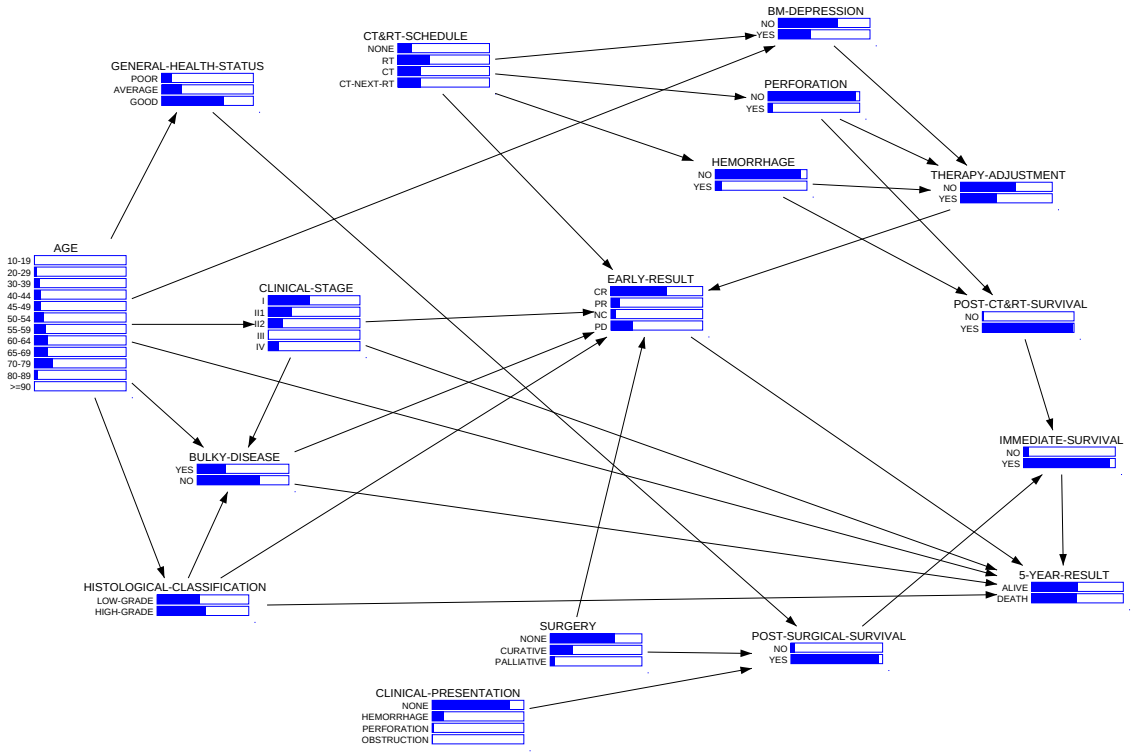


Figure 2. Bayesian-network model as designed with the help of medical experts.

In the following, we shall discuss a number of declarative and independent-form Bayesian models for non-Hodgkin lymphoma of the stomach, which sheds some light on the issues mentioned above.

3 Bayesian models of non-Hodgkin lymphoma of the stomach

A Bayesian model incorporating most factors relevant for the management of non-Hodgkin lymphoma of the stomach was developed in collaboration with clinical experts from the Netherlands Cancer Institute (NKI). The resulting model, shown in Figure 2, includes variables like age of the patient and clinical stage of the tumour (stage I is generally associated with good prognosis, whereas stage IV is generally associated with very poor prognosis). Some of the included variables concern patient information obtained from diagnostic procedures, which will be available prior to treatment selection. We shall refer to this information as *pretreatment* information. Another part of the model variables will only become available following treatment; examples are: EARLY-RESULT and 5-YEAR-RESULT. It is called the *posttreatment* part of the model. Finally, the model includes the variables SURGERY and CT&RT-SCHEDULE, which represent *treatment* variables with possible values: 'none', 'curative', 'palliative' for SURGERY, and 'none', 'chemotherapy', 'radiotherapy' or 'combination therapy' for CT&RT-SCHEDULE. Note that the declarative model of non-Hodgkin lymphoma of the stomach can indeed be used in the context of different tasks, such as prediction of prognosis, treatment selection – by comparing the likely outcomes of alternative treatment choices, possibly using preference information expressed as utilities – and generation of patient profiles [7].

The independent-form Bayesian network corresponding to the network shown in Figure 2 is depicted in Figure 3. Note that this model includes a single class variable: the posttreatment variable 5-YEAR-RESULT; in addition, all pretreatment and treatment variables are incorporated. The remainder of the posttreatment variables have been left out, since as these variables are always unknown for a patient, they are not immediately relevant for the prediction of 5-year survival in patients. This illustrates one difference between a declarative Bayesian model, as shown in Figure 2, and a task-specific model, as shown in Figure 3.

Taking the two topologies as a starting point, the probability distributions of the two models were learnt, adopting a number of different assumptions. The resulting models that will be discussed in this paper are briefly described in Table 1. Models with the letter S in their name are declarative or structured models; models with the letter I in their name are independent-form Bayesian networks. Learning took place using a dataset with patient data from the Netherlands Cancer Institute, comprising 137 cases, with some missing data. Missing data were either ignored (models S and I), distributed uniformly among the values of the variable for which a value was missing, or probability intervals instead of point probabilities were determined. The last approach thus views missing data as relaxing the constraints on a probability distribution. Model I_{ic1} was learnt starting with an initial uniform proba-

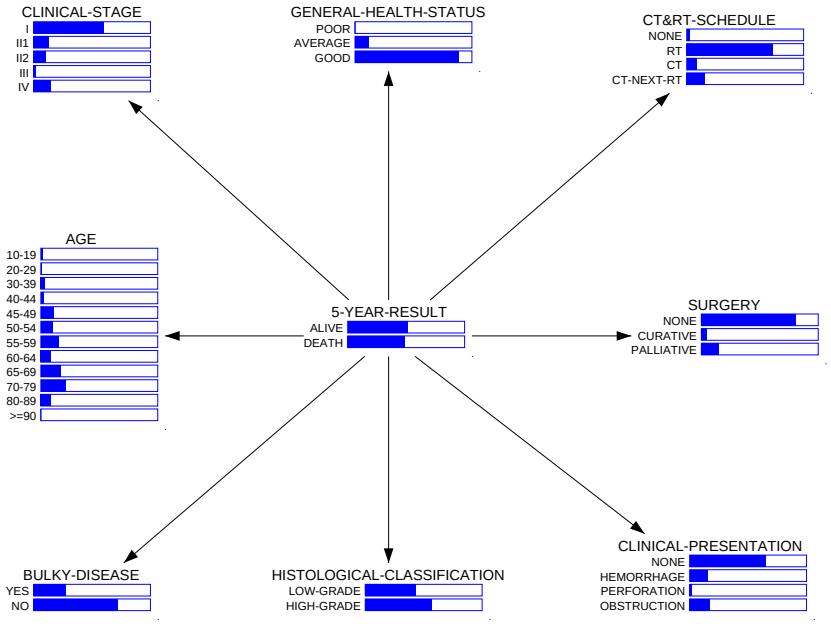


Figure 3. Independent-form Bayesian model.

Table 1. Bayesian models.

Model	Description	Topology	Missing Data
S_E	expert-assessed model	declarative	—
S	learnt model	declarative	ignored
S_u	learnt model	declarative	uniform distribution
I	learnt model	independent form	ignored
I_{ic1}	learnt model, sample size 22	independent form	interval calculus
I_{ic2}	learnt model, sample size 0	independent form	interval calculus
I_u	learnt model	independent form	uniform distribution

bility distribution, which was given an equivalence sample size of 22 cases. This means that the initial uniform probability distribution had a weight as if it was based on a sample of 22 cases. For model I_{ic2}, it was assumed that there were no such cases at all, i.e. no initial uniform distribution was assumed.

4 Evaluation and comparison

The Bayesian models were evaluated in a number of different ways. First, the a posteriori probability $\Pr(\text{5-YEAR-RESULT}|\mathcal{E})$ was computed, where $\mathcal{E}$ was the available evidence for each of the 137 patients with non-Hodgkin lymphoma of the stomach, restricted to values of all pretreatment and treatment variables. When $\Pr(\text{5-YEAR-RESULT} = \text{alive} \mid \mathcal{E}) > 0.5$, and the patient was known to have been alive more than 5 years following treatment, the model's prediction was considered correct; otherwise, it was classified as being incorrect. A similar decision procedure was used when the most likely prediction was that the patient would die within 5 years. No use was made of receiver-operator-characteristic curves [15], because this would not be entirely in line with the declarative nature of structured Bayesian networks.

For the two independent-form Bayesian networks learnt using the ROC system, the interval a posteriori probability distributions were interpreted using the *stochastic-dominance criterion* [13], i.e. it is assumed that the model predicts that class c is most likely if the minimum probability associated with this class value, i.e. $\Pr_{\min}(c \mid \mathcal{E})$, is larger than the max-

Table 2. Results for different Bayesian models. Percentages were computed by dividing the number of correct conclusions by the number of classified cases.

Model	Total	Classified (n) Incorrect	Correct	(%)	Unclassified (n)	Missing Data
S_E	137	43	94	(68.6)	—	—
S	137	22	115	(83.9)	—	—
S_u	137	20	117	(85.4)	—	—
I	137	40	97	(70.8)	—	—
I_{ic1}	137	28	98	(77.8)	11	stochastic dominance
	137	35	102	(74.5)	0	weak dominance
I_{ic2}	137	28	100	(78.1)	9	stochastic dominance
	137	32	105	(76.6)	0	weak dominance
I_u	137	39	98	(71.5)	—	—

imum probability associated with the other class values, i.e.

$$\Pr_{\min}(c \mid \mathcal{E}) > \max\{\Pr_{\max}(c' \mid \mathcal{E}) \mid c' \neq c\}$$

A disadvantage of the stochastic-dominance criterion is that often some cases remain unclassified. Therefore, a weaker criterion, called *weak dominance*, was used as well. This criterion associates a score $s_u(c \mid \mathcal{E})$ with each class value c; when C has two mutually exclusive class values, the score is defined as follows [13]:

$$s_u(c \mid \mathcal{E}) = \frac{\Pr_{\min}(c \mid \mathcal{E}) + \Pr_{\max}(c \mid \mathcal{E})}{2}$$

i.e. the interval midpoint is chosen. The class value with the highest score is selected. The results for the Bayesian-network models, as defined in Table 1, are given in Table 2.

A disadvantage of the straightforward method for comparing the quality of the Bayesian models, as described above, is that the actual a posteriori probabilities are ignored. A more precise impression of the behaviour of the Bayesian models would have been obtained if the resulting probabilities had been taken into account as well. For example, if patient Q was known to have survived more than 5 years following radiotherapy, and a Bayesian model had predicted this event with probability 0.8, this conclusion would seem intuitively better than the conclusion of another model which predicted this event with a probability of 0.6. A number of different scoring rules have been designed in the field of statistics that measure exactly this effect. One of the simplest scoring rules that can be given a statistical interpretation is the *logarithmic scoring rule* [2]. We shall briefly discuss this rule in the following.

Let D be a dataset, $|D| = p$, $p \geq 0$. With each prediction generated by a Bayesian model for case $r_k \in D$, with actual class value c_k, we associated a score:

$$S_k = -\ln \Pr(c_k \mid \mathcal{E})$$

which has the informal meaning of a penalty: when the probability $\Pr(c_k|\mathcal{E}) = 1$, then $S_k = 0$; otherwise, the score becomes rapidly larger than 0. The total score for an entire database D is now defined as follows:

$$S = \sum_{k=1}^{p} S_k$$

Since S is a stochastic quantity, it can be characterised further by means of central moments, such as the average E_k:

$$E_k = -\sum_{i=1}^{q} \Pr(c_i \mid \mathcal{E}) \ln \Pr(c_i \mid \mathcal{E})$$

with total average $E = \sum_{k=1}^{p} E_k$, and the variance V_k:

$$V_k = -\sum_{i=1}^{q} \Pr(c_i \mid \mathcal{E})(\ln \Pr(c_i \mid \mathcal{E}))^2 - E_k^2$$

with total variance $V = \sum_{k=1}^{p} V_k$.

The results obtained for five of the models are shown in Table 3.

Table 3. Logarithmic scores.

Name	S	E	V
S_E	81.28	71.47	30.65
S	57.05	65.93	25.65
S_u	48.92	58.88	25.91
I	67.50	51.31	28.20
I_u	69.50	54.07	31.99

The performance of the declarative model S_E, which incorporated expert-assessed probabilities, was lowest; the Bayesian networks S and S_u with the same topology as S_E, but with probabilities learnt from data, yielded the best results. It was not really surprising that model S_E was inferior to the other models, as its probability distribution was assessed taking into account recent changes in treatment policy as well as experience at other hospitals, as reported in the literature. In addition, some deviation of subjective probabilities from relative frequency information may always be expected to exist. In a sense, it was surprising that the performance of S_E was still near to that of model I.

The various independent-form Bayesian networks yielded results that were always below those of the two trained declarative models S and S_u. However, the results for the two independent-form models, where missing values were handled by means of interval calculus, were in turn better than those for the other two independent-form models. These results are consistent with previous results by M. Ramoni and P. Sebastiani [14], who also showed that using interval calculus may improve performance, although only to a slight extent when using the weak-dominance criterion. Stochastic dominance may leave difficult cases unclassified, which explains the

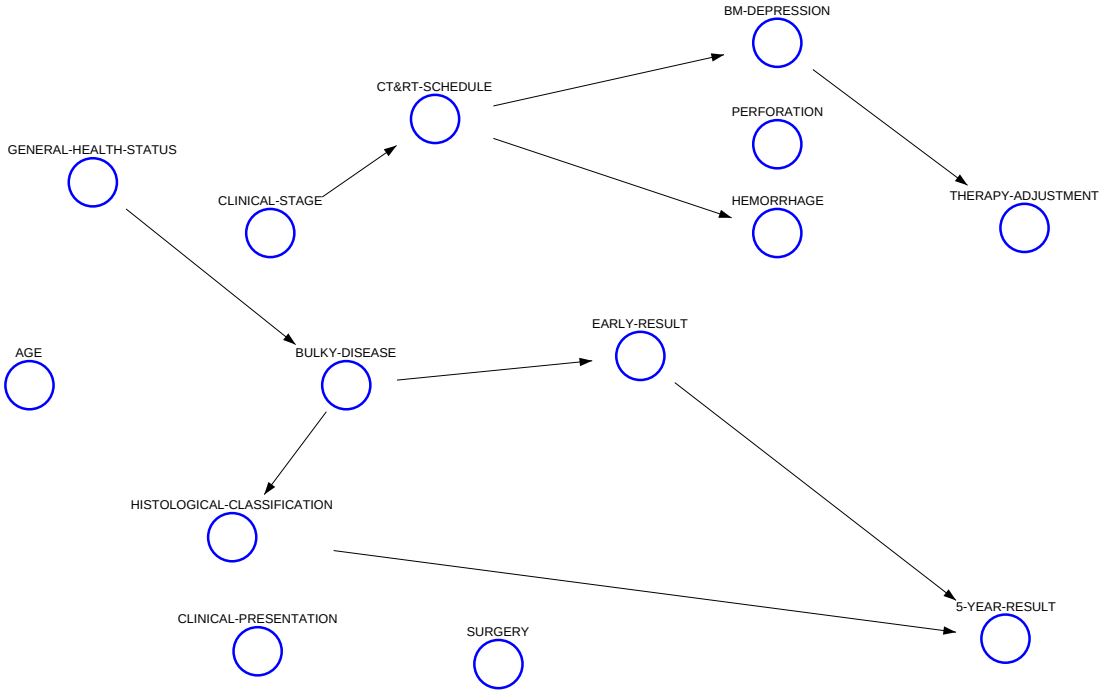


Figure 4. Bayesian-network structure as learnt using BKD.

better performance. However, this feature renders the latter criterion practically speaking less useful.

The Bayesian models in which missing values were uniformly distributed among values generally yielded slightly better results than the other models due to the artificial increase in sample size. The logarithmic scores shown in Table 3, however, indicate that this results in a slightly decreased accuracy of the a posterior probability for the independent-form Bayesian model; in contrast, the accuracy of the a posterior probabilities of the declarative model was improved. The latter effect is likely due to a reinforcement of the influence of the topology of the graph on the outcome; as the topology of the graph reflects clinical expertise, the quality of the results improves.

5 Checking the topology

Even though it is in principle possible to learn the topology of a Bayesian network from data, the dataset we had at our disposal on this occasion (a typical uncommon-disease dataset as mentioned in the Introduction) was just too small for this purpose. Nevertheless, experiments with a model (structure) selection algorithm were carried out in order to obtain more insight in how far one gets with such algorithms, given a small dataset. Use was made of the BKD (Bayesian Knowledge Discoverer) system, which implements a heuristic search method for finding the Bayesian-network structure S that best fits the data D according to the ratio $\Pr(D, S)/\Pr(D, S')$, where S' is an alternative structure [1, 11]. Furthermore, the system includes an algorithm for dealing with missing values, called *bound-and-collapse* [12]. This method first determines the bounds of a probability value using the data that is available, and finally collapses the set of values to a single probability using a convex combination of the extreme values of the set.

The resulting Bayesian network is shown in Figure 4. Only 6 of the 30 arcs in the expert-assessed network were predicted correctly; the arcs between the vertex GENERAL-HEALTH-STATUS and BULKY-DISEASE is clinically incorrect. The arc between HISTOLOGICAL-CLASSIFICATION and BULKY-DISEASE was reversed. Finally, the arc between the vertex CLINICAL-STAGE and CT&RT-SCHEDULE is correct, but left out in the original model, because it is assumed that therapy is always selected.

It is surprising that the algorithm was not able to find dependencies between the variable AGE and the other variables, as it is well known that many of the variables in the model are influenced by age. A similar observation holds for the variable CLINICAL-STAGE.

6 Discussion

The present paper is certainly not the first paper showing that human background knowledge may enhance machine learning. Earlier work in the areas of inductive logic programming (e.g. [4], but there are many other papers), and Bayesian networks [3], has demonstrated this. However, other researchers have primarily focused on the gathering of sufficient pieces of knowledge in order to enhance the learning of knowledge from data. In contrast, the present paper investigates the usefulness of extensive declarative models that were not developed for the purpose of learning in the first place. It appears that the effect of the topology of a Bayesian network as assessed by human experts, may be so strong that, even though a small

proportion of its probability tables are filled with probabilities obtained from data, a structured model still outperforms independent-form Bayesian networks. Note, furthermore, that is not true, as suggested by Pradhan et al. [10] that once given the structure of a Bayesian network, the incorporated probabilities are not relevant at all, as is demonstrated in this paper by the low performance of the expert-assessed network.

From this study, one can also conclude that dealing with missing values in a typical clinical research dataset may offer some advantages. In particular, it seems worthwhile to consider using interval calculus when standard data imputation techniques are expected to be unreliable. A clinical research database as used in this study will typically have relatively few missing values. Under these circumstances, the improvement resulting from dealing with missing values will only be moderate.

A limitation of the present study is that the resulting networks have not been evaluated using cross validation, which would be computationally quite intensive for the structured models. Given the techniques used – independent Bayes and a structured Bayesian network – it is clear that the declarative model is capable of better capturing the logic implicit in the data. It seems likely that this conclusion would stand further scrutiny.

Finally, although learning the structure of a Bayesian network from a small dataset is not feasible, checking its structure using one of the structure-learning algorithm might offer some insight. Unfortunately, when differences between the expert-derived and predicted topology cannot be explained clinically, there is no alternative than to stick to the original structure.

The results of this paper suggest that it may be worthwhile devoting even more time to the gathering of background knowledge, and to designing extensive domain models, than is usually done in the area of machine learning.

Acknowledgments. I am grateful to Derek Sleeman, who offered some useful comments to the original version of this paper.

REFERENCES

[1] G.F. Cooper, E. Herskovitz. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 1992; 9: 309–347.

[2] R.G. Cowell, A.P. Dawid, D. Spiegelhalter. Sequential model criticism in probabilistic expert systems. *PAMI* 1993; 15(3): 209–219.

[3] D. Heckerman, D. Geiger, D. Clickering. Learning Bayesian networks: the combination of knowledge and data. *Machine Learning* 1995; 20: 197–243.

[4] T. Horváth, G. Turván. Learning logic programs with structured background knowledge. In: L. De Raedt (ed.), *Advances in Inductive Logic Programming*, IOS Press, Amsterdam, 1996, pp. 172–191.

[5] S.L. Lauritzen, D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society (Series B)* 1987; 50: 157–224.

[6] P.J.F. Lucas, L.C. van der Gaag. *Principles of Expert Systems*. Addison-Wesley, Wokingham, 1991.

[7] P.J.F. Lucas, H. Boot, B.G. Taal. Computer-based decision-support in the management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine* 1998; 37: 206–219.

[8] T.M. Mitchell. *Machine Learning*. McGraw-Hill, New-York, 1997.

[9] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo, California, 1988.

[10] M. Pradhan, M. Henrion, G. Provan, B. Del Favero, K. Huang. The sensitivity of belief networks to imprecise probabilities: an experimental investigation. *Artificial Intelligence* 1996; 84(1-2): 363–397.

[11] M. Ramoni, P. Sebastiani. *Efficient Parameter Learning in Bayesian Networks from Incomplete Data*. Report KMi-TR-41, Knowledge Media Institute (KMI), Open University, 1997.

[12] M. Ramoni, P. Sebastiani. *Bayesian Knowledge Discoverer: reference manual*. Knowledge Media Institute (KMI), Open University, 1997.

[13] M. Ramoni, P. Sebastiani. *An introduction to the Robust Bayesian Classifier*. Report KMi-TR-79, Knowledge Media Institute (KMI), Open University, 1999.

[14] M. Ramoni, P. Sebastiani, R. Dybowski. Robust outcome prediction for intensive-care patients. In: A. Abu-Hanna and P.J.F. Lucas (eds.), Prognostic Models in Medicine: Artificial Intelligence and Decision-analytic approaches. *Workshop Notes AIMDM'99*, Aalborg, 1999.

[15] H.C. Sox, M.A. Blatt, M.C. Higgins, K.I. Marton. *Medical Decision Making*. Butterworths, Boston, 1988.

Abstraction and Representation of Repeated Patterns in High-Frequency Data

Silvia Miksch[1], **Andreas Seyfang**[1] and **Christian Popow**[2]

Abstract. Monitoring devices in intensive care units deliver enormous streams of data in the form of snapshot measurements. In contrast, physicians use high-level abstractions in reasoning about the parameters observed.

Standard data abstraction algorithms can only handle data which are more regular than many variables observed in medicine. A typical example is the ECG in intensive care, where electric currents are measured at the skin surface and displayed amplified in order to detect problems in the conduction system of the muscular contraction pattern.

We developed an algorithm to transform a curve constituted by a series of data points into a set of bends and lines in between them. The resulting qualitative representation of the curve can be expressed as a list of objects each describing a bend. In this format, it can easily be utilized in a Knowledge-Based System (KBS).

In addition, in the case of rhythmical data, comparing selected bends in all cycles of the oscillation yields new information. This comparison can be done by plotting derived data as separate graph beside the original one or by encoding the knowledge behind the reasoning in rules in the KBS.

Our algorithm performs best on curves which are rhythmical but too irregular to be analyzed by Fast Fourier Transformation or other standard methods aiming at describing regular patterns.

We demonstrate our approach by displaying heart rate and Q-S-distance graphically aside of ECG-data (to detect impeded conduction) and by showing example code for rules detecting pathological deviations from the standard based on the qualitative representation of the curve.

1 Introduction

In all fields of medicine one is confronted with rhythmical[3] data. By rhythmical we mean data which show repeated patterns which slightly vary from instance to instance but still have enough in common to make it interesting to compare them, like ECG.

If such patterns are strongly regular, they can easily be analyzed by Fast Fourier Transformations (FFT) [6], a widely used and fairly exploited method. The result of such a transformation is a spectrum of frequencies, describing the harmonics of an oscillation. While meaningful in some fields of applications, like music or signal processing, this type of information by itself may not be meaningful for medical experts because of the complexity of the results and the post hoc type of analysis and because the data available in medical domains, are rarely regular enough to yield useful results when analyzed by FFT.

In many domains, like ECG, there is a long tradition in analyzing graphs and thus a lot of – in part informal – knowledge about the appearance of a graph and the health status of the corresponding patient. The way a graph appears to an expert depends on the kind and pattern of the bends it makes (sharp or smooth), the direction and straightness of the lines in between them (steep, flat, up, down), and the relative position of characteristic points in the graph within one oscillation cycle.

These types of characteristic features are far away from conventional tools for the analysis of oscillating data, since they focus only on the mathematical aspects of the data like frequencies or other highly abstract parameters. It is nearly impossible to transform the experiences of human experts in analyzing a graph in their mind and the way they formulate their constraints into such a mathematical set of parameters.

To bridge this gap, we developed a method to abstract characteristics similar to those used by human experts from a graph. In particular, we decompose the graph into a series of repeated patterns. Each pattern is described by a set of bends and lines in between. A bend is a (typically short) section of the graph where it changes its direction. It has a position and a "sharpness" defining how rapid the change takes place. A line is placed in between each pair of bends in order to represent the data points in between. Its most important feature is its inclination.

There are two characteristics of a bend, its sharpness – which is necessary to consider it significant – and the minimum distance of neighboring bends – which is required to distinguish them from noise. These abstracted characteristics can be visualized as bar charts or graphs. They can also be used to match the graphs with the conditions of rules in a knowledge-base like "If the ascent of the first line exceeds that of the third then ..." or "If the distance of the 2^{nd} and 3^{rd} corner decreases by more than 50 % during the first minute of measurement, then ...".

Thus, existing knowledge about the interpretation of graphs can be utilized with significant less effort on information

[1] Institute of Software Technology, Vienna University of Technology Favoritenstraße 9-11/188, A-1040 Vienna, Austria, email: {silvia,seyfang}@ifs.tuwien.ac.at

[2] Department of Pediatrics, University of Vienna Währinger Gürtel 18-20, A-1090 Vienna, Austria, email: popow@akh-wien.ac.at

[3] In this paper, we denote data as rhythmical, if they exhibit repeated patterns, but is not regular enough to be considered periodical in the sense of signal processing.

transformation compared to the use of conventional tools which require highly abstract input.

Of course, such abstractions can be done retrospectively at best. If on-line, a significant delay between the time of measurement and the time of calculation since considering a sufficiently large time interval is indispensable for thorough analysis.

In section 2 we show how our approach differs from other work. In section 3 we explain our approach in depth. In section 4 we describe the application of the generated data for both building a bridge between monitoring and knowledge-based systems on the one hand and to give the user compact information on the other hand.

2 Related Work

On-line monitoring at Intensive Care Units (ICUs) produces a high volume of high-frequency data generated by monitoring devices. These data need to be analyzed and interpreted to reduce the information overload of the medical staff and to guarantee quality of patient care. The interpretation of time-series is a very challenging task. The temporal properties are very important aspects in the medical domain, particularly when dealing with the interpretation of continuously assessed data. The most common methods are time-series analysis techniques [2], control theory, probabilistic or fuzzy classifiers. However, these approaches have a lot of shortcomings, which lead to apply knowledge-based techniques to derive qualitative values or patterns of the current and the past situations of a patient, called *temporal data abstraction*. Several significant and encouraging approaches have been developed in the past years (e.g., $TrenD_x$ [9], RÉSUMÉ [18, 19], VIE-VENT [16], Larizza et al. [13], Keravnou [12], Belazzi [3]). A comprehensive selection of various approaches in intelligent data analysis in medicine can be found in [14].

These approaches rely on predefined qualitative descriptions or categories of temporal data abstractions. For example, the RÉSUMÉ project [18, 19] recommends to apply state, gradient, rate, and simple pattern abstractions, Larizza et al. [13] are using basic and complex abstraction, and the temporal data abstraction module in the VIE-VENT system [16] tries to arrive at unified, context-sensitive qualitative descriptions applying smoothing techniques of data oscillating and expectation-guided schemata for trend-curve fitting. In contrast, Calvelo et al. [4] seek to separate stable patients at an adult ICU from such in a critical situation by applying machine learning methods.

A comprehensive study about various approaches of intelligent data analysis for medical diagnosis using machine learning and temporal abstraction techniques can be found in [15]

However, we are going one step back and want to explore, which kinds of temporal data abstractions are needed for rhythmical data. We are demonstrating a way to acquire complex data abstraction methods to arrive at qualitative descriptions, like "the variability of the $\overline{PQ}$-distance increase significantly during the last 2 hours" which directly indicate medically relevant facts like – in this case – a problem in the excitation conduction from the atria to the ventricles.

A similar technique is the "Time Series Workbench" [11], which approximates data curves with a series of line-segments. However, we are going beyond the approximation by line-segments and take the particular characteristics of a graph into account, like the "sharpness" of a curve.

3 The Algorithm

While mathematicians might be horrified by the notion of a graph being a series of bends connected by rather straight lines this resembles the cognitive model most non-mathematicians use when looking at a graph. But how can we find a formal definition of such an informal entity as a bend?

There are two indications for bends in a curve: First, the second-order derivative of the curve shows a minimum in places where the original curve does a "bend to the right", i.e. changes from increase to decrease, and a maximum, where the original curve does a "bend to the left", i.e. changes from decrease to increase.

Second, we calculate linear regressions for a time window sliding over the curve as described in [17]. In places where the curve shows a bend, reducing the length of the interval will lead to a decrease in the standard error of the resulting linear regression. In places where there is no significant bend, shortening the time window will not decrease the standard error.

We will first explain both approaches in detail and then discuss which of them is more suitable for which type of data.

3.1 Using the Changes of the Derivative

Figure 1 shows an abstract example. A bend in the curve is characterized by a change in its derivative. The bigger the change in the derivative, the sharper the bend – and the bigger the absolute value of the second-order derivative.

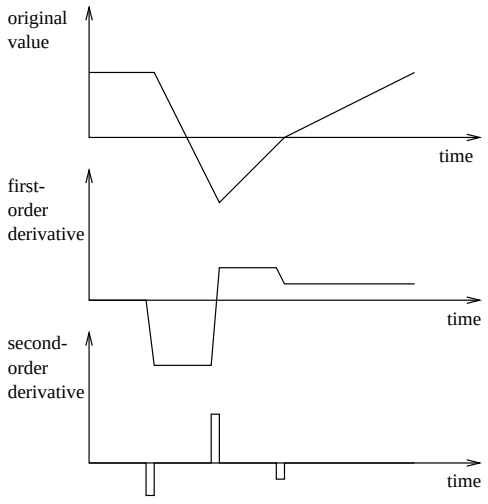


Figure 1. Abstract demonstration of bends in the original graph and its second-order derivative: In places where the original graph shows a bend, the first-order derivative's changes, which causes a peak in the second-order derivative.

While this notion is perfectly true for small derivatives, looking at changes in the absolute value of the derivative will overemphasize relatively small changes in places of high

derivative. If e.g. a derivative of 10 changes by 2, this might not seem too significant to an observer while a change from 0 to 2 certainly will. The second-order derivative is 2 in both cases. So its value will not reflect the users estimations. Figure 2 shows an example of a peak in the second-order derivative where a human would not see a significant change.

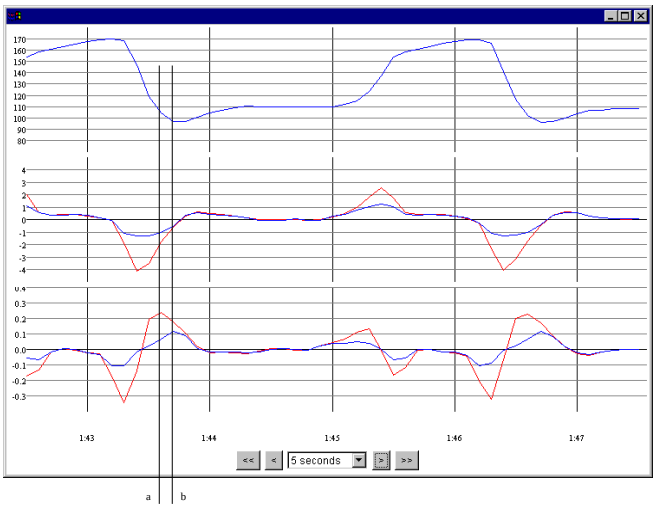


Figure 2. The topmost graph shows the original data. In the middle the gray graph (that with the bigger extrema) shows the absolute value of the first derivative and the black graph shows the angle of inclination. At the bottom, the derivatives of both graphs in the middle are shown. (The more moderate, black one is the derivative of the moderate one in the middle). While the change in absolute value of the first-order derivative is biggest in (a), the change in the angle associated with the derivative is biggest in (b). Human spectators seem to prefer (b) over (a) if asked to define the significant corner at this portion of the graph.

Using relative changes in the derivative only works for steep slopes and will overemphasize changes in flat regions of the curve. Instead, we are using the angle of the derivative. So instead of the derivative itself we calculate the angle α as $\tan\alpha = \frac{\Delta y}{\Delta x}$ and use the derivative of this function as an indicator of significant changes in the curve.

Figure 3 shows an example, where this function nicely reflects human perspective. The curve slightly but constantly turns up. So it is difficult to say, where a single corner should be. The derivative of the derivative's angle (i.e. the angle of inclination of the original curve) is constantly but slightly increasing at that part of the curve, reflecting the indecision of the observer.

In practical applications, calculating the derivative as the difference in the y-coordinate of two neighboring data points (divided by the difference in their x-coordinate) does not work on noisy input data, because the small erroneous oscillations of the curve might result in the derivative oscillating enough to hide the overall tendency of the curve. Comparing each point with the point following n points later instead of the ultimate neighbor (and placing the result in the middle between the two points) often yields sufficiently smooth graph for the derivative without the need to smoothen the original curve. The number of intermediate points n should be bigger then the typical wave length of erroneous oscillations or – for nondeterministic noise – simply big enough to suppress the

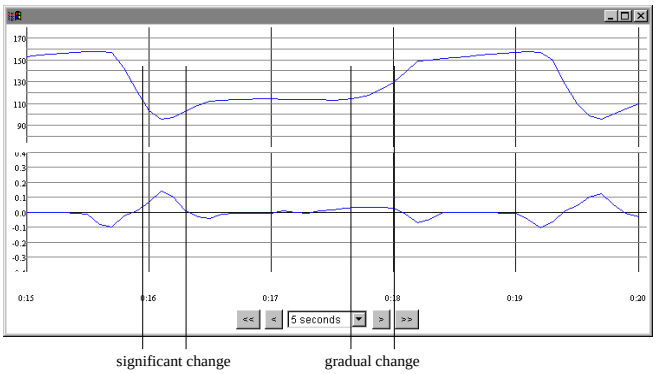


Figure 3. At gradual turns of the original curve (at the top), were a human observer has difficulties in pointing at the exact position of a single corner, the indication function (below) is trapezoidal reflecting the her indecision.

portion of noise in the result of

$$\begin{aligned} \textit{calculated derivative} &= \frac{n * \textit{real derivative} + \textit{noise}}{n} \\ &= \textit{real derivative} + \frac{\textit{noise}}{n} \end{aligned}$$

where *noise* is the average distant between a measured value and the real value, *real derivative* is the derivative of the ideal graph drawn from the real values (which is not known, of course) and *calculated derivative* is the value resulting from this calculation.

3.2 Using the Length of the Regression Line

The algorithm presented in the following seeks to detect bends in the graph by first calculating a linear regression for short sections of the graph and then checking whether reducing the size of the section reduces the standard error of the regression line.

The reason for applying linear regression lies in its ability to give an abstract representation of a set of data points and at the same time an estimate, how will this abstraction represents the actual data (by the standard error). If the regression line does not fit to the curve because it make a bend, then cutting the ends of the line results in a significantly reduced standard error. If the regression line does not fit the curve because the curve is noisy, a shorter regression line will have an equally high standard error as the original (full length) one. This distinction can be exploited to detect bends in a graph.

As illustrated by figure 4, we slide a window of consideration (*time window*), which is of fixed size over the time axis of the curve in small steps. For each instance of the time window, we calculate a linear regression for the data points contained in it. As opposed to [17], for this application the step width should equal the rate of data points (if there is one measurement per second, step width should be one second) and the length of the time window should be short enough to follow significant bends but much longer than erroneous oscillations.

So, for example, if the sampling rate is 1 measurement per second and the oscillations caused by noise show a wave length of up to 5 seconds, the step width will be one and the size of

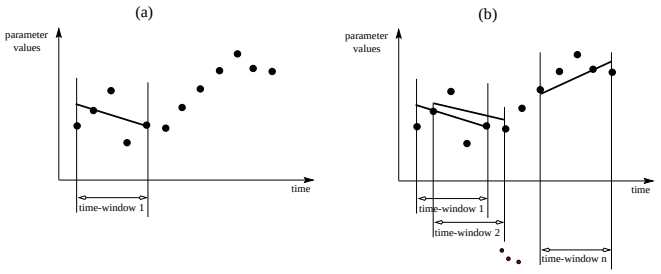


Figure 4. The calculation of the linear regression is done for a time window of fixed size sliding over the entire curve in small steps. (a) shows a single time window and the line calculated from the data points within it. (b) shows a sequence of overlapping time windows and the resulting lines.

the time window will be between about 7 and 10 seconds. We will thus receive one regression line per data point, calculated from its 7 to 10 neighbors.

The standard error of a linear regression line shows, how well the points, which are represented by that line fit together respectively to that line. The bigger the average distance, the bigger the standard error.

For each regression line, we take a look at its ends (see figure 5). On each end, there might be some neighboring points on the same side of the line. If a smooth curve takes a bend they will be numerous, if the graph is rather straight, but oscillating around the line, there will be very few points at the same side of the line.

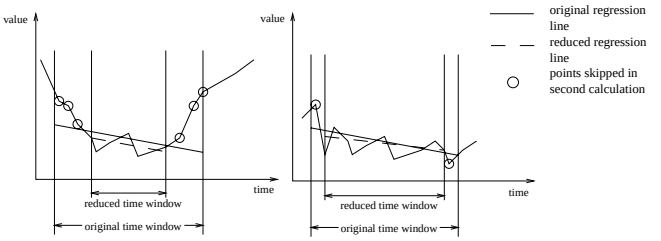


Figure 5. If the graph shows a bend in the interval under consideration (example on the left-hand side), there is a considerable number of data points on each end of the regression line which lie on the same side. Skipping them in the recalculation of the regression reduces the standard error to which the skipped points contributed significantly. If there is no bend (example on the right-hand side), skipping the few points on the ends does not reduce the standard error.

Next we shrink the time window to skip those groups of points on both ends which altogether lie on the same side of the curve and recalculate the linear regression for this second, smaller time window. If the distance of the skipped points exceeds the average distance of all points in the (first) time window to the (first) regression line, the standard error of the second regression line will be smaller than that of the first one. In this case we can assume that the deviation of the points on the ends of the line are not just an incident, but caused by a bend in the curve.

The difference in length between the first and the second time window as well as the decrease of the standard error are measures for the "sharpness" of the curve. Thus both of them can be used as indication function. Both only give positive values. The direction of the curve can be derived from the side of the regression line, on which the skipped data points lie. So we assign minus to bends to the right and plus the bends to the left and supply the absolute value of the indication function with this sign to produce an indication function compatible with the one described in section 3.1.

3.3 Common Issues of Both Approaches

In both cases (second-order derivative and length of the regression line) a bend in the curve will not yield only one high value at a single position on the time axis, but a more or less significant peak. Especially, bends with bigger radius result in a series of peaks or a long flat "hill" in the second-order derivative respectively a "valley" in the curve showing the length of the regression line.

To suppress such concurring peaks one can simply define a minimum distance (along the time axis) and only chose the highest peak out of several of them if their distance is below this threshold.

A better way is to consider both of two neighboring peaks only if they are separated by a local minimum of a certain depth. To see the difference to the above strategy consider the following cases: First, two sharp bends close to each other and second, a long slight bend.

Two sharp bends produce two high peaks with a clearly distinguishable minimum (of the absolute value) in between. If you only consider the distance on the time axis of the two peaks, you will have to ignore one of them, if you consider the minimum between them, you will accept both peaks to be significant.

A long slight bend results in a series of peaks with nearly no minimum between them. If you consider the distance along the time axis, the first and the last minor peaks might be far enough from each other to let both of them seem justified. If you look at the minima between them, you will ignore all but one of them.

Many curves of real data show small opposite bends which should be considered as a single straight line. A small threshold for the absolute value of the indication function does this job.

3.4 Matching the Two Approaches

The first approach – the change in the angle of inclination – is very intuitive if applied on smooth graphs. Applied on noisy input data, the graph of its indication function can get too distorted to be usable.

The second approach – the length of the regression line – is harder to compute than the first one. The outstanding advantage of linear regression is that it minimizes the influence of noise on the result. If the original graph shows numerous random peaks, they can fool the second algorithm because they might inhibit proper reduction of the regression line.

In such cases, a combination of both approaches performs best: The indication function is the change in the ascent of the regression lines.

1. The regression lines are calculated as described in section 3.2.
2. For each of them, the angle of inclination is calculated.

3. Then the resulting values are merged to a new function (replacing the first derivative in the first approach)
4. The derivative of this function is calculated as the indication function for detecting bends.

To summarize, given smooth input data, the first approach performs better. The more smoothing is necessary before or while calculating the derivative, the smaller this gain becomes.

3.5 The Resulting Representation

As results of the transformation of the discrete data points into bends and lines, we obtain three streams of different types of data: The bends, the corners of the original curve at those places where bends were detected, and the lines in between the bends.

3.5.1 Bends

By the term *bend* we subsume the abstract aspects of a turn in the original graph. Each bend is described by the position of its middle along the time axis, the height of the corresponding peak in the indication function (second-order derivative or length of regression line) and the area of the peak measured from one intersection with the zero-line to the next.

3.5.2 Corners

By the term *corner* we describe the position in which the lines neighboring a bend would meet. The x-coordinate of the corner clearly equals the middle of the bend. The y-value can be the y-coordinate of the nearest point in the original curve. To reduce the influence of noise, in most cases it is necessary to take the average of some of its neighbors into account too. Integrating too many of them in the calculation will distort the result towards the inner side of the bend. Thus, this parameter needs careful optimization.

3.5.3 The Lines in Between

The *lines* between the bends represent the data points of the original graph between two neighboring bends. They can either be drawn just as connections of the corners of the curve, or they are calculated as a linear regression of the points of the original curve between the bend.

3.6 Relating the Cycles in Oscillations

In many cases, looking at only one oscillation alone is not sufficient, but tendencies developing through a possibly long series of oscillations as well as deviations from the standard or average are interesting. To arrive at this, the following steps are performed:

1. Corresponding bends are found.
2. Their position relative to the start of the oscillation is calculated.
3. The data calculated in step 2 are related to the average of the previous instances of the oscillation.
4. Optionally, slope and standard deviation within a sliding time-window [17] is calculated for any of the values created in step 2 and 3.
5. Optionally, values calculated in steps 2 to 4 can be transformed into qualitative i.e. symbolic representation.

3.6.1 Relating Cycles within an Oscillation

First, for each cycle, a reference point must be found. This should be a point which is known to be invariant itself. A well-tried method is to take a position where the value changes rapidly and steadily, forming a steep slope. In the middle of this section of the graph, its value is chosen as a threshold. The point, where the graph of each oscillation intersects with the horizontal line of the threshold is the reference point of that oscillation. This way, the reference point is most invariant, even in oscillations with varying cycles.

To properly assign corners to groups, just looking at their ordinal number in the stream of corners constituted by each oscillation is not enough. Too often a corner is missing, because it was too small to exceed the threshold, or it is doubled due to errors in the data or the underlying data is simply irregular at that position.

Looking for nearest neighbors suffices only for curves, in which the positions of the same instance of a corner (e.g. the first one) do not change over time. If they do, it is useful to extrapolate the expected position of the next instance of a corner from the previous ones. For this purpose, both the x– and the y–coordinate are considered separately and a linear regression is calculated for each of them.

Often one coordinate (x or y) of a corner is much more regular than the other. In this case, weighting the deviations when looking for near neighbors (with or without extrapolation) helps to improve the result. The more regular coordinate is multiplied by a bigger weight than the other.

3.6.2 Numerical Absolute Values

For each bend, the above abstraction yields the following data:

- The corresponding maximum in the indication function shows how sharp the bend is.
- The area of the corresponding peak shows how significant the change of the original graph is.
- The x–coordinate shows the position of the corner on the time–axis.

For each line, its inclination is computed.

3.6.3 Numerical Relative Values

Each of the above values is measured against the average of previous instances in an interval of time defined by the user. The deviation is given both absolute and relative.

3.6.4 Qualitative Values

The quantities computed in the two steps above can be qualified using a set of tables. For each parameter, a table lists all qualitative values it can take and the numerical limits in between.

4 Fields of Application

In the following, we give some examples of how the data obtained in the previous section can be used.

4.1 Interfacing Knowledge-Based Systems

To bridge the gap between data analysis and knowledge-based systems (KBS) [7], we transform the output into clauses compatible with those use by a KBS.

4.1.1 Symbolic Representation of Features

The values describing bends, corners or lines can be expressed in a list to make them accessible to symbolic reasoners like knowledge-based systems or machine-learning tools.

To improve readability of the output, each corner and each line can be assigned a symbolic name (e.g. P, Q, R, S and T in an ECG) instead of its ordinal number to denote it in an intuitive fashion.

The following example describes a graph consisting of a line increasing by 20 degrees for 100 seconds followed by a narrow bend to the right and 30 seconds of decrease in Clips-Syntax [1]. In this example we omit the corners for clarity.

```
(graph-features(line (100sec up 20))
               (bow (right narrow))
               (line (30sec down 30)))
```

4.1.2 Applying Rules to Detect Patterns in Monitored Data

In the following, we show how knowledge about the interpretation of ECG [8] can be translated into rules and how the data abstracted as described before can be matched with such rules.

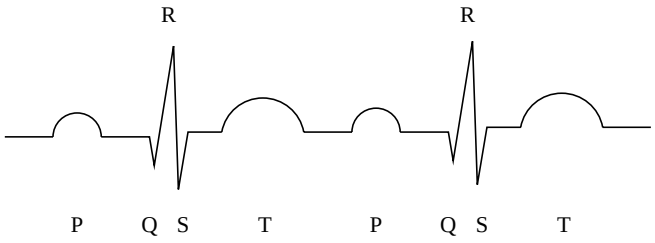


Figure 6. Idealized ECG. The features are labeled by letters. They are explained in Table 1.

Feature	Causality
P wave	excitation of the atria
QRS complex	excitation of the ventricles
$\overline{RR}$ distance	instantaneous heart rate
$\overline{PQ}$ distance	excitation conduction from the atria to the ventricles
$\overline{QS}$ distance	excitation of the ventricles prolonged in case of impeded conduction (heart block)

Table 1. Relation between features of the ECG and underlying causalities.

Figure 6 shows an idealized ECG. Table 1 compares some aspects of an ECG with underlying aspects of the heart and its possible problems. Figure 7 shows some rules in CLIPS [1]

```
(defrule ventricular-conductivity
   (distance (from Q)
             (to S)
             (value ?value):&(> ?value 55))
   (patient  (age ?age):&(and (< ?age 2)
                              (>= ?age 1))))
=>
   (assert (diagnosis (ventricular-conductivity
                       bundle-branch-block)))
```

Figure 7. CLIPS-rules applied on the symbolic representation of an ECG. Note that we measure the distance from peak Q to peak S and not the duration of the QRS-complex as a whole, which is approximately 15 msec longer.

to detect bundle branch blocks and AV-blocks on a symbolic level.

A more detailed analysis of ECG pathology in this setting remains speculative – except for the detection of (ventricular) extrasystoles – because the monitoring ECG is usually not derived under standardized conditions.

4.2 Visualization

The abstraction methods described above produce both qualitative and quantitative information, both on the level of single bends and as attributes of a cycle within the oscillation. In addition to these two dimensions, some features described are relatively rare, e.g., abnormalities in an ECG, others form a steady stream of data, calculated for every instant in time during measurement, e.g. the heart rate.

Sparse events and qualitative information tend to be visualized symbolically, e.g., as bars or markers, while qualitative information is commonly displayed as graphs. For a deeper discussion of visualization aspects see [20, 5].

In the following, we give some examples, namely the display of bend as bars, plotting features of an oscillation over a relatively long period of time and using markers to represent qualitative information.

4.2.1 Displaying Bends as Bars

Bends are features located at a certain position along the time axis and have several abstract attribute, the most important being sharpness and significance. These are derived from the height of the peak in the indication function and the area delimited by the graph indication function and the time axis.

Thus, it seems intuitive to visualize each bend with a significance above a certain threshold as a bar which is equal in height and area to the peak in the indication function.

Figure 8 illustrates this approach applied on data from ergonomic studies in rowing.

4.2.2 Displaying Relative Positions as Graphs

For each corner in the original curve, its position relative to the start of the oscillation or its distance to another corner in the same oscillation can be displayed as a graph.

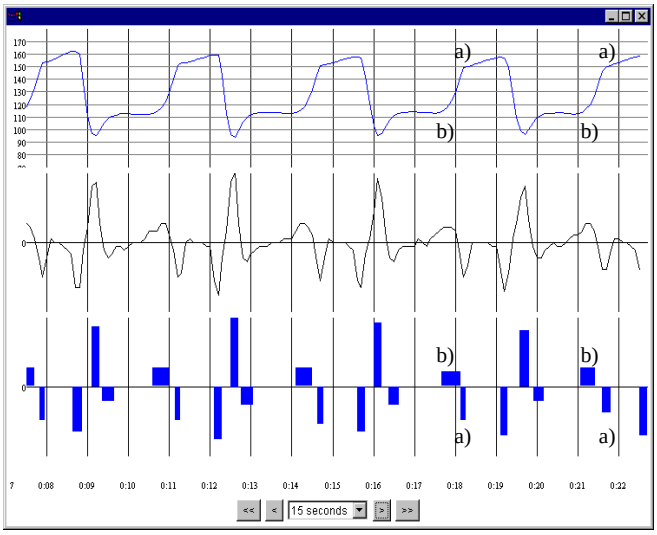


Figure 8. Starting at the top, we show the original graph, the indication function and the bars representing the significant bends. a) shows an example of an irregularity in the curve which a human could also dedect if concentrating on every detail: the bend in the right-most oscillation is not as sharp as the corresponding one in the other oscillations. b) draws our attention to a feature not perceptible by looking at the raw data: the long-spread bow to the left of the second oscillation from the right is not as sharp as the others as indicated by inferior height of the bar. The corresponding part of the original curve does not seem different by itself. The significance of the features found in a) and b) depends on the domain knowledge about the data represented by the curve.

Figure 9 gives an example from ECG monitoring. For demonstration purposes, we display the heart rate and the $\overline{QS}$ distance below the very condensed ECG graph.

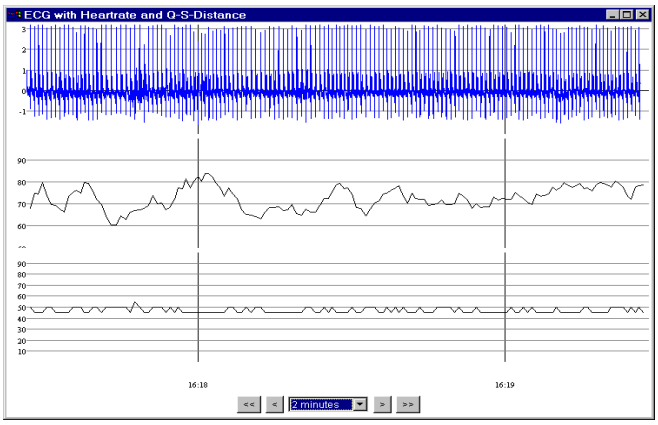


Figure 9. Two minutes ECG sample from an eight hours polygraphic recording of a two years old child. The first graph shows the condensed ECG. The second graph gives the instantaneous heart rate derived from the $\overline{RR}$ distance. The third graph gives the instantaneous $\overline{QS}$ distance. The apparent variability is caused by digitization inaccuracies (resolution is 200 Hz).

4.2.3 Displaying Qualitative Information along the Time Axis

Both the qualitative information derived from lookup-tables as described in section 3.6.4 as well as the results from knowledge-based reasoning as described in section 4.1.2 can be visualized graphically along the time axis.

For rare deviations in the data hinting at significant events, markers in different colors and shape, which are lined up below the original data, are most appropriate.

For continuous information, like qualitative status information, coloring a stripe below according to the values (e.g., blue for low, red for increased, green for normal) yields a dense and intuitive information representation.

Another way to visualize state information is the use of a set of symbols in different colors to represent two or three dimensions in one place. Such a technique has been successfully employed in the VIE-VENT-system [16] for instantaneous status information.

5 Conclusion

We have presented several methods to capture complex rhythmical curves by transforming them into series of bends, corners, and lines, based on the observation that a bend in the curve is synonym to a change in its inclination.

Our approach is applicable to data where Fast Fourier Transformation fails, because the oscillation is not regular enough for such a strictly numerical algorithm. Furthermore, a frequency spectrum is a less intuitive representation of a curve than series of corners and lines in many medical domains.

The abstraction of characteristics from a stream of raw data points offers the following opportunities:

Compact Visualization. Displaying only the important features of a graph in an abstract form in addition to the original graph allows for easy detection of trends and outliers which otherwise would be buried in the overwhelming impression of countless oscillations.

Bridge to Knowledge Representation. The abstracted characteristics extracted by our algorithm can be matched against conditions in a rule base. So the curves can be tagged according to a set of classifications stored in a knowledge base. This aspect is crucial for the integration of high-frequency data and symbolic systems such as symbolic machine learning, knowledge-based systems for intelligent alarming and a guideline execution system like as developed in the Asgaard project.

The algorithms described have been implemented in Java^{TM} in an experimental setting to allow their evaluation. Future work will be devoted to the acquisition of rules for the automatic interpretation of clinical data and in the implementation of several modes of graphical display to meet the practical requirements under various settings.

Acknowledgments

We thank Michael Urschitz and Klaus Hammermüller for supplying data and Maria Dobner, Georg Duftschmid, Werner Horn, and Robert Kosara for their useful comments. This

project is supported by "Fonds zur Förderung der wissenschaftlichen Forschung - FWF" (Austrian Science Funds), P12797-INF.

REFERENCES

[1] 'Clips reference manual', Technical report, Software Technology Branch, Lyndon B. Johnson Space Center, NASA, (1993).
[2] R.K. Avent and J.D. Charlton, 'A critical review of trend-detection methologies for biomedical monitoring systems', *Critical Reviews in Biomedical Engineering*, **17**(6), 621–659, (1990).
[3] R. Bellazzi, C. Larizza, P. Magni, S. Montani, and G. De Nicolao, 'Intelligent analysis of clinical time series by combining structural filtering and temporal abstractions', In Horn et al. [10], pp. 261–270.
[4] Daniel Calvelo, Marie-C. Chambrin, Denis Pomorski, and Pierre Ravaux, 'ICU patient state characterisation using machine learning in a time series framework', In Horn et al. [10], pp. 356–360.
[5] Stuart K. Card, *Readings in Information Visualization: Using Vision to Think*, Series in Interactive Technologies, Academic Press/Morgan Kaufmann, 1999.
[6] J. W. Cooley and J. W. Tukey, 'An algorithm for the machine calculation of complex fourier series', *Mathematics of Computation*, **19**(90), 297–301, (1965).
[7] J. Giarratano and G. Riley, *Expert Systems - Principles and Programming*, PWS Publishing Company, Boston, second edn., 1994.
[8] W. G. Guntheroth, *Pediatric Electrocardiography*, W. B. Saunders, Philadelphia-London, 1996.
[9] I. J. Haimowitz and I. S. Kohane, 'Managing temporal worlds for medical tread diagnosis', *Artificial Intelligence in Medicine, Special Issue Temporal Reasoning in Medicine*, **8**(3), 299–321, (1996).
[10] Werner Horn, Yuval Shahar, Greger Lindberg, Steen Andreassen, and Jeremy Wyatt, eds. volume 1620 of *Lecture Note in Artificial Intelligence*, Aalborg, Denmark, June 1999. Springer.
[11] Jim Hunter and Neil McIntosh, 'Knowledge-based event detection of complex time series data', In Horn et al. [10], pp. 271–280.
[12] E. T. Keravnou, 'Temporal abstraction of medical data: Deriving periodicity', In Lavcač et al. [14], 61–79.
[13] C. Larizza, R. Bellazzi, and A. Riva, 'Temporal abstractions for diabetic patients management', in *Proceedings of the Artificial Intelligence in Medicine, 6th Conference on Artificial Intelligence in Medicine Europe (AIME-97)*, pp. 319–30, Berlin, (1997). Springer.
[14] Nada Lavcač, Elpida Keravnou, and Blaž Zupan, eds., Kluwer Academic Publishers, Boston/Dordrecht/London, 1997.
[15] Nada Lavcač, Igor Kononenko, Elpida Keravnou, Matjaž Kukar, and Blaž Zupan, 'Intelligent data analysis for medical diagnosis: Using machine learning and temporal abstraction', *AI Communications*, **11**(3,4), 191–218, (1998).
[16] S. Miksch, W. Horn, C. Popow, and F. Paky, 'Utilizing temporal data abstraction for data validation and therapy planning for artificially ventilated newborn infants', *Artificial Intelligence in Medicine*, **8**(6), 543–576, (1996).
[17] S. Miksch, A. Seyfang, W. Horn, and Popow C., 'Abstracting steady qualitative descriptions over time from noisy, high-frequency data', In Horn et al. [10].
[18] Y. Shahar, 'A framework for knowledge-based temporal abstraction', *Artificial Intelligence*, **90**(1-2), 267–98, (1997).
[19] Y. Shahar and M. A. Musen, 'Knowledge-based temporal abstraction in clinical domains', *Artificial Intelligence in Medicine, Special Issue Temporal Reasoning in Medicine*, **8**(3), 267–98, (1996).
[20] Edward R. Tufte, *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut, 1983.

Analysis of Primary Care Data

Ying-Lie O[1]

Abstract. Intelligent data analysis methods have been used to determine the knowledge models for medical decision support. In primary care all related data are entered in the computer-based patient record (CPR) as events in the journal. The overall characterisation of the care provision is based on patient groups with specific healthcare related conditions and needs.

The development consists of the following steps: problem formulation, database configuration, and data analysis. The features are chosen using a heuristic strategy: initially based on domain knowledge, and then the contribution of the remaining attributes is tested.

The data-set for analysis is count-based. The patient groups are obtained using a modified nearest neighbour cluster analysis method. The proposed approach is mainly data-driven. Only a very limited domain knowledge has been used for the initial selection of features, correction of outliers, and interpretation of the results.

1 INTRODUCTION

Intelligent data analysis methods [3, 7] have been used to determine the knowledge models for medical decision support. Most of the use concern specific health conditions [4, 8, 1], such as chronic diseases, or critical care. The required information is generally extracted from data-sets that contain cases. Each case consists of attributes with values that represents a possible condition associated to the disease.

A computer-based patient record (CPR) supports the care provision as a journal of events. The lack of standards in vocabulary, incompleteness, and inaccuracies limits the overall analysis of the data. The composition of a data-set for analysis requires well designed processing.

The analysis methods are bound by the inexact nature of the data. This means that "soft" methods prevail above logic methods. Appropriate methods would be association rules from data mining and cluster analysis from pattern recognition, and for a limited number of variables also regression, interpolation, and neural networks.

2 PROBLEM STATEMENT

The provision of primary care is generally provided in local health care centres or out-patient clinics of regional hospitals. These health care centres are typically staffed by GPs (general practitioners) and practice nurses.

The typical setting of routine primary care is the current visit and possible follow-up visits. All related data are entered in the CPR as events in the *journal*. The journal is a layered composition of care activities, short notes, diagnostic tests, and other data. The main entry consists of activity types and further reference to detailed information. Primary patient characteristics are stored in the patient identification record. To allow analysis, the layered event-based structure must be converted to a data-set in a single layered structure or *universal relation* as opposed to the highly normalised database.

In order to provide proper health care, an overall characterisation of the required care provision is needed. The characterisation is based on patient groups with specific health care related conditions and needs. Each group is distinguished by health care related features, some of which specifically determine outcomes. All features are specified by variables that are attributes of the *item-set*.

In a top-down approach, first of all the most general patient groups that can be directly determined from the main entries of the journal are determined. Analysis of more detailed information not necessarily yield a hierarchical structure, because the structure of the data may be different from the functional association. For instance, a disease-based item-set consists of all attributes from different levels of detail related to the disease.

3 OVERALL CHARACTERISATION OF PATIENT GROUPS

The overall characterisation is determined from the main entries of the journal in conjunction with primary patient characteristics in the patient identification record.

In this study, the posed question is: "What is the amount of provided care activities according to general patient characteristics?".

The development consists of the following steps: problem formulation, database configuration, and data analysis.

3.1 Problem Formulation

The above posed question is a common data mining problem [2]. With respect to the covered data, the *problem formulation* encompasses

- *Description* divides most of the data into a limited number of large partitions that gives the dominant concepts for feature selection.
- *Grouping* divides most of the data into a reasonable number of partitions with clear concepts that specifies the features.

The selected features is a subset of available attributes that directly contribute to health care activities: $\{age, gender, history, activitytype\}$. In this item-set, $activitytype$ is the *outcome*, and all other variables are considered to contribute to this outcome.

The problem is formulated as follows: find the patient groups based on the association $\{age, gender, history, activitytype\} \rightarrow \#activities$, where $\#$ stands for the number of counted instances.

[1] Julius Centre for General Practice and Patient-Oriented Research, University Medical Centre Utrecht, the Netherlands, email: y.o@jc.azu.nl

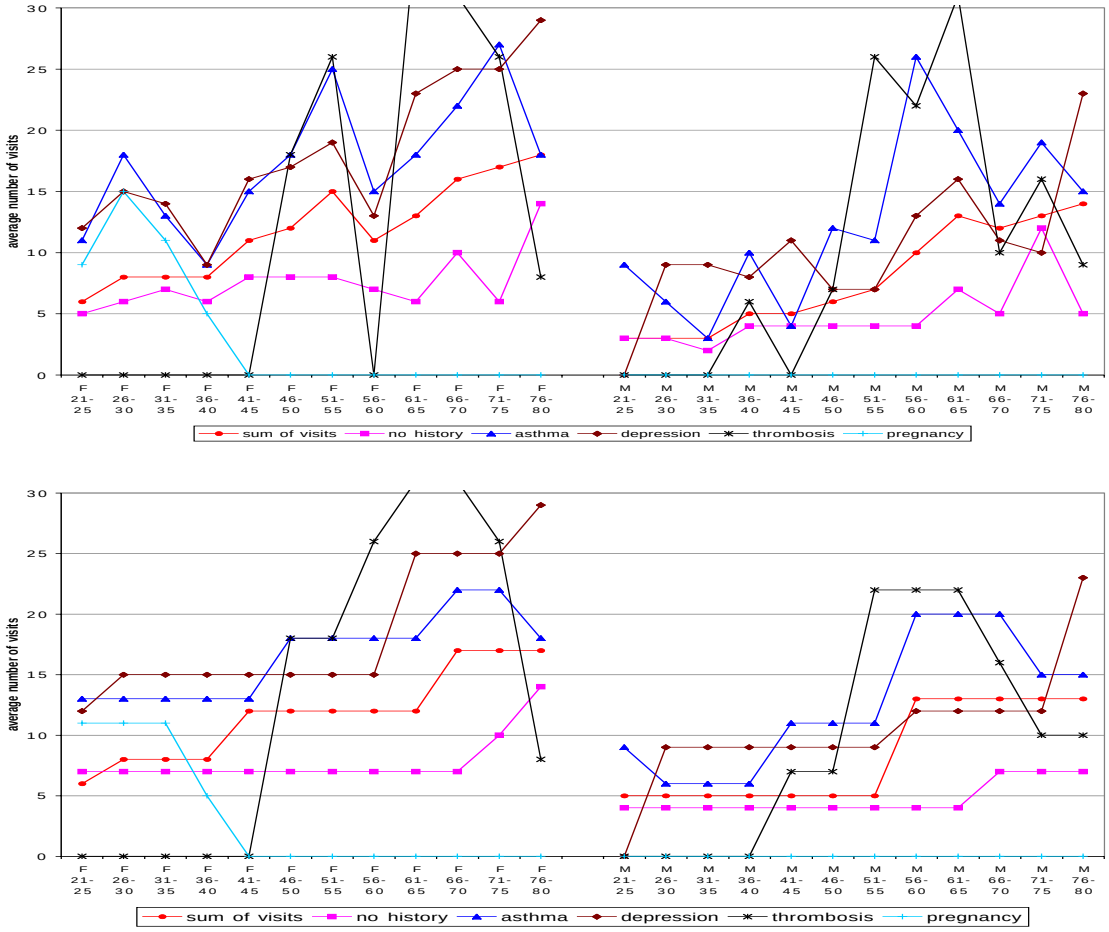


Figure 1. Upper: Data-set containing average numbers of visits for different history conditions in age groups of 5 years for females (left) and males (right). Lower: Analysed data-set depicting the clusters of patient groups.

Instead of an attribute set-based approach, a *count-based* solution is applied by replacement of the above association by the number of counted instances in the item-set $\{age, gender, history, activitytype\}$.

The description problem regards a summary of occurrences in the item-set and the grouping problem concerns the division of into a reasonable number of patient groups.

3.2 Database Configuration

Database configuration is the process of making a data-set that is suitable for analysis according to the problem formulation. This includes the definition of the data structure according to the item-set, retrieval, pre-analysis for formatting, correction and preparation of the data, and pre-analyses regarding the choice of attributes. These tasks are mainly database operations using queries and built-in functions.

The selection of the attributes can be performed in two ways: by analysis of the contribution or association of all attributes to the outcome, or a heuristic strategy.

In the *heuristic* strategy, first a hypothesis is posed, then the alternatives are tested. Failing the test implies adaptation of the initial hypothesis. In case of the item-set, first the predefined attributes are considered, then the contribution of the remaining attributes are analysed. If the contribution is significant, then the attribute is included.

The first action is the creation of a *count-based* table from the original normalised database tables. Preliminary analyses have shown that several attributes do not affect the number of activities. Also, a limited number of activity types is given to a majority of patients characterised by age and gender. The most important activity type are the regular visits. Hence, the item-set is reduced to the desired attributes $\{age, gender, history, activitytype = visit\}$.

The genders have significantly different behaviour. There is a functional tendency in the age dependence, but it is disturbed by fluctuations. Therefore, the ages are stratified into bins of 5 years, and ages below 20 and above 80 are excluded because of different dependencies. Age is treated as the running variable, separately for each gender.

It is generally not possible to correct for outliers, unless it is an obvious error such as a value in the midst of zeroes, or the condition “pregnancy” for the gender “male”. Missing values can be corrected by statistical imputation, or nonlinear filtering such as median or opening (closing) of mathematical morphology. Nonlinear filtering removes outliers and smoothes fluctuations. Closing fills ditches related to neighbouring values, while opening cuts peaks.

The description problem regards the distinction between different history conditions. History conditions are generally only registered in case of a chronic disease, or if the patient has been treated. For the grouping problem several different types of history conditions with high numbers of instances are selected: $history = \{nohistory, asthma, depression, thrombosis, pregnancy\}$.

Thus, the final data-set in a count-based table containing the number of visits for different history conditions according to the item-set $\{age = \{binsof5year\}, gender = \{F, M\}, \{nohistory = \#visits, \ldots, pregnancy = \#visits\}\}$.

The reduction of the amount of original data to the final data-set is shown in Table 1.

Table 1. The number of occurrences in each data-set

	original in journal normalised	associated to visits count-based	selected age bins count-based	final data-set count-based
activities	390945	81086	62143	31775
patients	11218	8921	6506	4452

3.3 Data Analysis

The pre-analyses in the database configuration part, the selection of features and attributes, and the resulting item-set and data-set can be considered as an answer to the description problem.

Grouping divides the data into partitions that are not necessarily non-overlapping. In this case, the number of clusters are not known beforehand (unsupervised).

Clustering can be performed by association rules [6] or cluster analysis [5]. Analysis reveal a large variety of different history of health conditions, with low fractions of occurrence that are not discriminative to distinguish different clusters.

Therefore, the grouping problem will be based on the transformed data-set $\#data = f(\#visits, \#patients)$, where the function f is the rounded *specific average* of the number of visits $\#visits$ on the number of patients $\#patients$ for each bin. The resulting data-set is shown in Figure 1.

Cluster algorithms are based minimising the disimilarity between items "within" the clusters, and if non-overlapping maximixing the disimilarity "between" clusters. It is common to use a metric as disimilarity measure. For this purpose, the d_1 metric $d(x, y) = \sum |x - y|$ will be used. The clustering algotrithm is a modified *nearest-neighbour* algorithm that takes into account that the items are values instead of spatial points.

1. Initialisation: Specify the nearest-neigbour threshold t, typically 1 or 2 times the standard deviation. Select a number of initial clusters points, for instance by taking local maxima.
2. Nearest neighbour: $\forall$ item x_i in a cluster, find its nearest neighbour x_j.
3. Assignment: If $d(x_i, x_j) < t$, assign x_j to the cluster, otherwise assign x_j to a new cluster.
4. Stopping criteria: If every item has been assigned to a cluster, stop. Else repeat the process and go to step 2.
5. Completion: Remove small or outlier clusters, and assign the items to the clusters of its nearest neighbours if $d(x_i, x_j) < 3t$, else assign to a new cluster. If every item has been reassigned, stop, and if desirable assign the median value of the clusters to the items in the cluster.

The resulting clusters are shown in Figure 1. In comparison with the rather disturbed data, the different patient groups are now clearly distinguished including the tendencies.

The overall number of visits increases with age for both genders, women have a higher contribution. If no history is registered, then it is constant with a slight increase for the elderly.

Asthma and depression occur at the whole range of ages, and is increasing with age. Thrombosis occurs at a certain age, has a peak, and then decreases. As can be expected, pregnancy only occurs for females in the reproductive age.

4 CONCLUSION

A strategy for the selection of features, and a modified nearest neighbour clustering method for count-based data-sets have been proposed. The approach seems to be appropriate for the analysis CPR data.

The heuristic strategy of first choosing the features based on domain knowledge, and then testing the contribution of the remaining attributes is satisfactory. Only a very limited domain knowledge has been used for the initial selection of features, correction of outliers, and interpretation of the results.

The clustering algorithm gives promissing results, its choice is motivated by the failure of the much used association rules method. This is caused by the main property of this data set: large number of variables with relatively small instances. Using a specific average compensates for fluctuations in the data. The nearest-neighbour clustering is robust and able to handle "bad" data and favors the grouping into homogeneous age groups.

This study is a pilot to gain experience in the application of analysis methods on primary care data in particular and CPR data in general. Future work should include the automatic selection of useful features for the item set, inclusion of domain knowledge, and correction for outliers. Also, for each type of data, an adaptive cluster algorithm should be designed.

ACKNOWLEDGEMENTS

The author is indebted to Dr. M.C. de Bruijne who posed the question, and kindly provided the data and additional domain information.

REFERENCES

[1] R. Belazzi, B. Zupan, S. Andreassen, E. Keravnou, W. Horn, C. Larizza, N. Lavrac, X.H. Liu, S. Miksch, and C. Popow, editors. *Intelligent Data Analysis In Medicine and Pharmacology*, 1998.

[2] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press / MIT Press, 1996.

[3] D.J. Hand, J.N. Kok, and M.R. Berthold, editors. *Advances in Intelligent Data Analysis*, LNCS 1642. Springer, 1999.

[4] W. Horn, Y. Sharar, S. Lindberg, G. Andreassen, and J. Wyatt, editors. *Artificial Intelligence in Medicine*, LNAI 1620. Springer, 1999.

[5] A.K. Jain, R.P.W. Duin, and Mao J.-C. Statistical pattern recognition: a review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[6] W.A. Kosters, E. Marchiori, and A.A.J. Oerlemans. Mining clusgters with association rules. In Hand et al. [3], pages 39–50.

[7] X.H. Liu, P.R. Cohen, and M.R. Berthold, editors. *Advances in Intelligent Data Analysis: Reasoning about data*, LNCS 1280. Springer, 1997.

[8] Y. Sharar, S. Anand, S. Andreassen, L. Asker, R. Belazzi, W. Horn, E. Keravnou, C. Larizza, N. Lavrac, X.H. Liu, S. Miksch, C. Popow, and B. Zupan, editors. *Intelligent Data Analysis In Medicine and Pharmacology*, 1999.

A Generic Architecture for Knowledge Acquisition Tools in Cardiology

Káthia Maral de Oliveira[1], **Antônio A. Ximenes**[2], **Stan Matwin**[3], **Guilherme Travassos**[1] and **Ana Regina Rocha**[1]

Abstract. Knowledge-acquisition is well known to be a bottleneck activity in the development of knowledge-based systems. Several techniques and tools were proposed to support this process. However, knowledge engineers still have difficulties to understand the problem domain, to apply these techniques and to interact with the experts. Considering that domain-specific tools can be useful for knowledge acquisition, we defined a generic architecture of knowledge acquisition tools and, based on that, we built KED, a Knowledge Editor for Diagnosis in the cardiology domain. KED is part of a general environment that aims at supporting the software development in cardiology domain.

1 INTRODUCTION

Knowledge acquisition is one of the longest and most difficult activities in the development of knowledge-based systems [8]. We could verify this while developing SEC [7], an expert system for diagnosis of acute myocardial infarction. At the beginning of this project the computer science team had a one-day course on basic cardiology concepts to make it possible to start the knowledge acquisition process. Although we used various techniques, we had several problems like the difficulty to schedule a meeting with the experts, or the difficulty to interact with them using their own jargon. One way to assist in this process is to use knowledge acquisition tools. In this context, domain-specific tools [3] are very useful because they can interact directly with the experts using the terminology of the domain and provide basic knowledge for the knowledge engineers in the knowledge acquisition process.

With the goal of developing new expert systems and believing in the importance of assisting in the knowledge acquisition process by using domain-specific tools we defined a generic architecture for knowledge acquisition tools for the cardiology domain. The basic feature of this architecture is to organize independently the domain knowledge and the tasks. Using this architecture we built KED, a knowledge editor for cardiology specialized in diagnosis.

KED is one of the tools of an environment that provides support for the construction, management and maintenance of software products for Cardiology (see [10]). This environment uses embedded domain-knowledge to guide the software developers across the several phases of the software process. It is made of a set of domain-specific tools like KED. These tools differ from more classical ones because they use the vocabulary of the domain (specified in the domain ontology) for all their interactions.

In the following sections, we will first describe how the cardiology knowledge is organized in our environment and how the domain-specific tools use it (section 2). In section 3, we describe the generic architecture of the knowledge acquisition tools exemplified with KED. Finally, in section 4, we present our conclusions and future works.

2 THE DOMAIN-KNOWLEDGE

Knowledge acquisition involves knowledge elicitation and representation. Ontology [1, 2] can be used during this process to facilitate the communication between the experts and the knowledge engineers by establishing a common vocabulary and a semantic interpretation of terms. Ontology is defined as a coherent set of representation terms, together with textual and formal definitions, that embodies a set of representation design choices. It can be used in knowledge-acquisition tools that directly interact with domain experts and effectively avoid errors in the acquired knowledge by constantly verifying constraints on the form and content of the knowledge. Ontology can also provide sharability of knowledge bases and knowledge organization.

To define the ontology for cardiology we used a methodology defined in [5]. To simplify this activity and better organize the domain model, we divided the cardiology domain in sub-domains. Each sub-domain has a group of concepts and relations among them sharing the same semantic context. The sub-domains also have relations between themselves to compose the whole domain. We have identified the following sub-domains: (i) heart anatomy (concepts about the heart structure and the physiology); (ii) findings (concepts used in the physician's investigation process); (iii) therapy (kinds of therapies and their features); (iv) diagnosis (concepts and characteristics that identify syndrome and etiology diagnoses); and (v) pathologies (representing different situations of heart components whose classification and features are important for the purpose of the domain theory of cardiology in the environment). Figure 1 shows some of the concepts of these sub-domains. The ontology was validated by cardiologists and formalized using first order logic and Prolog.

We also identified potential tasks of the domain and mapped them with the domain knowledge. These tasks represent activities such as diagnosis or interpretation that happen in the domain, but are domain-independent (e.g. diagnosis of diseases and diagnosis of machine failures). They are important to specify tools. For cardiology

[1] COPPE / UFRJ - Graduate School of Engineering - Federal University of Rio de Janeiro. Cx Postal 685111 Cep 22945-970. Rio de Janeiro - RJ - Brazil - email: kathia, ght, darocha@cos.ufrj.br

[2] UCCV/FBC- Unit of Cardiology and Cardiovascular Surgery - Fundao Bahiana de Cardiologia - Federal University of Bahia - Rua Augusto Viana s/n Cep 40140-060 Brazil

[3] SITE - School of Information Technology and Engineering University of Ottawa - Canada - email: stan@csi.uottawa.ca

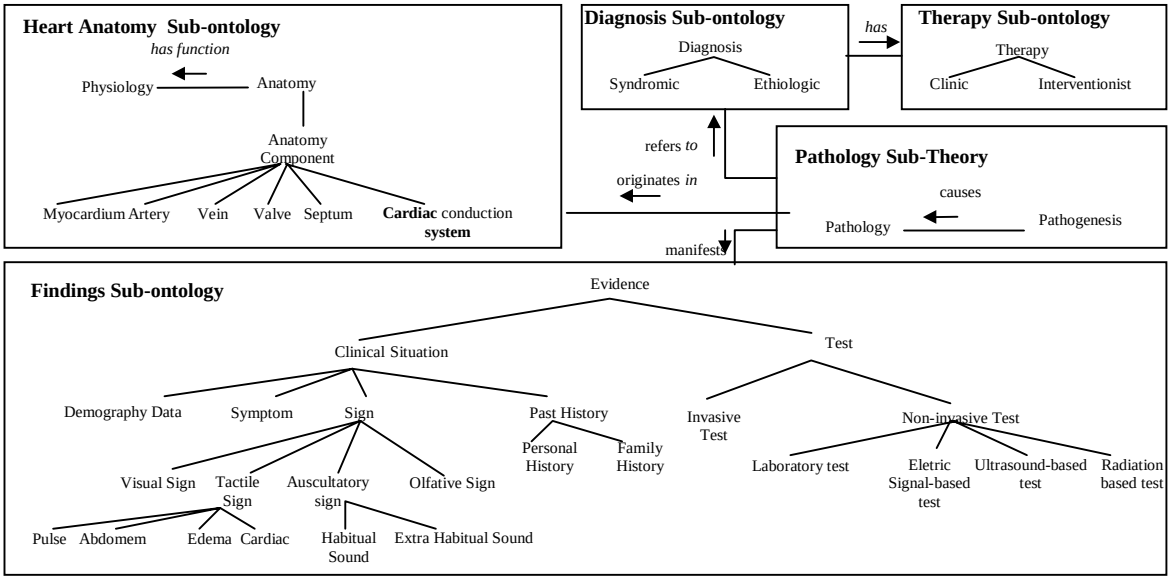


Figure 1. Domain Knowledge for Cardiology

we consider the following main tasks: diagnosis, therapeutic planning, simulation and monitoring. The mapping between the task and the ontology gives an idea of the concepts more closely related to each task. For the diagnosis task, for instance, it is important to consider the findings, pathology and diagnosis sub-ontologies. Similarly, for therapeutic planning, the important sub-ontologies are therapy and pathology; for monitoring, the heart anatomy and findings sub-ontologies; and for simulation, the heart anatomy sub-ontology.

3 KED: A KNOWLEDGE EDITOR FOR DIAGNOSIS IN CARDIOLOGY

Using this domain ontology, we defined a generic architecture to build knowledge acquisition tools. This architecture is based on interactions with the experts (asking them to enter cases) and is composed of three levels (Figure 2): the Knowledge Representation Level, the Generic Process Level and the Operational Level. The *Knowledge Representation Level* is the ontology for cardiology defining the structure and content of the knowledge domain. The *Generic Process Level* is a model of the task specific to each tool (e.g. for KED it will be diagnosis). The *Operational Level* records a set of cases, used to define the generic process model and to find out the knowledge. The tools following this generic architecture can ask information about cases to the expert using the domain language (concepts from the ontology). These cases are stored in a database and can be used later for testing. They are also used to instantiate the domain ontology and, if possible, to define general rules for the specific task.

Using this architecture we built KED, the knowledge editor for cardiology, specialized in diagnosis task. Figure 3 shows KED's conceptual model for the three-level architecture. We can see that the first level consists in the domain sub-ontologies related to the diagnosis task (findings, pathology and diagnosis). The Generic Process Level and the Operational Level are based on analysis patterns defined by Fowler [4]. We also consider the generic task model of systematic diagnosis by causal tracing proposed by KADS [6] to better understand and define the types in the Generic Process Level. In this model, concepts are caused by other concepts and these in turn by others, etc., in a causal hierarchy. This hierarchy is represented in the Generic Process Level by defining general rules (*Norms*) for diagnosis using the cases (set of *qualitative* and *quantitative observations* from the patient) entered in the Operational Level. These rules are formed by justifications of the diagnoses entered by the cardiologists, they are simple diagnostic associations. While entering new justifications the tool generalizes them with the previous ones, so that, in the end, some general rules for the diagnosis of the pathologies are available for future use in interviews during the elicitation process.

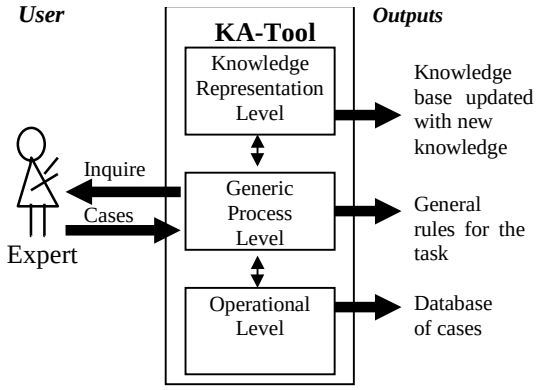


Figure 2. Three -level architecture of the Knowledge Acquisition Tool

KED supports knowledge elicitation in five steps. In the first one, it inquires *observations* from one patient case using the information from the ontology. In the second step, the expert enters a diagnostic for that patient specifying which *observations* are significant for his decision. The third and fourth steps are performed starting from the second case. In this situation, KED analyzes the important observations (chosen in the second step) and generalizes the previous associations (defined for others patient cases for the same pathology) using simple machine learning rules [9]. Then, KED shows this generalization and the cardiologist validates it (fourth step). If he/she disagree,

Table 1. Example of Knowledge Elicitation with KED

	Case 1	Case 2
1st step (expert enters the required information)	Sex =Male, Age=45, Weigh=70, Symptom=chest pain	Sex=Female, Age=60, Weigh=55, Symptom=chest pain
2nd step (expert gives the diagnosis and chooses the significant observations)	Diagnosis Acute myocardial infarction, Sex=Male, Age=45, Symptom=chest pain	Diagnosis Acute myocardial infarction, Sex=Female, Age=60, Symptom=chest pain
3rd step (Ked generalizes the cases entered)		(Sex=female or Sex=male) and Age$\geq$45 and Symptom=chest pain
4th step (expert evaluates the generalization)		Disagree
5rd step (expert provides justifications)	Sex=Male and Age$\geq$45 and Symptom=chest pain	Sex=female and Age$\geq$55 and Symptom=chest pain

KED learns that this generalization should never be proposed again. Finally, the expert gives justifications for that case (that will be generalized in the next cases) and new justifications when he/she did not agree with the generalization in the previous step. Table 1 shows a simple example of this process considering the information sex, age, weigh and symptom. The generalization in the third step was done with the justification provided in case 1. This is just one kind of generalization that the KED uses. In the fifth step the expert provides an association to correct the one gave by the system.

The associations can be translated to Prolog to be used as a first draft knowledge base. They can also be used, by the knowledge engineers, to continue the knowledge acquisition.

KED was validated with several cases collected at UCCV/FBC and considering associations provided by cardiologists.

4 CONCLUSION

Knowledge acquisition tools are an essential support for the development of knowledge-based systems. They can be used directly by the experts and can provide a valuable starting point for the knowledge engineers in the knowledge elicitation process. This paper presents a generic architecture for building knowledge acquisition tools in the cardiology domain along with a specific example: KED, the Knowledge Editor for Diagnosis in Cardiology. The generic architecture relies on a single domain ontology for cardiology and a model of the desired task for each knowledge acquisition tool. In this paper we present how we organize the knowledge and how these tools interact by describing KED.

We are working on this tool to manage conflicting associations defined by the experts. Besides this, we will define tools for other tasks to help in the development of expert systems to be used in intelligent tutors under development at UCCV/FBC.

ACKNOWLEDGEMENTS

We thank Dr. Alvaro Rabelo for his valuable contribution and CNPq, the Brazilian Government financial support for this project.

References

[1] A. Abu-Hanna and W. Jansweijer, 'Modelling domain knowledge using explicit conceptualization', *IEEE Expert*, 53–63, (oct 1994).

[2] B. Chandrasekaran, R. Johsephson, and V. R. Bejamins, '"what are ontologies, and why do we need them?"', *IEEE Software*, 20–27, (Jan/Feb 1999).

[3] H. Eriksson and M. Musen, 'Metatools for knowledge acquisition', *IEEE Software*, 23–29, (may 1993).

[4] M. Fowler, *Analysis Patterns: Reusable Object Models*, Addison-Wesley, 1997.

[5] A. Gómez-Pérez, M. Fernandez, and A. J. Vicente, '"toward a method to conceptualize domain ontologies"', in *Proceedings of Workshop on Ontological Engineering / ECAI96*, (Aug. 1996).

[6] F. Hickman, J. Killin, L. Land, T. Mulhall, D. Porte, and R. M. Taylor, *Analysis for Knowledge-Based Systems: A practical Guide for KADS, Methodology*, Ellis Horwood, 1992.

[7] Rabelo Jr, A. R. Rocha, and K. M. *et al* Oliveira, 'An expert system for diagnosis of acute myocardial infarction with ecg analysis', *Artificial Intelligence in Medicine*, **10**, 75–92, (1997).

[8] K. McGraw and K. Harbison-Briggs, *Knowledge Acquisition: Principles and Guidelines*, Prentice Hall, 1989.

[9] R. S. Michalski, *Machine Learning - An Artificial Intelligence Approach*, volume I, chapter A Theory and Methodology of Inductive Learning, Morgan Kaufmann Publishers, 1986.

[10] K. M. Oliveira, A. R. Rocha, and G. H. Travassos, 'A domain-oriented software development environment for cardiology', in *Proceedings of America Medical Informatics Association conference - AMIA*, Washington, D.C., USA, (Nov. 1999). AMIA.

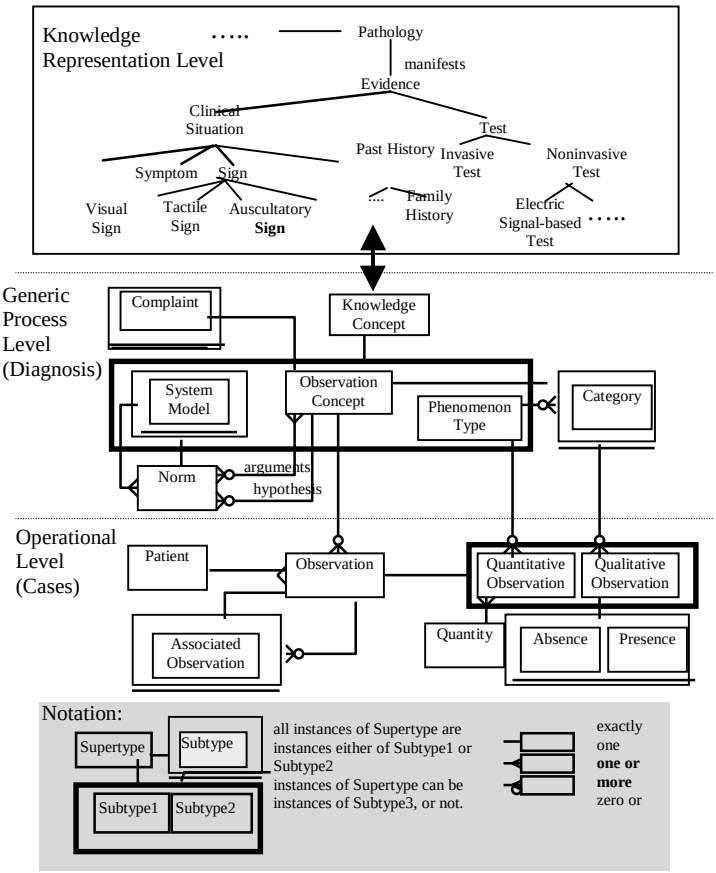


Figure 3. KED Conceptual Model

Mining Knowledge in X-Ray Images for Lung Cancer Diagnosis

Petra Perner

Abstract. Availability of digital data within picture archiving and communication systems raises a possibility of health care and research enhancement associated with manipulation, processing and handling of data by computers. That is the basis for computer-assisted radiology development. Further development of computer-assisted radiology is associated with the use of new intelligent capabilities such as multimedia support and data mining in order to discover the relevant knowledge for diagnosis. In this paper, we present our work on data mining in medical picture archiving systems. We use decision tree induction in order to learn the knowledge for computer-assisted image analysis. We are applying our method to interpretation of x-ray images for lung cancer diagnosis. We are describing our methodology on how to perform data mining on picture archiving systems and our tool for data mining. Results are given. The method has shown very good results so that we are going on to apply it to other medical image diagnosis tasks such as lymph node diagnosis in MRI and investigation of breast MRI.

1 INTRODUCTION

Radiology departments are at the center of a massive change in technology. The ubiquitous radiographic film that has been the basis of image management for almost 100 years is being displaced by new digital imaging modalities such as 1. computed tomography (CT); 2. magnetic resonance (MR); 3. nuclear medicine (NM); 4. ultrasound (US); 5. digital radiography (DF); 6. computed radiography (CR) using storage phosphor imaging plates or film digitizers, 7. digital angiography (DA); 8. MR spectoscopy (MRS); 9. electron emission radiography (EMR).

These digital modalities are continuously refined and new digital applications are being developed. The scientific prognosis is that about 80 of patient imaging examination in radiology department will be performed using digital imaging modalities already at the end of this century [1].

Digital image management systems are under development now to handle images in digital form. These systems are termed Picture Archiving and Communication Systems (PACS)[2][3]. The PACS are based on the integration of different technologies that form a system for image acquisition, storage, transmission, processing and display of images for their analysis and further diagnosis. The main objective of such systems is to provide a more efficient and cost-effective means of examining, storing, and retrieving diagnostic images. These systems must supply the user with easy, fast, reliable access to images and associated diagnostic information.

Availability of digital data within the PACS raises a possibilities of health care and research enhancements associated with manipulation, processing and handling of data by computers. That is a basis for computer-assisted radiology development. However, that will only work if the systems are carefully designed so that they supply sufficient data for the development of decision support systems. Often this aspect has not been considered when implementing a radiology information system.

In this paper, we present our work on data mining for picture archiving systems in medicine. We explain our methodology for performing data mining in picture archiving systems. The experiment described here has provided the methodology for other medical image diagnosis experiments. The experiment has also lead to a variety of other data mining application in medical image diagnosis such as lymph node diagnosis in MRI and investigation of breast MRI[4]. It also has shown the advantage of data mining over other techniques for improving the quality of image diagnosis in medical applications and it provides in the long run the opportunity for the development of fully automatically image diagnosis systems.

In Section 2, we describe the problems by the development of computer-assisted medical diagnosis systems and the application itself. The method used for data mining are described in Section 3 and results are given in Section 4. In another experiment, we use feature subset selection before applying decision tree induction. The method and the results are described in Section 5. Finally, we summarize our experience in Section 6.

2 APPLICATION

Knowledge acquisition is the first step in developing an image interpretation system. The kind of method used for knowledge acquisition depends on the inference method the image interpretation system is based on.

[1] Institute of Computer Vision and Applied Computer Sciences IBaI, Arno-Nitzsche-Str. 45, 04277 Leipzig, Germany, email: ibaiperner@aol.com

The knowledge acquisition process for rule-based system is usually manually done by interviewing a human expert [5]. or by employing interactive knowledge acquisition tools such as e.g. repertory grid [6].

In model-based systems, the knowledge about the objects is represented based on semantic nets that structure the knowledge into concepts and their relations. The language of the semantic net determines the way new knowledge is elicitated. Kehoe et al.[7]. describe a model based system for defect classification of welding seams. The knowledge base is manually maintained by specializing or generalizing the defect classes, their attributes, and attribute values. Schröder et al.[8]. described a system where knowledge acquisition is done automatically based on the language of the semantic net. Although semantic nets seem to be the most convenient way of representing and eliciting knowledge, this method requires a deep understanding of the domain, which is not given a-priori for all applications.

When generalized knowledge is lacking, then case based reasoning [9]. seems to be a proper method. This system is based on a case base consisting of a set of cases. An interpretation is made by determining the closest case or cases in the case base to the actual case and by displaying the value of the closeness measure and the interpretation associated with the similar case of the case base. How the closeness measure should be interpreted is left to the user. The limited explanation capabilities are the main drawback of case based reasoning systems.

We want to develop a knowledge acquisition method for such applications where no generalized knowledge about the domain is available but a large data base of images associated with expert description and interpretation. If we think of the recent trend to picture archiving systems in medicine and other domains, tasks such as these become quite important.

Therefore, the aim of our project is development of knowledge acquisition methods for medical image diagnosis, which can help to solve some cognitive, theoretical and practical problems:

1. Decision model of an expert for specific tasks solution will be reproduced and displayed.
2. It will show the pathway of human reasoning and classification. Image features which are basic for correct decision by expert will be discovered.
3. Developed model will be used as a tool to support decision-making of physician, who is not an expert in a specific field of knowledge. It can be used for teaching of decision-making.

The application of data mining will help to get some additional knowledge about specific features of different classes and the way in which they are expressed in the image (can help to find some inherent non-evident links between classes and their imaging in the picture).

It can help to get some nontrivial conclusions and predictions can be made on the base of image analysis.

For our experiment, we used a database of tomograms of 250 patients with verified diagnoses (80 cases with benign disease and 138 cases with cancer of lung). Patients with small pulmonary nodules (up to 5 cm) were selected for this test. Conventional (linear) coronal plane tomograms with 1 mm thickness of section were used for specific diagnosis.

Original linear tomograms were digitized with step of 100 micron (5,0 line pairs per millimeter) to get 1024 x 1024 x 8 bits matrices with 256 levels of gray.

The use of linear tomograms and such a digitization enabled an acquisition of high spatial resolution of anatomical details that were necessary for the specific diagnosis of lung nodules.

To improve results of specific diagnosis of small solitary pulmonary nodules we used optimal digital filtering and analysis of post-processed images. The processing emphasized diagnostically important details of the nodule and thus helped to improve the reliability of image analysis: the physician was more certain in feature reading and interpretation. The radiologist worked as an expert on this system.

3 METHODS AND APPROACH TAKEN

First, an attribute list was set up together with the expert, which covered all possible attributes used for diagnosis by the expert as well as the corresponding attribute values, see Table 1. We learned our lesson from another exeriment and created an attribute list having no more than three attribute values. Otherwise, the resulting decision tree is hard to interpret and the tree building process stops very soon because of the splitting of the data set into subsets according to the number of attribute values.

Then, the expert collected the database and communicated with a computer answering to its requests. He determined whether the whole tomogram or its part had to be processed and outlined the area of interest with overlay lines and he also outlined the nodule margins. The parameters of optimal filter were then calculated automatically. A radiologist watched the processed image (see Fig. 1) displayed on-line on a TV monitor, evaluated its specific features (character of boundary, shape of the nodule, specific objects, details and structures inside and outside the nodule, etc.), interpreted these features according to the list of attributes and inputted the codes of appropriate attribute values into the database program. Hard copies of the previously processed images from the archive have been used in this work as well.

The collected data set was given as a dBase-file to the inductive machine learning tool.

For the data mining experiment we used our tool DECISION_MASTER [10]. It can create binary and n-ary decision trees from the data. It has several options which makes it possible to specify how numerical features should be partitioned [11] and what method should be used for feature selection. Evaluation of the results can be done by test-and-train and n-fold crossvalidation.

Attribute	Short Name	Attribute Values
Class	*CLASS*	1 malignant 2 benign
Structure inside the nodule	*STRINSNOD*	1 Homogeneous 2 Inhomogeneous
Regularity of Structure inside the nodule	*REGSTRINS*	1 Irregular Structures 2 Regular orderly
Cavitation	*CAVITATIO*	0 None 1 Cavities
Areas with calcifications inside he nodule	*ARWCAL*	0 None 1 Areas with calcifications
Scar-like changes inside the nodule	*SCARINSNOD*	0 None 1 Possibly exists 2 Irregular fragmentary dense shadow
Shape	*SHAPE*	1 Nonround 2 Round 3 Oval
Sharpness of margins	*SHARPMAR*	1 NonSharp 2 MixedSharp 3 Sharp
Smoothness of margins	*SMOMAR*	1 NonSmooth 2 MixedSmooth 3 Smooth
Lobularity of margins	*LOBMAR*	0 NonLobular 1 Lobular
Angularity of margins	*ANGMAR*	0 Nonangular 1 Angular
Convergence of vessels	*CONVVESS*	1 Vessels constantly 2 Vessels are forced away the nodule 3 None
Vascular Outgoing Shadows	*VASCSHAD*	0 None 1 Chiefly vascular
Outgoing sharp thin tape-lines	*OUTSHTHIN*	0 None 1 Outgoing sharp thin tape-lines
Invasion into surrounding tissues	*INVSOURTIS*	0 None 1 Invasion into surrounding tissues
Character of the lung pleura	*CHARLUNG*	0 No Pleura 1 Pleura is visible
Thickening of lung pleura	*THLUNGPL*	0 None 1 Thickening
Withdrawing of lung pleura	*WITHLUPL*	0 None 1 Withdrawing
Size of Nodule	*SIOFNOD*	Numbers (eg, 1.2)in cm

Table 1 Attribute List

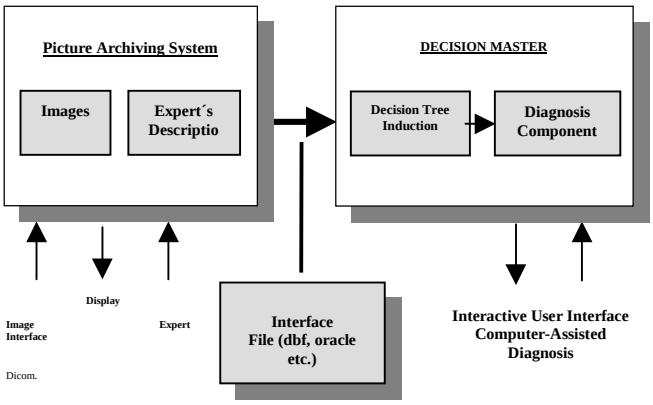


Figure 1 Architecture

Missing values can be handled by different strategies. The tool also provides functions for outlier detections. Once the diagnosis knowledge has been learnt, the rules are provided weather in txt-format for further use in an expert system or the expert can use the diagnosis component of DECISION_MASTER for interactive work. The tool is written in C++ and runs under Windows95 and Windows NT. It has a user-friendly interface and is set up in such a way that it can be handled very easily by non-computer specialists.

Figure 1 shows our structure for a Picture Archiving System combined with the data mining tool.

We used a decision tree induction method for our experiment which creates binary-trees based on maximum entropy criteria [12]. Pruning is done based on reduced-error pruning technique [13]. Evaluation was done by crossvalidation. Besides the error rate we calculate Sensitivity for Class_1 and Specificity for Class_2, which are error criteria normal required for medical applications: $E_{sens} = S_{c1m} / N_{c1}$ $E_{Spec} = S_{C2m} / N_{C2}$ with S_{c1m} the number of misclassified samples of class 1 and N_{C1} the number of all samples of class 1 and S_{C2m} and N_{C2} respectively

4 Results

The unpruned tree consists of 20 leaves. Therefore, the resulting tree has not been shown in this paper. The pruned tree consists of 6 leaves, see Figure 2. Our expert liked the unpruned tree much more since nearly all attributes he is using for decision making appeared in the tree. The expert told us that the attribute *Structure* is very important, also the *attribute Scar-like changes inside the nodule.*

However the expert wonders why other features such as ***Structure*** and some others didn't work for classification. The expert told us that he usually analyzes a nodule starting with its ***Structure***, then tests ***Scar-like changes inside the nodule***, then ***Shape*** and ***Margin***, then ***Convergence of Vessels*** and ***Outgoing Shadow in Surrounding tissues***. Although decision trees represent the decision in a comprehensible format to human, the decision tree might not represent the strategy used by an expert since it is always the attribute appearing first in database and satisfying the splitting criteria that is chosen.

Therefore, we investigated the error rate as main criterion, see Tab. 2 and Tab. 3.

Before pruning	*After pruning*	
	Error Rate	**Error Rate**
(1)	6,857 %	7,428 %
(2)	6,30%	7,3 %

Table 2 Result (1) and Evaluation of Decision tree on Test Data (2)

Accuracy		*Sensitivity/Specifity*			
Human	DT	Class 1		Class 2	
		Human	DT	Human	DT
94,4%	93,2 %	97,5 %	93 %	91,4 %	90 %

Table 3 Comparison between Human Expert and Decision Tree Classification

We did not come close to the expert's performance. One reason might be the choice of attribute values. For some categorical attributes, there are too many categorical values. That causes that during the tree building process the training set is split up into too many subsets with few data samples. As a result the tree building process will stop very soon since no discrimination power is left in the remaining data samples.

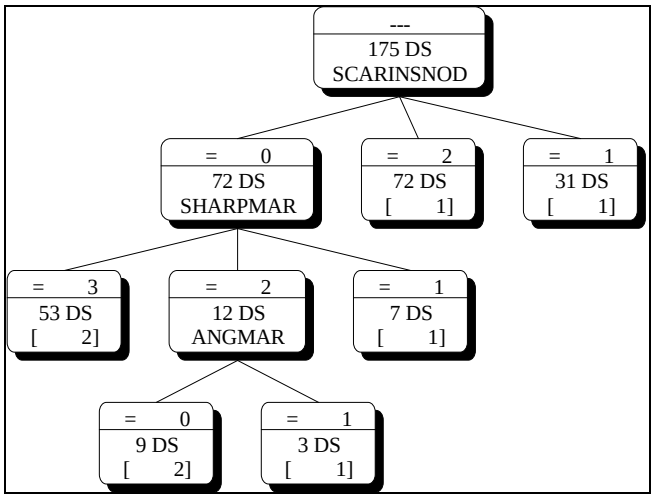


Figure 2. Prunend Decision Tree

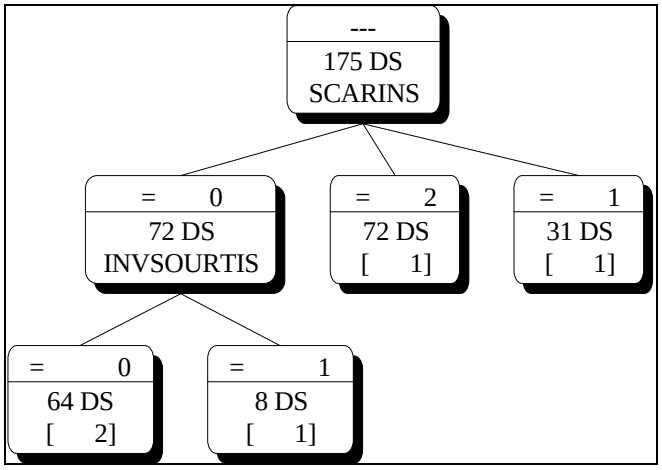


Figure 3. Decision Tree with Feature Subset Selection

5 Feature Subset Selection

Ideally, decision tree induction should use only the subset of features that leads to the best performance.

Induction algorithm usually conduct a heuristic search in the space of possible hypotheses. This heuristic search may lead to induced concepts which depend on irrelevant features, redundant, or correlated features.

The problem of feature subset selection involves finding a "good" set of features under some objective function.

We consider our feature subset selection problems as a problem of finding the set of features which are most dissimilar to each other. Two Features having high similarity value are used by the experts in the same manner for the target concept. They are redundant and can be removed from the feature set.

For our experiment, we used Kruskal´s tau [14] as similarity function and single linkage method to visualize the similarity relation. We observed that *Character of Lung Pleura* and *Within Lung Pleura* are more or less used in the same manner. The expert confirmed this observation. However, on his opinion it is necessary to have both features since sometimes one feature does exist and the other does not exist.

From the dendrogram we created different subsets of features with 10, 13, 15, and 17 features by selecting the most dissimilar features. The first subset included the following 10 features: REGSTRINS, ARWCAL, LOBMAR, CONVVESS, VASCSHAD, OUTSHTHIN, INVSOURTIS, SIOFNOD, SPICMAR, and SHAPE. The next subset of features included three more features with high dissimilarity value and so on. From these subsets Decision Master induced decision trees and calculated the error rate based on cross validation.

We observed that a better error rate can be reached if the decision tree is only induced from a subset of features, see Table 4 and Figure 4. The method used in this paper does not tell us what is the right number of features. This, we can only find out by running the experiment.

Another side effect is that the resulting decision tree is more compact, see Figure 3.

Feature Number	Unprunend Decision Tree Error Rate	pruned Decision Tree Error Rate
19	6,8571	7,428
10	10,85	14,85
13	7,4286	4,5714
15	7,429	4,5714
17	10,28	7,42

Table 4 Error Rate for different Feature Subsets

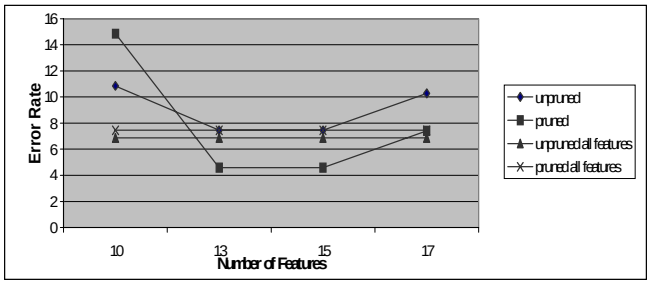


Figure 4 Diagramm Error Rate for different Feature Subsets

6 CONCLUSION AND FUTURE WORK

In this paper, we presented our methodology for data mining in picture archiving systems. The basis for our study is a sufficiently large database with images and expert descriptions. Such databases result from the broad use of picture archiving systems in medical domains.

We were able to learn the important attributes needed for interpretation and the way in which they were used for decision making from this database by applying data mining methods. We showed how the domain vocabulary should be set up in order to get good results and which techniques could be used in order to check reliability of the chosen features.

The explanation capability of the induced tree was reasonable. The attributes included in the tree represented the expert knowledge.

Finally, we can say that picture archiving systems in combination with data mining methods open the possibility of advanced computer-assisted medical diagnosis systems. However, it will not give the expected result if the PACS have not been set up in the right way. Pictures and experts descriptions have to be store in a standard format in the system for further analysis. Since standard vocabulary and very good experts are available for many medical diagnosis tasks this should be possible. If the vocabulary is not a priori available, then vocabulary can be determined by a methodology based repertory grid. What is left is to introduce this method to the medical community, which we are going on to do recently for mammography image analysis and lymph nodule diagnosis. Unfortunately, it is not possible to provide an image analysis systems, which can extract features for all kind of images. Often it is the case that it is not clear how to describe a particular feature by automatic image feature extraction procedures. The expert's description will still be necessary for a long time. However, once the most discriminating features have been found the result can lead in the long run to fully automatic image diagnosis system which is set up for specific type of image diagnosis.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments which helped improve this paper.

REFERENCES

[1] H.U. Lemke, Medical imaging and computer assisted radiology, Tutorial Notes of CAR'91 (The 5-th International Symposium on Computer Assisted Radiology) Berlin, July 2-3 1991.

[2] D.M. Chaney, Coordinate your plans for EMR and PACS investments, Health Management Technology (Aug. 199) vol. 20, no.7, p.24, 26-7.

[3] K. Adelhard, S. Nissen-Meyer, C. Pistitsch, U. Fink, M. Reiser, Funtional requirements for a HIS-RIS-PACS-interface design, including integration of "old" modalities, Methods of Information in Medicine (March 1999) vol. 38, no. 1, p. 1-8.

[4] S. Heywang-Köbrunner, P. Perner, Optimized Computer-Assistet Diagnosis based on Data Mining, Expert Knowledge and Histological Verification, IBaI Report August 1998 ISSN 1431-2360.

[5] P. Perner, „A knowledge-based image inspection system for automatic defect recognition, classification, and process diagnosis.“ Int. Journal on Machine Vision and Applications, 7 (1994): 135-147.

[6] J.H. Boose, D.B. Shema, and J.M. Bradshaw, „Recent progress in Aquinas: a knowledge acquisition workbench,“ Knowledge Acquisition 1 (1989): 185-214.

[7] A. Kehoe and G.A. Parker, „An IKB defect classification system for automated industrial radiographic inspection,“ IEEE Expert Systems (1991) 8, pp. 149-157.

[8] S. Schröder, H. Niemann, G. Sagerer, „Knowledge acquisition for a knowledge based image analysis system,“ In: Proc. of the European Knowledge-Acquisition Workshop (EKAW 88), Bosse, J. and Gaines, B. (ed.), GMD-Studien Nr. 143, Sankt Augustin, 1988.

[9] P. Perner, Case-Based Reasoning for the Low-level and High-level Unit of an Image Interpretation System, Sameer Singh (Eds.), Advances in Patter Recognition, Springer Verlag 1998, p. 45-54.S.

[10] DECISION MASTER http://www.ibai-solutions.de

[11] P. Perner and S. Trautzsch, Multinterval Discretization for Decision Tree Learning, In: Advances in Pattern Recognition, A. Amin, D. Dori, P. Pudil, and H. Freeman (Eds.), LNCS 1451, Springer Verlag 1998, S. 475-482

[12] H.S. Baird and C.L. Mallows, „Bounded-Error Preclassification Trees,“ In: Shape, Structure and Pattern Recognition, Eds. D. Dori and A. Bruckstein, World Scientific Publishing Co., 1995, pp.100-110.

[13] J.R. Quinlain, „Simplifying decision tree,“ Intern. Journal on Man-Machine Studies, 27 (1987): 221-234.

[14] A.Agresti, „Categorical Data Analysis,“ John Wiley Inc. 1990

Patient survival estimation with multiple attributes: adaptation of Cox's regression to give an individual's point prediction

Ann E. Smith*, Sarabjot S. Anand*.

Abstract. In the field of medical prognosis, Cox's regression techniques have been traditionally used to discover "hazardous" attributes for survival in databases where there are multiple covariates and the additional complication of censored patients. These statistical techniques have not been used to provide point estimates which are required for the multiple evaluation of cases and comparison with other outputs. For example, neural networks (NNs) are able to give individual survival predictions in specific times (with errors) for each patient in a data set. To this end, an evaluation of predictive ability for individuals was sought by an adaptation of the Cox's regression output. A formula to transform the output from Cox's regression into a survival estimate for each individual patient was evolved.

The results thus obtained were compared with those of a neural network trained on the same data set. This may have wide applicability for the performance evaluation of other forms of Artificial Intelligence and their acceptance within the domain of medicine.

1 INTRODUCTION AND RATIONALE

Evidence-based medicine, with its attendant requirement for accountability in treatment strategy at the individual patient level, requires that we try to obtain a quantitative assessment of the prognosis for a patient. This prognosis is based on multiple relevant factors or attributes collected about that patient. Extremely precise predictions at the individual level are, of course, impossible because of unexplained variability of individual outcomes [1]. Cox's regression technique [2] has been the standard statistical tool utilised where "censored" patients exist. Censorship means that the event does not occur during the period of observation and the time of event is unknown, but these cases are incorporated into the analysis. Those whose event is unknown, or who are lost to the study (right censored) or new patients introduced into the study (left censored), add to the information on patients whose event time is known (uncensored), at each time interval. Cox's regression is used to derive the hazard ratio, and hazard regression co-efficients, where there are multiple attributes associated with survival and a variable time to an event, e.g. death.

* Faculty of Informatics, University of Ulster at Jordanstown, Newtownabbey, Co. Antrim, N. Ireland. BT37 0QT
Email:{ae.smith,ss.anand}@ulst.ac.uk

However, criticism has been made of traditional statistics in providing prognostic outcomes for individual patients [3]. Neural networks have entered this field [4] and have been shown to possess some advantages in overcoming the drawbacks. These include being able to give point estimates on multiple cases as a system output whilst not having to depend on assumptions of linearity, or the proportionality with time, of the hazard variables. The authors were carrying out research into A.I. methods of modelling the prognosis of colorectal cancer patients so a comparison and evaluation between neural networks and Cox's regression, was sought [5].

We have evolved a formula to calculate outputs of an individual point estimate for each patient, using Cox's regression as a basis for handling both uncensored and censored data, to give outcomes of survival times.

2 THE DATA SET

For purposes of illustration of the methodology, rather than any particularly good prognostic model, we used a local database of clinico-pathological attributes on 216 colorectal cancer patients, which contained details of both uncensored and censored patients. The uncensored patients had a time of death noted in intervals of 1 month, up to 60 months, and censored patients had only records of attendance at clinics after 60 months, which gave a minimum survival. The data collection instrument contained questions of patients demographic details as well as pathological co-variates, such as polarity; tubule configuration; tumour pattern; pathological type; lymphocytic infiltration; fibrosis; venous invation; mitotic count; penetration; differentiation; Dukes stage; obstruction and site.

However, the database could be any large validated data set for any disease process that results in an "event".

3 COX'S REGRESSION

Application of Cox's regression, in SPSS [6] to the multiple covariates has produced a parsimonious model for the hazard, or death rate, h(t), with significant variables of Dukes stage, patient age and fibrosis category, transforming

a hazard baseline, where the covariates are set to zero, which changes with time:-

$$h(t) = [h_0(t)] * \exp \sum_{i=1}^{n} (B_i X_i) \quad (1)$$

where n is the number of explanatory variables, B_i s are the partial regression co-efficients, X_i's the values of the covariates for each patient and $h_0(t)$ is the hazard baseline. It is possible to obtain a survival baseline $s_0(t)$ in discrete time intervals, accounting for censorship, and a survival function s(t):-

$$s(t) = s_0(t)^{\exp \sum_{i=1}^{n} (B_i X_i)} \quad (2)$$

This equation gives a probability of survival at each time interval. Note that the exponential is as before, but is the power term rather than a multiplicative factor. In general, Cox's regression has more commonly been used for comparing hazards to survival in two populations, e.g. patients undergoing different treatment modalities. Cox's regression does not give an output of a direct point survival estimate for each patient, but a survival function in discrete time intervals of h(t). An exact 50% (or a median) chance of survival, is not possible as an output other than by manually reading off an individual survival curve at the approximate half-life. This is because the non-parametric estimates of s(t) are step functions. This does not lend itself to direct comparability with other systems where a time in months, years etc., as an **automated output** from a system where many cases are involved, is required.

4 THE TRANSFORMATION

The aim was to transform this survival function into a formula for a survival estimate, the time term $\hat{t}$, for comparison with other approaches, such as neural networks (NN) which give a point estimate as an output from the system, for each patient in the data set.

Cox' regression makes two main assumptions about the hazard function. Firstly, it assumes that the covariates are independent. Secondly, it assumes proportionality of the hazard covariates with time. The output is a step function in time, which gives the hazard ratio per time interval for all patients with these particular covariates. This does not predict survival (time to event) for individual patients. Thus, we propose a method for providing a point estimate of survival for individual patients given a particular threshold probability of survival. This method, rather than assuming a particular form for the baseline survival, fits a curve to the discrete data points of survival. In the simplest form, the baseline could be linear, however, non- linear baselines may also be investigated in the future. The important issue here is ensuring a "good" fit.

4.1 The hypothesis

If the discrete survival baseline, $s_0(t)$, for uncensored patients, can be shown to be linear by curve-fitting, then it is possible to suggest that $\mathbf{s_0(t) = mt+1}$, where **m** is the slope of the fitted line and **t** is the time variable, and the constant is the probability of survival when t = 0. Note that for $S_0(t)$ to be a survival baseline, $t \in [0, 1/|m|]$

Assuming linearity of $S_0(t)$ and using equation 2 with S(t) = p. In the general case where p is the probability of event, of the survival distribution, we get:-

$$p = (mt + 1)^{\exp \sum_{i=1}^{n} (B_i X_i)} \quad (3)$$

Performing simple algebraic transformations, and taking the logarithm of both sides and extracting t, we now are able to provide a formula for the point estimate of t, as shown below.

$$\hat{t} = [\exp(\exp(-\sum_{i=1}^{n} B_i X_i) * \ln p) - 1] / m \quad (4)$$

Note that this formula (Point Cox) for $\hat{t}$ is only valid if it can be demonstrated that the baseline is linear by empirical testing.

5 AN EVALUATION OF THE APPLICATION

In this section we used the cancer data set as a validation data set for the method proposed above. Firstly, we selected all uncensored patients, where we know the actual survival, in one month intervals up to 60 months, in order to do direct comparisons of time to event.

Here, the important result is to show linearity of the survival baseline and compare the individual patient survival from our method with those of a NN. The overall accuracy of the prognostic model is not important as it is not an accurate measure for the evaluation of the suitability of the covariates for modelling the survival. The overall accuracy is heavily dependent on the quality of the data. This is an evaluation of the comparability of the paradigms.

The first stage of evaluating the method proposed is to show that the survival baseline is linear. For validating the results obtained, 10-fold cross validation was used. Curve fitting to fit the survival baseline of all the patient data, in each of the cross-folds, to a linear curve gave a mean $\mathbf{R^2}$ (goodness-of-fit measure of a linear model) of 0.989. Clearly, the cross-validation of this curve fit implies that the survival baseline for the cancer data set is sufficiently close to linearity. The slope of this was estimated as -5.30704E-05 for this model. The survival baseline for the data set as a whole is given in Figure 1.

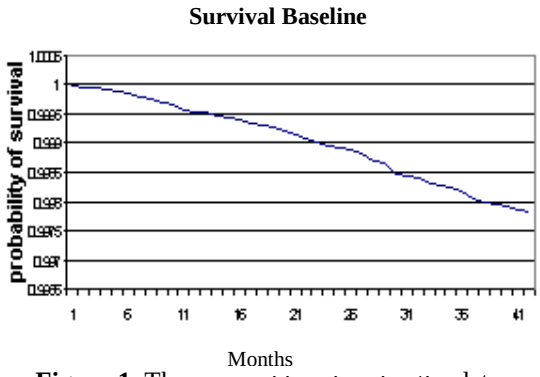


Figure 1. The survival baseline for the data

Note that this is the survival baseline, not the distribution of actual survival of patients. The model's survival baseline slope along with the individual patient's multiple covariates must be used for arriving at that individual patient's survival point estimate.

Having confirmed linearity of the survival baseline, the Point Cox method was then applied to each patient's attributes. Individual predictions for survival of each patient were then generated. Table 1 gives some summary statistics of predicted survivals and actual survivals for those within the uncensored range for which direct comparisons were possible. The actual survivals have a distribution (in months) of mean 18.4, median 15.0, s.d. 13.4, range of 57.0 and skewness of 0.785. The neural net used in this study is part of Clementine Data Mining tool [7]. The learning algorithm used was a variant of the back propagation network, that attempts to optimise the topology of the neural net by starting with a large network and iteratively pruning out nodes in input and output layers.

Table 1 Summary statistics for predicted survivals for the uncensored patients, as output by the two methodologies

	N	Mean	Median	S.D.	Range	Skewness
Point Cox	105	22.8	20.3	13.8	55.6	0.945
NN	105	19.1	17.0	10.1	49.0	0.733

A Wilcoxon's signed rank test comparing directly the Point Cox and the neural network outputs indicated no really significant differences between the two approaches (p= 0.052).

A histogram of the actual survivals recorded, the estimates from the Point Cox method and a neural network are given in Figure 2 for illustration and clarification of the above.

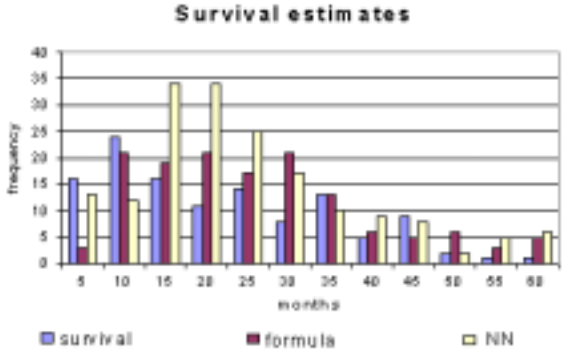


Figure 2. Histogram of survival estimation for the point Cox method, NN output and the actual survivals (uncensored patients)

This particular clinico-pathological model had regression co-efficients of Dukes stage (0.67840), Age (0.03396), Fibrosis category (0.21160). When command syntax was written within SPSS and the Point Cox method was applied to these regression co-efficients and patient covariate values, a survival estimation for each patient was generated. As an example, the distribution output from Cox's for a 54 year old patient whose pathological attributes had the highest grade of Dukes staging (6) and the highest Fibrosis category (6) is shown in Figure 3. The half-life of this is about 13 months, the same value produced from the Point Cox method for the point estimate $\hat{t}$.

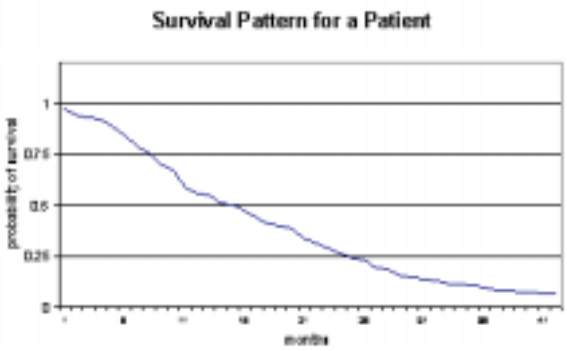


Figure 3. Survival distribution with PointCox method for a patient with covariates as described in text.

The overall results of applying the Point Cox method, within the colorectal cancer data set are given in the first row in Table 2 :-

Table 2 Comparing the actual survivals with both Point Cox method an NN, for uncensored patients.

Model	**Mean |survival-predicted|**	**95% C.I.**	**Wilcoxn**
Point Cox	15.9 months	(7.0, 24.8) months	p = 0.021
NN	14.3 months	(7.2, 21.4) months	p = 0.020

The second column is the mean absolute error between the predicted survival and the actual survival in months for those uncensored patients for whom an actual survival is recorded. The mean actual survival is 18.4 months. It indicates the magnitude of the error in prediction. The Wilcoxon signed ranks test gives a measure of the probability that there is no difference between the observed and predicted survival distributions for the paired data. The results from this test indicate that there are no significant differences in both the Point Cox method and the neural net, when compared with actual survivals.

The errors of the estimates are large, as indicated by the confidence interval (C.I.) but this model only contained limited attributes and no indication of the treatment that the patient received, e.g. surgery, radiotherapy or chemotherapy, was included. We are aware that this particular model is not sufficient in itself for any prediction of survival in a patient, but we would expect the inclusion of information on such factors, if available, to improve the model and reduce the errors.

In the case of the censored patients, the observed survival values are a minimal estimate since they are only attendances recorded at a clinic. The degree of overestimate however, in this situation is valid, but cannot be used for assessment of the performance, as it is not possible to get an exact measure of

the error. The underestimate can be stated in minimal errors, for example, if a censored patient has a minimum survival recorded of 100 months and the predicted survival is 45 months, it can be assessed as an error in prediction of at least 55 months.

Table 3 gives a comparison of the minimal errors given by the Point Cox estimate and a neural network to see if the differences in the paradigms applied to censored patients are significant.

Table 3 T test of results when applying the Point Cox method compared with a neural network, for censored patients.

Model	% Underestimate	Mean error	T value	d.f.	Sig
Point Cox	63.2%	46.8	0.333	9	0.749
NN	93.5%	45.0			

This comparison indicates no significant difference between the two paradigms.

6 DISCUSSION

This paper presents an original technique, for deriving point estimates from Cox' Regression, aimed at enabling the evaluation of other methods e.g. regression trees or NN against the standard technique used in the medical domain. The authors believe that the presented Point Cox method can be used as a simple approach and incorporated into Cox's Regression to obtain point survival estimates for individual patients in any disease process which results in an event, and contains censored patients. This possibly has wide applicability for the prognosis of new patients and can certainly be used for direct comparisons with, and evaluation of, other applications and systems, such as emerging intelligent systems, where many cases and many attributes demand an automated output.

Alternative parametric methods may be employed, eg. the Breslow [8] approach, or other methods of smoothing of survival curves to give a median estimate, Collett [9]. However, we believe that the current method is more easily integrated into programs and understandable to non-statisticians, when using the Cox's model for multivariate analysis.

We believe that the empirically defined linear function allows a practicable simulation of the continuous function that exists in reality with patients dying. Once a model has been derived, from any similar database, with co-efficients for the attributes and slope of the baseline defined, the significant attribute values for each new or current patient can be inserted into the Point Cox to obtain a point estimate. The attributes are defined in the model set up by the user and can be attributes shown to be hazardous for any disease process which results in an "event", such as death. By changing the probability of survival from 0.5 in the formula, representing 50% chance, statements can also be made as to 20%, 80% etc. chance that the patient will live to a certain time. These statements are, of course constrained by the inherent limitations of the model of attributes applied, plus the uncertainty of patient variability. However, this paradigm may be useful for the planning of treatment profiles.

This Point Cox, derived from Cox' regression, is novel in its attempt to overcome the failure of Cox's regression so far to provide point estimates as a direct output. It, however, depends heavily on the assumptions of linearity of the survival baseline produced for both uncensored and censored data up to censoring time, using the regression co-efficients for the attributes hazardous to survival. However, it may be possible to include non-linear functions if curve fitting can show that the survival baseline can be approximated adequately by a specific function.

The analysis of more survival data sets for observing the extent to which linear baselines can be expected, so that this current constraint can be lifted, is left as a future research goal.

ACKNOWLEDGEMENTS

We thank Dr. Peter Hamilton and others at the Dept. of Pathology, Royal Victoria Hospital, Belfast, U.K. for the use of the database on colorectal cancer patients.

We also thank Ms Adele Marshall for expert help on survival analysis.

We acknowledge the Grant, in the form of a Fellowship, from the Medical Research Council, which has enabled continuation of this work.

REFERENCES

[1] M Buyse, P Piedbois. Comment. *Lancet*, **350 (9085),** 1175 1176, (1997)

[2] DR Cox. Regression models and life tables. *J R Stat Soc*, **34**, 187- 220 (1972)

[3] L Bottaci, PJ Drew, JE Hartley, MB Hadfield, R Farouk, PWR Lee, et al. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. *Lancet*, **350 (9076)**, 469-472, (1997)

[4] D Farragi, R Simon. A neural network model for survival data. *Stats Med*, **350**, 72-82, (1995)

[5] S S Anand, A E Smith, P W Hamilton, J.S Anand, J G Hughes, P Bartel. An evaluation of Intelligent prognostic systems for colorectal cancer. *J Art Int Med*,**15(2),** 193-213, (1999).

[6] SPSS for Windows, Release **9.0,** *SPSS inc. Chicago*. 1996

[7] Clementine Data Mining, *User Manual. Integral Solutions Ltd*. 1997

[8] P Breslow. Covariance analysis of censored survival data. *Biometrika* **57**, 579-594, (1974)

[9] D Collett. *Modelling survival data in medical research*. Texts in Statistical Science. Chapman & Hill, London. 1994

Intelligent Data Mining for Medical Quality Management

Wolf Stühlinger[1], **Oliver Hogl**[2], **Herbert Stoyan**[3], and **Michael Müller**[2]

Abstract. In the healthcare sector cost pressure is growing, quality demands are rising and the competitive situation amongst suppliers is mounting. These developments confront hospitals more than ever with the necessities of critically reviewing their own efficiency under both medical and economical aspects. At the same time growing capture of medical data and integration of distributed and heterogeneous databases create a completely new base for medical quality and cost management. Against this background we applied intelligent data mining methods to patient data from several clinics and from years 1996 to 1998. This includes data-driven as well as interest-driven analyses. Questions were targeted on the quality of data, of standards, of plans, and of treatments. For these issues in the field of medical quality management interesting data mining results were discovered in this project.

1 INTRODUCTION

Reforms in the healthcare sector have caused a continuously rising cost pressure during the last years. At the same time quality demands on hospitals and other suppliers of medical services are increasing. Along with an aggravating competitive situation the demand for intensive cost and quality management in all fields of the healthcare sector, in diagnostics as well as in therapeutics and administration, is growing, aiming at the exploitation of efficiency potentials [6]. On the other hand, the introduction of integrated hospital information systems (HIS) and the step-by-step conversion to electronic patient data files enable the capture of large amounts of data and thus a comprehensive documentation of historical diagnostic and therapeutic information on cases. Electronic documentation goes along with standardization efforts, e.g., the definition of diagnostic keys ICD-9 and ICD-10 [1]. Henceforth, distributed, heterogeneous, operative databases can be integrated, consolidated in a data warehouse or hospital information system, and made accessible within clinics or beyond. Rising costs and quality pressure on the one hand and new technologies of data processing on the other hand create both the necessity and the opportunity of a data based quality management in the health care sector.

Objective of a study launched by TILAK, the holding-organization of the hospitals in the Tirol region, and FORWISS was the inspection of supposedly quality relevant criteria for medical quality management in patient data as well as the discovery of new criteria. This includes the search for indices in the data enabling the detection of quality relevant differences in care patterns. Above that, we attempt to discriminate normal clinic stays from stays with complications and good from less good documentation. Applying this knowledge efficiency potentials can be found, countermeasures taken, and compulsory quality standards or guidelines created.

Intelligent data mining incorporates advantages of both knowledge acquisition from data, also known as data mining [2], and knowledge acquisition from experts. This technology integrates domain experts intensely into the knowledge discovery process by acquiring domain knowledge and using it to focus the analyses as well as to filter the findings [5].

Roughly 60.000 treatment cases were available from the clinics administered by TILAK for each of the years 1996 to 1998. Analyses have been carried out within patient groups of selected clinics: eye clinic, dermatological, gynaecological, neurological, and urological clinic.

The study represents a first approach to the application of data mining in patient data in the field of medical quality management in order to discover evidence to improvement potentials regarding the efficiency and quality of clinical processes. It could be shown that substantial implicit knowledge is hidden in the available data which cannot be discovered with conventional methods of analysis straight away.

2 THE TASKS OF MEDICAL QUALITY MANAGEMENT

The tasks of medical quality managers can be described as optimization of clinical processes in terms of medical and administrative quality as well as cost-benefit ratio. The core issues of medical quality management processes are the quality of data, of standards, of plans, and of treatments. These qualities can be measured with different indices. In the case of standards for the length of stay of patients in the clinic, a standard is expressed by an interval for a certain primary diagnosis. For example, for the primary diagnosis "coronary atherosclerosis" (ICD-9-code: 414.0) the standard interval for the length of stay is 2 to 5 days. An adequate index in this case is the ratio of cases which meet the standard in contrast to all cases. Data mining can be deployed by quality managers to solve the following tasks:

- Discover new hypotheses for quality indices for data, standards, plans, and treatments.
- Check whether given quality indices for data, standards, plans,

[1] TILAK – Tiroler Landeskrankenanstalten Ges.m.b.H., Quality Management, Anichstrasse 35, A-6020 Innsbruck, Austria, email: wolf.stuehlinger@tilak.or.at

[2] FORWISS – Bavarian Research Center for Knowledge-Based Systems, Knowledge Acquisition Group, Am Weichselgarten 7, D-91058 Erlangen, Germany, email: {oliver.hogl, michael.mueller}@forwiss.de

[3] University Erlangen-Nuremberg, Department of Computer Science 8, (Artificial Intelligence), Am Weichselgarten 9, D-91058 Erlangen, Germany, email: hstoyan@informatik.uni-erlangen.de

and treatments are still valid.

- Refine, coarsen, and adjust given quality indices for data, standards, plans, and treatments.

Most of these tasks can be supported by means of data mining if the existing knowledge in the domain is intensely considered in the data mining process.

3 THE PROCESS OF INTELLIGENT DATA MINING

Intelligent data mining requires a tight corporation between domain experts, in this case medical quality managers, and data mining experts and consists of data-driven as well as interest-driven analyses. The work is supported by our data mining tool, the *Knowledge Discovery Assistant* (KDA) [4].

3.1 Interest-driven data mining

The interest-driven process can be decomposed into seven core phases.

Acquisition of domain knowledge. Domain knowledge is intensely acquired for intelligent data mining with structured expert interviews before the data mining step. This relates amongst others to the formation of interesting groups and concepts, suppositions and questions which can be derived from the tasks of medical quality management. This knowledge is used to define search spaces, to limit the search as well as to filter and sort results. Thus, the user is given quick access to the most interesting results and the deployment of the gained knowledge is made easier.

Formulation of business questions. We used the new *Knowledge Discovery Question Language* (KDQL) which is designed to enable business users in the medical quality management domain to represent business questions in order to focus data mining queries and to structure data mining results. KDQL abstracts from database terminology, e.g., attribute names, and data mining terminology, e.g., names of data mining algorithms. As just one example the following question has been formulated:

How does personal information influence the deviation from standards?

Refinement of business questions. Most KDQL questions formulated by business users will not be initially translatable because they contain concepts which are not part of the data mining world such as attribute groups or attribute value groups. In order to make those questions processable by data mining methods they have to be refined by the KDA using various concept taxonomies for the components of a question. The above question will be refined into the following questions

How does the age influence the deviation from standards?

How does the sex influence the deviation from standards?

where the attribute group "personal information" is replaced by relevant concepts which correspond to attributes in the database.

Transformation of business questions into data mining queries. Once a question has been made translatable by a set of refinements, the transformation into data mining queries can be started. The transformation of the question object is done by the KDA and corresponds to a mapping of the object to one or many data mining methods or statistical tests. Suitable mappings are determined by a number of criteria:

- Criteria concerning the demands of the user (e.g., simplicity, accuracy)
- Criteria concerning the data (e.g., volume of data, scale types of attributes)
- Process criteria (e.g., stage of analysis, level of iteration)

If more than one data mining method or statistical test appear appropriate a question will be mapped parallel on all of them creating a corresponding number of data mining queries.

Execution of data mining queries. The KDA executes the data mining queries and returns structured findings in relation to the questions of the quality manager.

Processing data mining findings. Most data mining methods overwhelm the user with a flood of results. Therefore, the KDA enriches each finding with a value called interestingness which allows business users focused and flexible access to the large volumes of findings produced by data mining methods [3]. The KDA also supports the user by generating visualizations of findings and allows to navigate in structures of findings, to search and sort the findings as well as to filter uninteresting and select interesting findings which can help solve the tasks of the medical quality management.

Transformation of data mining results into answers. Corresponding to questions in the language of the business user, natural language answers are produced by abstracting data mining results. However, the transformation is not yet supported by the current implementation of the KDA.

Triggering new questions. Viewing the answers often causes the quality manager to come up with follow-up questions which may result in the formulation of new questions.

Figure 1 illustrates this language-oriented model for intelligent data mining.

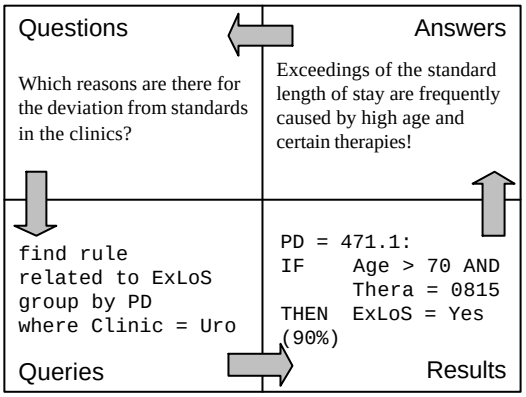


Figure 1. The process model

3.2 Data-driven data mining

Purely interest-driven analyses tend to overlook unexpected patterns in the data. To avoid this shortcoming we also use data-driven types of analyses in addition to the interest-driven analyses. Association rules are mainly deployed for these analyses. Also for data-driven analyses, the business user's interests are applied to structure the findings.

This hybrid approach ensures, that on the one hand users are not overwhelmed by floods of findings which are far beyond their interests and that on the other hand also unexpected patterns do not escape their notice.

4 THE RESULTS FOR MEDICAL QUALITY MANAGEMENT

To give a rough impression of the results which have been found in the analyses we show one example for an interest-driven discovery and one for data-driven discoveries in the following sections.

4.1 Interest-driven discoveries

To the question

> *Which conspicuous subgroups can be identified within differentiable patient groups such as patients with identical diagnoses, who - discernible by the frequent occurrence of complications - should be treated differently, but have not been treated specifically?*

for patients with the diagnosis "fragments of torsion dystonia" (ICD-9-code: 333.8) in the neurological clinic in 1996 the following result has been discovered.

Table 1. Number of exceedings of the standard length of stay in dependency of age for the diagnosis "fragments of torsion dystonia" (ICD-9-code: 333.8)

	Age			
	15-59	60-74	>75	total cases
standard length exceeded	8	9	0	17
standard length not exceeded	11	2	3	16
total cases	19	11	3	33

Here it is obvious, that for "fragments of torsion dystonia" the standard length of stay is exceeded more frequently by patients between 15 and 59 years old than by those between 60 and 74 years old. This is a clear deviation from expectations as one would suppose that elder patients stay longer than younger ones. In the scope of etiology it has to be investigated, if it is in fact the age which can be held responsible for the longer duration of stay. If that is the case, for patients of this age a modified treatment should be considered in the future. If not, the knowledge about the standard length of stay has to be refined for this diagnosis by the corresponding exception.

4.2 Data-driven discoveries

In addition to the above interest-driven data evaluation, data-driven analyses produced another set of interesting findings. Thus, strong associations, e.g., within diagnoses and medical treatments have been discovered. The discovery

IF *one of the single medical treatments [SMT] was "continuous ventriculometry",*

THEN *further SMTs were "respirator therapy" and "burr-hole trepanation". (77%)*

in the neurological clinic in 1997 gives evidence which could increase the reliability of plans. Furthermore, conspicuousnesses regarding the length of stay with several influence factors, such as

IF *one of the SMTs was "combined strabotomy",*

THEN *the overall length of stay is between 2 and 6 days. (92%).*

and

IF *primary diagnosis category was "benign neoplasias of skin" (ICD-9-code: 216),*

THEN *the case fell below the lower limit of length of stay, because it was an out-patient treatment. (100%)*

have been discovered in the eye clinic in 1996. The first rule can aid again an increased reliability of plans for the management of resources and the second allows conclusions that the lower limit of length of stay for primary diagnosis category "benign neoplasias of skin" is not adequate.

5 CONCLUSIONS AND FUTURE WORK

We have shown, that intelligent data mining in addition to conventional analyses and statistical studies in patient data can deliver further evidence for medical quality management. In detail, measures for the quality of data, of standards, of plans, and of treatments can be improved by intelligent data mining. The compliance with given measures can be evaluated, given measures can be modified to better suit the requirements and new measures can be found.

However, the quality management department is usually only in charge of providing the indices. Mostly, measures have to be taken by physicians in the clinics, and it is their acceptance of the results which is required to ensure that they are put into action. Therefore, understandable results in a clear and adequate presentation are indispensable.

Further works concern primarily the statistical consolidation of the results, support for the conversion of conclusions as well as the development of an adapted automated data mining process and the introduction of this technology as part of a comprehensive and permanent set of controlling instruments.

REFERENCES

[1] Brown, F. (ed.): *ICD-9-CM Coding Handbook*, American Hospital Association, Chicago, 1999.

[2] Fayyad, U.; Piatetsky-Shapiro, G. et al.: *From Data Mining To Knowledge Discovery: An Overview*, in: Fayyad, U.; Piatetsky-Shapiro, G. et al. (eds.): *Advances in Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, California, 1996, pp. 1-34.

[3] Hausdorf, C.; Müller, M.: *A Theory of Interestingness for Knowledge Discovery in Databases Exemplified in Medicine*, in: Lavrac, N.; Keravnou, E. et al. (eds.): *First International Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-96)*, Budapest, Hungary, August 1996, pp. 31-36.

[4] Hogl, O.; Stoyan, H. et al.: *The Knowledge Discovery Assistant: Making Data Mining Available for Business Users*, in: Gunopulos, D.; Rastogi, R. (eds.): *Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD-2000)*, Dallas, Texas, May 2000, pp. 96-105.

[5] Müller, M.: *Interessantheit bei der Entdeckung von Wissen in Datenbanken*, PhD thesis, University Erlangen-Nuremberg, Erlangen, Germany, 1998.

[6] Tweet, A.; Gavin-Marciano, K.: *The Guide to Benchmarking in Healthcare: Practical Lessons from the Field*, American Hospital Publications, Chicago, 1998.

Confounding Values in Decision Trees Constructed for Six Otoneurological Diseases

Kati Viikki[1], **Erna Kentala**[2], **Martti Juhola**[1] and **Ilmari Pyykkö**[3]

Abstract. In this study, we examined the effect of example cases with confounding values on decision trees constructed for six otoneurological diseases involving vertigo. The six diseases were benign positional vertigo, Menière's disease, sudden deafness, traumatic vertigo, vestibular neuritis, and vestibular schwannoma. Patient cases with confounding values were inserted into original vertigo data and decision trees were constructed. Confounding values made classification tasks more difficult and decreased true positive rates and accuracies of decision trees. Despite decreased true positive rates and accuracies, new decision trees organised confounding values in a reasonable way into the reasoning process. The occurrence of confounding values simulates better the real life classification tasks.

1 INTRODUCTION

Diagnosis of a vertiginous patient is a difficult task even for specialised otologists. To assist this task, an otoneurological expert system ONE [1] was developed. In addition to diagnostic purposes, it can be used as a tutorial guide for medical students. ONE is beneficial for research, too, due to its capability to store data. In addition to expert systems, we have been interested in other computerized methods, such as genetic algorithms [2], decision trees [3], and neural networks [4], which can be used to support decision making process. We have applied these methods to acquire diagnostic knowledge for benign positional vertigo, Menière's disease, sudden deafness, traumatic vertigo, vestibular neuritis, and vestibular schwannoma [5-7], which are the six largest patient groups in the database of ONE [8].

Due to symbolic knowledge representation, decision trees are suitable for medical classification tasks, in which explanations for decisions are needed. Decision trees generated in the previous study [6] were intelligible and, overall, their true positive rates and accuracies were high. Sudden deafness was the most difficult disease to identify, partly due to the small number of these cases (only 3.7% of all the cases). Recently, data collection was continued to increase the number of example cases especially in the group of sudden deafness. New patient cases were inserted into the database of expert system ONE and subsequently retrieved for decision tree induction. New patient cases had confounding values, which weakened learning results [9] compared with results of the earlier study [6]. Because of the quite small number of cases with confounding values, these values did not occur in the constructed decision trees. This led us to a hypothesis that giving a large enough number of example cases for the decision tree program may result in decision trees which incorporate confounding values. In this study, we retrieved more cases with confounding values from the database of the expert system ONE and constructed decision trees using the See5 decision tree program [10], a descendant of C4.5 [11]. The constructed decision trees were compared with the trees of the earlier study [6].

[1] Department of Computer and Information Sciences, FIN-33014 University of Tampere, Finland, email: Kati.Viikki@mail.cs.uta.fi
[2] Department of Otorhinolaryngology, 00029 Helsinki University Central Hospital, Finland
[3] Department of Otorhinolaryngology, Karolinska Hospital, 17176 Stockholm, Sweden

2 MATERIALS AND METHODS

The data were collected at the Vestibular Laboratory of the Department of Otorhinolaryngology, in Helsinki University Central Hospital, Finland. In the data collection, the expert system ONE was employed. The database of ONE [12], stores a large amount of detailed information about patients. The information is presented in a total of 170 attributes. Use of the expert system does not require the user to answer all the questions, and accordingly, there are some attributes having a lot of missing information. The attributes can be divided into the following categories [12]:

- Patient demographics and referring physician.
- Symptoms: vertigo, hearing loss, tinnitus, unsteadiness, headache, anxiety, and neurological symptoms.
- Medical history: use of ototoxic drugs, head trauma, ear trauma and noise injury, ear infections and operations, specific infections and examinations, and other diseases.
- Findings: clinical findings, otoneurological data, audiometric data, imaging data, and fistula testing.

During the development of ONE, a database of 1167 vertiginous patients was collected prospectively [8]. Otologists were able to confirm the diagnosis of 872 patients, of which 746 belonged to the six largest patients groups. A linear discriminant analysis done by the otologists revealed that some patients had confounding values [8]. By confounding values we mean symptoms and signs that are not related to the current disease, but are rather caused by earlier diseases, medication, or some other factor [8]. Patients had age-related hearing loss, chronic noise exposure combined with hearing loss, chronic diseases such as diabetes, cardiac arrhythmia or major stroke. These diseases and processes caused signs and symptoms that interfered with the discrimination of diseases. Exclusion of patient cases having confounding values finally resulted in an 'original' data set with 564 cases [8]. This data set was used in the earlier studies with machine

learning methods [5-7]. Recently, 76 new cases were inserted into the database of ONE. Especially new benign positional vertigo (BPV) cases (16) and vestibular neuritis (VNE) cases (8) had confounding values. For example, in the original data set, BPV cases did not have hearing loss symptoms. Of the 16 new BPV cases, again, five cases had hearing loss symptoms (Table 1).

Table 1. Distributions for attribute hearing loss.

Benign positional vertigo	Hearing loss	N	%
59 original cases	No	59	100.0
	Yes	0	0.0
16 new cases	No	11	68.8
	Yes	5	31.2
48 cases in extended data	No	17	36.2
	Yes	30	63.8
	Missing	1	

Decision tree tests with the 76 new cases and the 564 previous cases revealed that confounding values reduced the classification ability of decision trees [9]. Confounding values did not, however, occur in the constructed trees because of the quite small number of cases having these values. For this study, 48 BPV cases and 40 VNE cases, discarded in the earlier study [8] because of the confounding values, were retrieved from the database of ONE. These cases were combined to the original data set and to the 76 new cases resulting in the extended data set of 728 cases (Table 2).

Table 2. Data sets.

		Original		Extended	
Diagnosis		N	%	N	%
Benign positional vertigo	BPV	59	10.5	123	16.9
Menière's disease	MEN	243	43.1	283	38.9
Sudden deafness	SUD	21	3.7	30	4.1
Traumatic vertigo	TRA	53	9.4	56	7.7
Vestibular neuritis	VNE	60	10.6	108	14.8
Vestibular schwannoma	VSC	128	22.7	128	17.6
Total		564	100.0	728	100.0

Decision trees were constructed in the form of one disease (positive cases) versus other diseases (negative cases) using the decision tree generator See5 [10]. From the 170 attributes of the expert system ONE, 110 [6] were used in the decision tree construction. The group of 110 attributes was formed on the basis of the physicians' knowledge; for example, attributes having a large number of missing values were excluded [6].

3 RESULTS

The measures used to evaluate the performance of the constructed decision trees were true positive rate (*TPR*) and accuracy (*ACC*):

$$TPR = (Tpos / Pos) \cdot 100\%,$$

where *Tpos* is the number of correctly classified positive cases and *Pos* is the total number of positive cases, and

$$ACC = ((Tpos + Tneg) / N) \cdot 100\%,$$

where *Tneg* is the number of correctly classified negative cases and *N* is the number of all cases. Table 3 presents estimated true positive rates and accuracies given by 10-fold cross-validation [11]. These estimates were compared with the results of the original data set [6]. Furthermore, the experienced physician evaluated decision trees by scrutinising them branch by branch.

Table 3. True positive rates, accuracies and number of attributes for decision trees constructed from original (ORG) and extended (EXT) data sets.

	True positive rate %		Accuracy %		Attributes (N)	
	ORG	EXT	ORG	EXT	ORG	EXT
BPV	98.3	71.5	99.3	92.0	3	18
MEN	98.8	95.1	94.1	91.7	15	18
SUD	52.4	56.7	98.0	98.1	7	5
TRA	96.2	87.5	99.3	98.1	3	6
VNE	98.3	80.6	99.5	96.2	4	9
VSC	82.0	78.1	95.2	95.3	9	11

The true positive rate for benign positional vertigo reduced from 98.3% to 71.5%, and the accuracy from 99.3% to 92.0%. The number of attributes used in the decision tree increased from 3 to 18 (Table 3). In the decision tree constructed from the original data set, the root attribute was hearing loss. This tree classified all cases having hearing loss as not having BPV. In the branch corresponding cases not having hearing loss, frequency of vertigo attacks and occurrence of head injury were tested. In the new decision tree (Figure 1), effects of confounding values are seen in both subtrees branching from the root. For patients not having hearing loss (hearing loss = 0), the attribute concerning Tumarkin-type drop attacks was used to discriminate BPV patients from patients with Menière's disease or vestibular schwannoma. For patients having hearing loss, normal finding in electronystagmography and posturography confirmed BPV diagnosis, whereas deviant findings fit better for vestibular schwannoma and Menière's disease. The audiometry in these BPV patients revealed mostly a mild hearing loss.

For Menière's disease, the true positive rate reduced from 98.8% to 95.1%, and the accuracy from 94.1% to 91.7%. The number of attributes increased from 15 to 18. An essential triad of hearing loss, vertigo, and tinnitus [8] was found in both trees. Further, strong nausea, Tumarkin-type sudden slipps of falls, and fluctuation in hearing appeared in a sensible way in the trees. Some attributes of the old tree were replaced in the new tree by new attributes more valuable in the diagnostic work-up. These new attributes held descriptive characteristics of vertigo and visual blurring.

The true positive rate of the decision tree constructed for sudden deafness increased from 52.4% to 56.7%. The accuracy remained almost the same, 98.1%. Type of hearing loss, which is an essential attribute in diagnosing sudden deafness [8], occurred in both trees, as well as the fluctuation in hearing. Other attributes concerned injury and detailed information about hearing loss. The new decision tree contained five attributes, two attributes less than the tree constructed from the original data set.

The true positive rate for traumatic vertigo reduced from 96.2% to 87.5%, and the accuracy from 99.3% to 98.1%. The number of attributes used in the decision tree constructed for traumatic vertigo increased from three to six. Both trees tested attributes concerning head injury. New attributes concerned nausea and detailed information about head trauma.

For vestibular neuritis, the true positive rate reduced from 98.3% to 80.6%, and the accuracy from 99.5% to 96.2%. The decision tree contained nine attributes, five attributes more than the old tree. An important attribute in both decision trees concerned the low frequency of vertigo attacks. Duration of vertigo attack and hearing loss appeared in both trees, also. New attributes concerned

descriptive characteristics of vertigo, movement difficulties, occurrence of head injury, gain latency in pursuit eye movements, and tone burst audiometry at the frequency of 2 kHz. Neurological findings appearing in the new decision tree were also important.

```
Hearing loss = 0:
:...Duration of vertigo attack > 2: not
:  Duration of vertigo attack <= 2:
:  :...Injury = 1:
:      :...Duration of tinnitus <= 5: not
:      :  Duration of tinnitus > 5: bpv
:      Injury = 0:
:      :...Frequency of vertigo attacks <= 1: not
:         Frequency of vertigo attacks > 1:
:         :...Cranial nerve palsy = 0: bpv
:            Cranial nerve palsy = 1:
:            :...Tumarkin-type drop attacks <= 1: bpv
:               Tumarkin-type drop attacks > 1: not
Hearing loss = 1:
:...Position induced vertigo <= 39.5: not
   Position induced vertigo > 39.5:
   :...Duration of vertigo attack > 2: not
      Duration of vertigo attack <= 2:
      :...Stapedectomy = 1: bpv
         Stapedectomy = 0:
         :...Duration of hearing loss <= 4: not
            Duration of hearing loss > 4:
            :...Head trauma = 1: not
               Head trauma = 0:
               :... Injury = 1:
                  :...Duration of tinnitus > 5: bpv
                  :  Duration of tinnitus <= 5:
                  :  :...Rotational vertigo <= 40: not
                  :      Rotational vertigo > 40: bpv
                  Injury = 0:
                  :...Ear illness = 1:
                     :... Audiometry at 500 Hz, right ear <= 60: bpv
                     :    Audiometry at 500 Hz, right ear > 60: not
                     Ear illness = 0:
                     :...Spontaneous nystagmus > 0: not
                        Spontaneous nystagmus <= 0:
                        :...Age at first vertigo symptoms <= 52:
                           :...Posturography, eyes open <= 0.75: bpv
                           :  Posturography, eyes open > 0.75: not
                           Age at first vertigo symptoms > 52:
                           :...Audiometry at 2000 Hz, right ear <= 35: bpv
                              Audiometry at 2000 Hz, right ear > 35: not
```

Figure 1. New decision tree for benign positional vertigo (bpv) and other diseases (not).

The true positive rate for vestibular schwannoma reduced from 82.0% to 78.1%. The accuracy remained almost the same, 95.3%. An important attribute occurring in both decision trees was caloric asymmetry: in vestibular schwannoma, this value is high. The attribute concerning Tumarkin-type drop attacks discriminated VSC patients, who seldom have these drop attacks, from Menière's disease patients, who have them often. Attributes concerning vertigo, hearing loss, use of ototoxic or vestibulotoxic drugs, and results of computerized tomography appeared also in both trees. Three new attributes were hearing loss of left ear, headache, and tone burst audiometry at the frequency of 1 kHz. The new decision tree showed that about half of VSC patients do not have vertigo symptoms.

4 DISCUSSION

Overall, the generated decision trees were intelligible, although we found some strange branches formed by chance. New attributes occurring in the decision trees were sensible. Attributes concerning confounding values were found especially in the decision tree constructed for benign positional vertigo. In the previous studies [6,8], test results were of minor value in classification of these six diseases. Results of this study suggest that they are important when cases with confounding values are classified.

True positive rates decreased for all diseases, except for sudden deafness, which had the true positive rate of 56.7%, 4.3% more than for the decision tree constructed from the original data set. Sudden deafness seems to be a difficult disease to identify, partly due to the small number of cases, partly due to its nature [5-9]. For Menière's disease (95.1%), traumatic vertigo (87.5%), and vestibular schwannoma (78.1%), the decrease in the true positive rate was less than 10%. The largest decreases (26.8% and 17.7%) were found for benign positional vertigo (TPR of 71.5%) and vestibular neuritis (TPR of 80.6%), respectively. These large decreases can be explained by confounding values, which made BPV and VNE cases to resemble more Menière's disease cases. Overall, classification accuracies reduced slightly, varying from 91.7% to 98.1%. Reductions in true positive rates and accuracies agreed with results of the linear discriminant analysis and the expert system ONE [8,9].

Confounding values made classification tasks more difficult. Although true positive rates and accuracies decreased, the constructed decision trees are valuable. The benefit of the new trees is that they simulate more the real life situation with patients who have confounding symptoms. The value of otoneurological tests increases in the diagnostic work. In the future, our aim is to find different ways to handle confounding values in the reasoning process.

REFERENCES

[1] E. Kentala *et al.*, 'Otoneurological expert system', *Annals of Otology, Rhinology & Laryngology*, **105**, 654-658, (1996).

[2] D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA, USA, 1989.

[3] J.R. Quinlan, 'Induction of decision trees', *Machine Learning* **1**, 81-106, (1986).

[4] L.M. Fu, *Neural Networks in Computer Intelligence*, McGraw-Hill, Singapore, 1994.

[5] E. Kentala *et al.*, 'Discovering diagnostic rules from a neurotologic database with genetic algorithms', *Annals of Otology, Rhinology & Laryngology*, **108**, 948-954, (1999).

[6] K. Viikki *et al.*, 'Decision tree induction in the diagnosis of otoneurological diseases', *Medical Informatics & The Internet in Medicine*, **24**, 277-289, (1999).

[7] M. Juhola *et al.*, Neural network recognition of otoneurological vertigo diseases with comparison of some other classification methods. In W. Horn *et al.* (Eds.) *Artificial Intelligence in Medicine, Lecture Notes in Artificial Intelligence 1620*, Springer-Verlag, Berlin Heidelberg, pp. 217-226, 1999.

[8] E. Kentala *et al.*, 'Characteristics of six otologic diseases involving vertigo', *American Journal of Otology*, **17**, 883-892, (1996).

[9] K. Viikki *et al.*, Building training data for decision tree induction in the subspecialty of otoneurology, accepted to Medical Informatics Europe 2000, Hannover, Germany, August 27 - September 1, 2000.

[10] J.R. Quinlan, See5, version 1.07a, http://www.rulequest.com/, 1998.

[11] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, USA, 1993.

[12] E. Kentala *et al.*, 'Database for vertigo', *Otolaryngology, Head and Neck Surgery*, **112**, 383-390, (1995).

DOPAMINE – A tool for visualizing clinical properties of generic drugs

C.J. Wroe[1], W.D. Solomon[1], A.L. Rector[1] and J.E. Rogers[1]

Abstract. Visualization tools are becoming more important as ontologies increase in size and complexity. One example of a large ontology is the Drug Ontology developed at the Medical Informatics Group using GALEN classification techniques. It will provide a reference terminology for PRODIGY; a project providing computerized medicine prescribing guidelines for primary healthcare in the UK. While developing the Drug Ontology we have needed a customized visualization tool for authoring and checking ontology entries. The software tool (*Drug Ontology Production And MaIntenance eNvironmEnt or DOPAMINE*) provides a tabular summary of sets of ontology entries.

DOPAMINE has improved the consistency and speed of authoring entries, and provided a means of highlighting, previously hard to detect, authoring errors. Although specific to the drug ontology, the techniques used could be generalized for other ontologies.

1 INTRODUCTION

Tabular summaries are an effective means of visualizing structured data, and are the mainstay of database and spreadsheet user interfaces. Several research groups have used tables to present summaries of knowledge base entries.

The Generic Knowledge Base Editor (GKB – Editor) is being developed at SRI International's Artificial Intelligence Center. It is a tool for 'graphically browsing and editing knowledge bases across multiple Frame Representation Systems (FRSs) in a uniform manner' (see [1]). It has a spreadsheet viewer facility but is limited to instances of a knowledge base. The Drug Ontology in contrast describes only classes of drug concepts, such as 'All drugs containing atenolol hydrochloride'.

Conceptually Oriented Description Environment (CODE4) has been developed at the Department of Computer Science, University of Ottawa [2]. Using a 'Property Comparison Matrix', it is possible to visualize the similarities and differences between the properties of concepts. These functions are very similar to those needed by DOPAMINE. A related display is the 'Property Inheritance Matrix' which displays values of a properties as they change down a subsumption hierarchy. The tool can be used for both retrieval and authoring of information.

Protégé is an environment developed at Stanford Medical Informatics for building reusable ontologies and problem solving methods [3]. The environment includes the ability to use a table widget. The table widget is designed to simplify authoring of instances in a knowledge base [4]. The table is regarded as an integral part of the knowledge base, rather than an independent method of visualising the knowledge.

Within our department, Gary Ngg, a PhD student, has been developing tools for visualizing ontologies authored in Grail. Tables show summaries of Grail properties. 'Lenses' can then be superimposed over the table to visualise the result of user defined queries on the table [5].

The development of DOPAMINE was born out of a necessity to provide a user interface with which large and verbose drug ontology class descriptions could be authored and checked easily and rapidly. It draws on previous research on tabular based summaries and applies it to the specific task within the Drug Ontology project.

1.1 Rational behind the drug ontology

The Prodigy Project is developing prescribing guidelines for UK primary healthcare [6]. These guidelines are triggered when the doctor enters a diagnosis, and guide the doctor through prescribing decisions. The project is now in its third phase and is developing more complex guidelines for chronic diseases. These guidelines are active over several months of a patient's chronic condition and need to reference terms in the patient record, for example, to detect the patient's current medication. With this in mind, they recognized the need for a scalable terminology solution.

Through the GALEN and GALEN In USE programs [7], we have been developing methods for creating and maintaining scalable terminologies in the medical domain. The work centres on the use of ontologies - explicit formal representations of concepts [8]. These allow communication, reuse, and representation of medical concepts in a logical system [9]. A description logic, named Grail, has been used to implement the ontology [10]. Description logic provides classification and consistency services for the ontology. To date, ontologies have been created that cover anatomy, basic physiology, basic pathology and basic medical devices. Together these form the Common Reference Model (CRM) [11].

The 'Drug Ontology' builds on existing ontologies of pathology and physiology to create formal descriptions of a generic drug's clinical properties. These properties comprise: ingredients, formulation, indications, contraindications, cautions, mechanism of action, interactions, side effects and clinically relevant pharmacokinetics [12].

[1] Medical Informatics Group, Department of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL, UK, email: cwroe@cs.man.ac.uk
& Open Galen Organisation, http://www.OpenGalen.org.

The project is producing descriptions of primary care relevant generic drugs in the British National Formulary (BNF) [13]. There are approximately 1500 such entries in the BNF. During the project we have found that the explicit formal ontology descriptions, soon become verbose and difficult to read by human readers. Information has to be included in the descriptions that human readers would automatically infer, and so appear redundant.

1.2 Intermediate Representation

To help simplify authoring ontology descriptions an Intermediate Representation (IR) was previously developed in the GALEN in USE Project. IR is a simplified formal language used to describe the definition and properties of a concept [14][15]. The IR is automatically translated into the lower level language Grail, with which the description logic classifier operates. Translation rules are authored for each domain, allowing some degree of ambiguity in each Intermediate Representation.

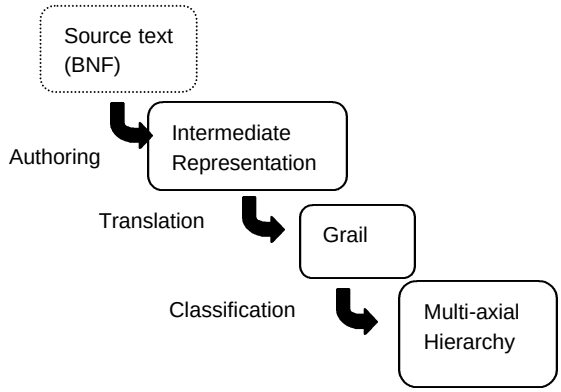


Figure 1. Processes involved in producing a drug ontology description using intermediate representation.

Each term description (called a dissection) starts with the keyword MAIN which is interpreted in a domain specific way. For example, in the drug ontology, it is translated to Grail as 'Drug which hasIngredient'.

A set of terms (called descriptors) and semantic links, follow the MAIN keyword, which specify the terminological definition and properties of the concept being described. Indentation of the links is used to specify which descriptors are being linked.

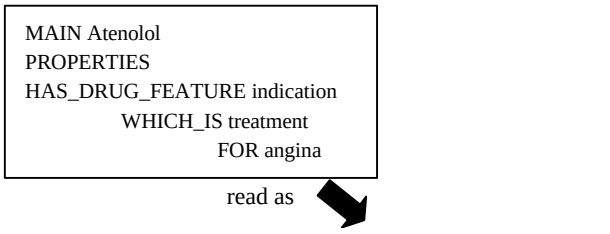


Figure 2. How Intermediate Representation should be interpreted.

If ambiguity exists *within* a domain, as with the term 'indication' in the drug domain, the exact meaning of the term has to be specified at the level of the Intermediate Representation. As the number of properties of a drug increases it can become difficult to read (see Figure 3).

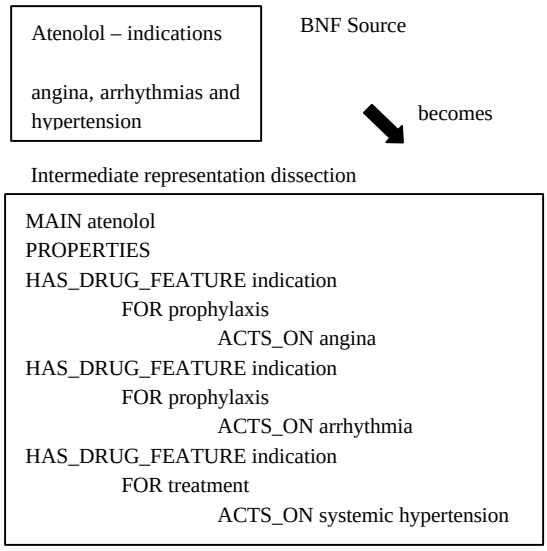


Figure 3. The result of authoring an Intermediate Representation dissection for a sample of source text.

2 METHOD

2.1 Define queries

The description of each type of property follows a stereotyped pattern. For example, all indication properties begin 'HAS_DRUG_FEATURE indication FOR ..'. The first step is to manually author a query. This will be used by the application to organise properties of the same type, for example, all indications. These queries are specified using a purpose built entry tool.

2.2 Execute queries

The user selects a group of drug descriptions to be analyzed. For example, the set of all beta-adrenoceptor blocking drugs. The drug descriptions within the set are checked to see if any properties match one of the predefined queries. If a property does match it is added to a list of detected properties to be presented to the user. For example an indication query will detect and collate all properties beginning with the structure 'HAS_DRUG_FEATURE indication FOR ..'

2.3 Present results as a view with which the user can interact

A table is constructed with a list of properties detected along the vertical axis and the list of drugs examined on the horizontal axis. The list of properties is grouped by the queries that the user has previously defined. Properties are further grouped by common initial structure. This process is explained in figure 4. The top table shows the view that would be presented if the common structure was repeated. The bottom table is an extract from the DOPAMINE tool and shows indications for Atenolol. Users can also add to drug ontology descriptions using the tool.

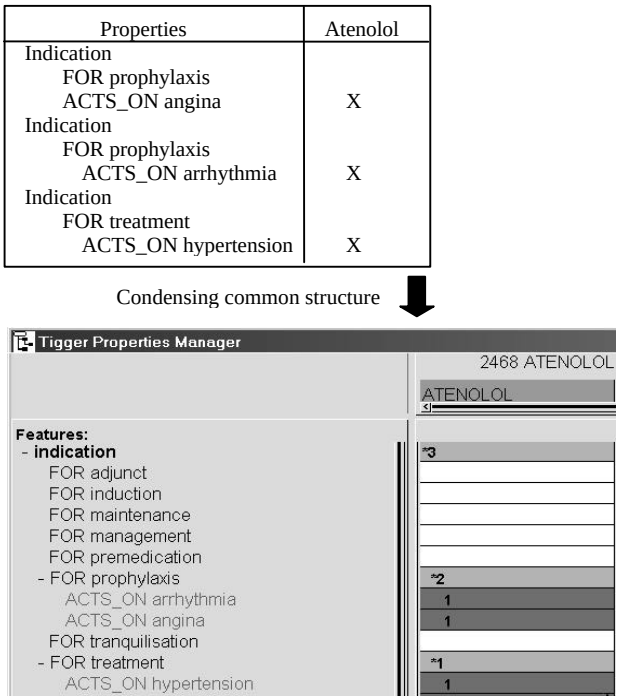


Figure 4. Presentation of properties in a condensed format.

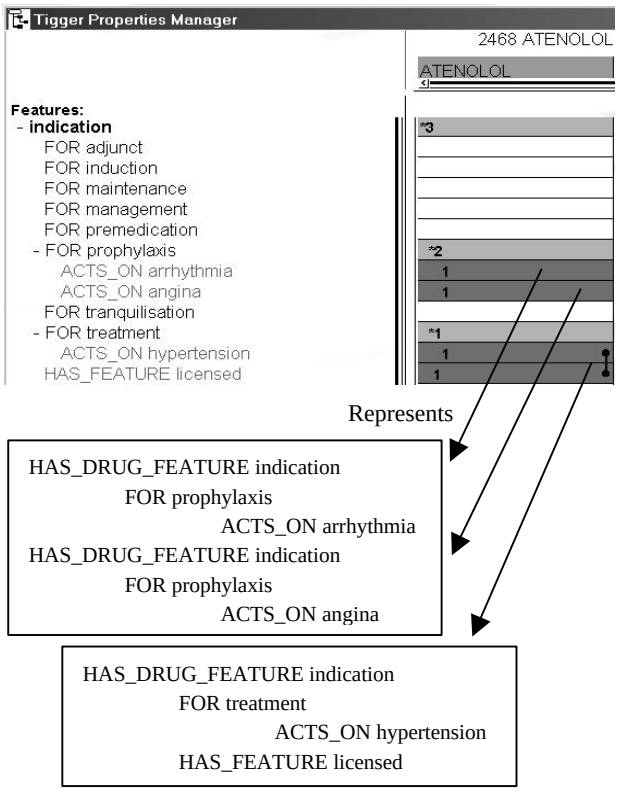


Figure 5. Extract from a DOPAMINE view showing indications of Atenolol with the use of bars.

This method of presentation produces a problem if the structure of the property includes two links attached to the same descriptor. An additional notification is necessary to distinguish multiple semantic links attached to one descriptor in contrast with two separate properties that have a common initial section. This is achieved by the use of bars. These group together semantic links that belong to the same property (see figure 5 for an example). When authoring with the tool it is necessary for the user to indicate whether a bar is necessary.

3 EVALUATION

3.1 Authoring

The Drug Ontology project is producing drug descriptions of primary care relevant generic drugs in the BNF. Initial authoring of the 1500 descriptions is now complete. The 'DOPAMINE' authoring tool was developed after initial authoring experience and was operational mid-way through the project. It has therefore been used to author 700 descriptions. The average size of a description is 21 properties.

The major rate-limiting step in authoring with the tool was creating the left-hand candidate property list. A property can only be added to a description if it appears along the vertical axis of the table. The source text is semi-structured containing a large number of comma-separated lists. A major increase in authoring speed was therefore gained by simply separating candidate terms using punctuation.

BNF source text descriptions of drugs in the same therapeutic class, often share sets of properties such as contraindications. A mechanism therefore had to be added which allowed copying of sets from one description to another. The visual representation of the information provided users with instant confirmation that the process had produced the correct result.

3.2 Checking

Of the 1500 descriptions, 300 have been released to the PRODIGY guideline developers for external appraisal. Before release these descriptions were checked using the DOPAMINE tool. Therapeutic classes of drugs were examined as individual sets. Their properties were compared and checked against the source text.

The tool allowed visualization of information not readily apparent in the textural representations of the descriptions.

Previously, visualizing the classification of drug concepts was the main mechanism for checking the validity of drug descriptions. Users found it easy to detect mis-classifications based on incorrect ontology descriptions. For example, 'Atenolol tablet' clearly should not be a child of 'Atenolol injection'. However, it was much more difficult to detect *missed* classification due to omissions in ontology descriptions. For example, 'Atenolol intravenous injection' should be a child of 'Atenolol injection' but may in fact have been placed in another area of the hierarchy due to an omission in its description.

Using the tabular view, the authors were able to detect omission errors, not detected by checking the classification of drug concepts. These errors showed up as sparse areas in the table for one drug compared with a dense corresponding area for a neighbouring related drug. The omission errors were often due to a failure to copy a set of properties from one drug description to another.

Using the tabular view, the authors were also able to detect multiple use of similar terms in the source text to represent the same property.

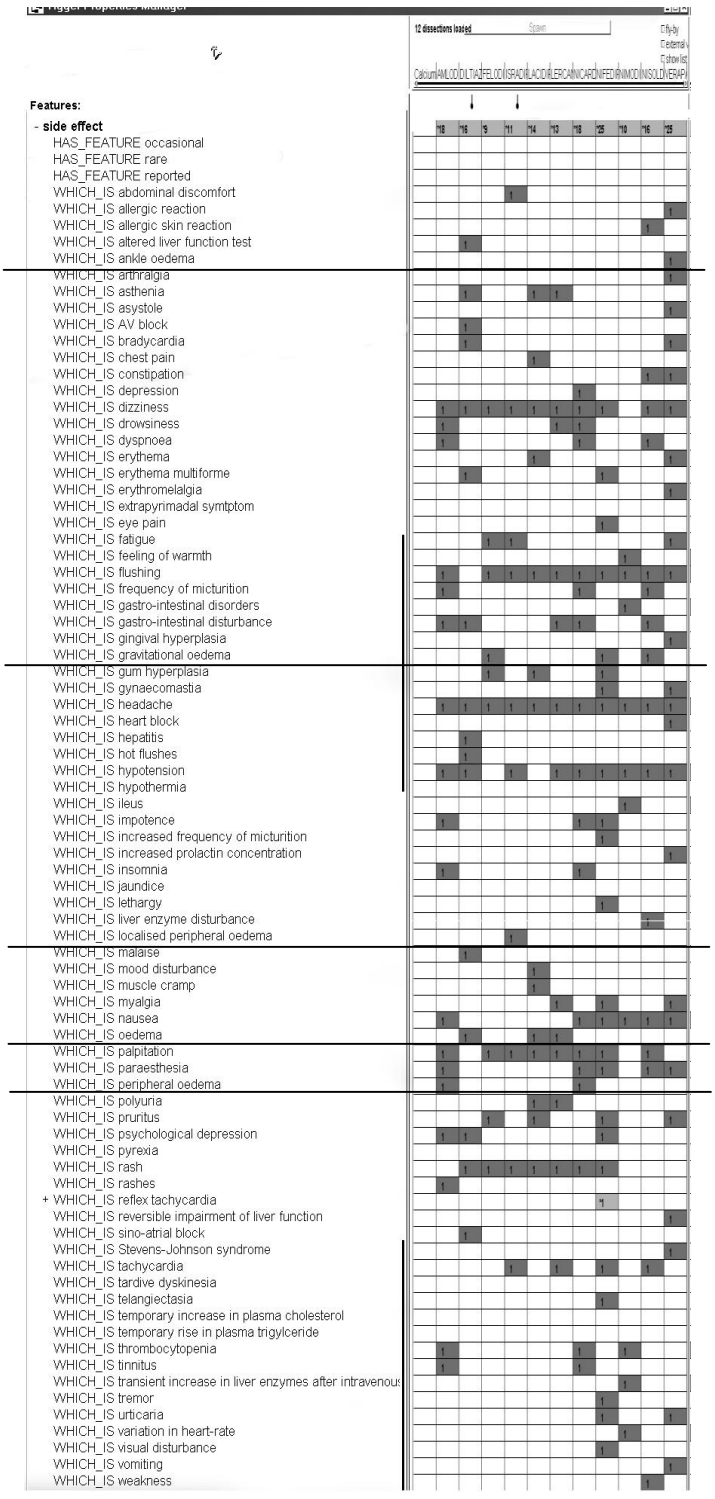


Figure 6. Extract of a DOPAMAINE view showing side effects of calcium channel blockers. Columns show properties of the calcium channel blocker class description, Amlodipine, Diltiazem, Felodipine, Isradipine, Lacidipine, Lercanidipine, Nicardipine, Nifedipine, Nimodipine, Nisoldipine, Verapamil. Horizontal lines have been added to highlight similar terms.

In the calcium channel antagonist section, five related terms have been used in the source text to state that specific calcium channel blockers have a 'peripheral oedema' side effect ('Ankle oedema', 'gravitational oedema', 'localized peripheral oedema', 'peripheral oedema', and 'oedema') (see Figure 6).

Additional organization to group similar terms together is discussed later (see section 4.1), but even with this alphabetical organization of concepts it is possible to identify this possible inconsistency in the drug descriptions. Although inconsistencies of this nature may not be important for professional readers of the source text, automated decision support relies on consistency of terms for correct operation.

As a complete therapeutic class of drugs can be visualized, the table can be used to detect properties which are true for all members. In this case it may be useful to promote this property to the class description and reduce redundancy [16].

Figure 7 shows that all members of the clofibrate class posses breast feeding and pregnancy contraindications. These contraindications could therefore be promoted to the description of the clofibrate drug class.

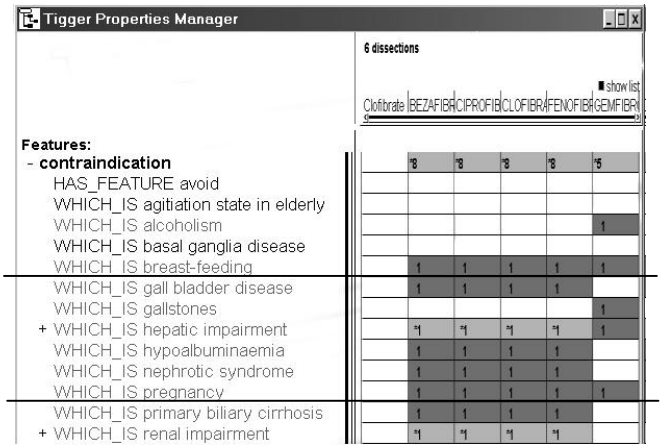


Figure 7. Extract of a DOPAMAINE view showing contraindications of the clofibrate class of drugs. Columns show properties of clofibrate class, Bezafibrate, Ciprofibrate, Clofibrate, Fenofibrate and Gemfibrozil. Horizontal lines have been added to highlight properties true of all class members.

4 DISCUSSION

4.1 General tool design

At present, properties are organized based solely on the information in the structure of the IR drug descriptions. If a drug has a list of 30 side effects that do not differ in structure but only in the terms used, they will only be organized as a flat list (see figure 6). In the initial stages of authoring this is the only information available. However, as the terms in the drug descriptions are mapped to concepts in the Common Reference Model, a more detailed conceptual organisation is possible. Conceptually similar terms will be grouped under common parents. For example the flat list of 30 side effects could be grouped by which body system they involve. Further work is needed to make this organisation possible and explore any difficulties of visualisation raised with a mutliaxial organisation.

The structure of the queries used to group related features is very specific to this application. To allow use in other domains the specification of queries would need to be made more general. This may affect the way properties are displayed and their impact would need to be explored.

4.2 Authoring

Extracting terms form the source text and placing them on the vertical axes of the table was the major factor which determined speed of authoring. While it is not the aim of this project to completely automate the extraction process, the semi-structured nature of the source text could be used to more effectively provide a small candidate list of terms from which the author can choose.

Visualization of properties is most effective when working on one therapeutic class of drugs. Selection of the set is entirely manual at present, which could lead to a drug description being incorrectly missed out of a set. An automatic query or classification-based selection may be more reliable.

4.3 Checking

Errors can arise when the IR descriptions are translated into Grail and classified. As the focus of the project moves away from bulk authoring to bulk validation, it will be necessary for the tool to use the compiled Grail model as a source for visualization. Tools have already been developed to reverse the translation process and produce IR from Grail. It is therefore possible to provide a view based directly on the underlying Grail representation. Future work involves testing this approach.

5 SUMMARY

For automatic classification of drug terms to be successful the formal description of those terms needs to be unambiguous and so verbose. This limits the productivity of authors. We have produced more concise views for authors to interact with and in doing so have provided novel opportunities to visualize clinical drug information. These views have been successfully used to both author the drug descriptions and increase the consistency of the information within those drug descriptions. As the amount of information about each drug grows, even these views are becoming too complex. Further work is needed to make use of the domain information that is available in the related ontologies of the Common Reference Model to further organise the presentation of drug properties.

ACKNOWLEDGEMENTS

The Drug Ontology Project is funded by the NHS Executive. We would like to thank the PRODIGY team at the Sowerby Center For Health Informatics, University of Newcastle and the British National Formulary for their support.

REFERENCES

[1] GKB-Editor Home Page.
URL: http://www.ai.sri.com/~gkb/welcome.shtml

[2] Doug Skuce, Timothy C. Lethbridge. *CODE4: a unified system for managing conceptual knowledge*. International Journal of Human-Computer Studies, Vol. 42, No. 4, Apr 1995, pp. 413-451

[3] W. E. Grosso, H. Eriksson, R. W. Fergerson, J. H. Gennari, S. W. Tu, & M. A. Musen. *Knowledge Modeling at the Millennium (The Design and Evolution of Protege-2000).* 1999. URL:
http://smi.web.stanford.edu/pubs/SMI_Abstracts/SMI-1999-0801.html

[4] Protégé 2000 Tutorial.
URL: http://smi-web.stanford.edu/projects/protege/protege-2000/doc/tutorial/index.html

[5] Gary Ngg. *Interactive Visualisation Techniques For Ontology Development.* PhD Thesis submitted for examination at the University of Manchester UK, May 2000.

[6] I Purves. *Prodigy interim report.* Journal of Informatics in Primary Care 1996:2-8.

[7] A.L. Rector, P.E. Zanstra, W.D. Solomon, J. E. Rogers et el. (1998). *Reconciling Users' Needs and Formal Requirements: Issues in developing a Re-Usable Ontology for Medicine.* IEEE Transactions on Information Technology in BioMedicine, Special issue on the EU Healthcare telematics programme; Vol. 2 No.4 pp. 229-242

[8] N. Guarino., P. Giaretta. *Ontologies and Knowledge Bases: Towards a Terminological Clarification.* In N. J. I. Mars (ed.), Towards Very Large Knowledge Bases, IOS Press 1995.

[9] Mike Uschold, Martin King. *Towards a Methodology for Building Ontologies.* Technical Report AIAI-TR-183, Artificial Intelligence Applications Institute, Edinburgh University,UK, 1995.

[10] A.L. Rector, S.K. Bechhofer, C.A. Goble, I. Horrocks, W.A.Nowlan, W.D. Solomon. *The GRAIL Concept Modelling Language for Medical Terminology.* Artificial Intelligence in Medicine, Volume 9, 1997.

[11] Open Galen Common Reference Model
URL: http://www.openGalen.org

[12] W.D Solomon, C.J. Wroe, J.E Rogers, and A.L. Rector. *A reference terminology for drugs.* Journal of the American Medical Informatics Association 1999(1999 Special Conference Issue).

[13] British National Formulary. British Medical Association and the Royal Pharmaceutical Society of Great Britain, 2000.

[14] J.E. Rogers, W.D. Solomon, A.L. Rector, P.E. Zanstra. *From rubrics to dissections to GRAIL to classifications.* Medical Informatics Europe (MIE-97). Thesalonika, Greece: IOS Press, 1997:241-245.

[15] W.D. Solomon,, J Rogers, A. Rector, E J van der Haring, P.E. Zanstra. *Supporting the use of the GALEN Inermediate Representation.* Journal of the American Medical Informatics Association 1998(1998 Special Conference Issue).

[16] C.J. Wroe, W.D. Solomon, J.E. Rogers, A.L. Rector. *Inheritance of Drug Information.* Submitted to Journal of the American Medical Informatics Association 2000 (2000 Special Conference Issue).

Methods for Clustering Mass Spectrometry Data in Drug Development

Huiru Zheng[1], Sarabjot Singh Anand[1], John G Hughes[1] and Norman D Black[1]

Abstract. Isolation and purification of the active principle within natural compounds plays an important role in drug development. MS (mass spectrometry) is used as a detector in HPLC (high performance liquid chromatography) systems to aid the determination of novel compound structures. Clustering techniques provide useful tools for intelligent data analysis within this context. In this paper, we analyse some representative clustering algorithms, describe the complexities of the mass spectrometry data generated by HPLC-MS systems, and provide a new algorithm for clustering, based on the needs of drug development. This new algorithm is based on the definition of a dynamic window within the instance space.

1 INTRODUCTION

Chemists make literally thousands of analogues in order to develop drugs from natural sources [4]. Figure 1 shows a simplistic view of the drug development process.

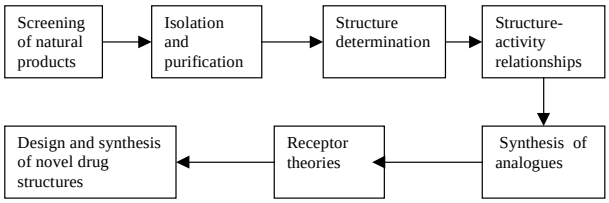


Figure 1. A Simplistic view of the drug development

Screening natural compounds for biological activity continues today in the never-ending quest to find new compounds for discovering drugs. Isolating and purifying the new compounds for the active function plays an important role in the pattern. Next, the structures of the new compounds have to be determined before the structure-activity relationships (SARs) can be analysed. Although the new compounds have useful biological active function, chemists have to use synthetic analogues to reduce any serious side effects, and finally, develop the novel drugs.

Our project is targeted at the isolation, structure elucidation and synthesis of biologically active substances of potential pharmacological significance from the venom of frogs. Several isolation techniques have been developed such as freeze-drying, filtration, centrifugation, and, in particular chromatography have been developed. MS (mass spectrometry), a technique used to characterize and separate ions by virtue of their mass/charge (*m/z*) ratios [7], can be helpful in structure determination as the fragmentation can give useful clues about the structure.

In our experiment, we use a HPLC (high performance liquid chromatography) system to isolate the natural product. The chromatographic separations are based upon the distribution of components of a mixture between a liquid phase and a stationary phase. The MS (mass spectrometry) is used as the detector. We use a number of different samples of the frog's venom that result in millions of two-dimensional records describing the mass and the separation time of isolated molecules. Clustering is applied to discover the functional molecules for determining structures of the pharmacological compounds.

The aim of this paper is to ascertain whether current clustering algorithms are suitable for the data generated by HPLC-MS. In section 2 we describe the features of the mass spectrometry data generated by our experiments in drug discovery. In section 3 we give an analysis of current algorithms. In section 4, we describe a new algorithm for clustering on the needs of drug development, which is based on the dynamic window concept. In section 5, we will provide the result and discuss it for further work.

2 FEATURES OF THE MASS SPECTRO-METRY DATA

The venom of frog is collected and injected into the HPLC. The entire isolation time is about 80 minutes. Every 0.03 or 0.04 minutes during these 80 minutes, the system will perform a scan to isolate the molecules from the venom analyte. The mass spectrometry detects the isolated molecules. However, it takes an isolated molecule about 1 minute to travel from one end of the column in the HPLC system to the other end. One molecule could be detected in about 33 (if the scan time interval is 0.03 minute) or 25 (if the scan time interval is 0.04 minute) scans. Therefore, data obtained from the HPLC-MS contains a lot of noise, which needs to be eliminated. For each frog, the amount of the original data is large; for example, frog "Aur" has about 446,900 records. Therefore, clustering technique can be applied to smooth the noise data and get the real data, i.e. identify the true values of mass/charge ratios and isolation time of the molecules.

We aim to search for the same functional molecules between the pharmacological frogs. For each frog, even after eliminating noise, the amount of data is still large, for examples, frog "Aur" contains 24,549 points and frog "Inf" has 50,772. There are different

[1] Faculty of Informatics, University of Ulster at Jordanstown, Newtownabbey, Co. Antrim,N. Ireland
Email:{H.Zheng,ss.anand,jg.hughes,nd.black}@ulst.ac.uk

species of frogs, each species has different types, and for each type, there are hundreds of different frogs. Therefore, in order to mine for the functional molecules, clustering techniques are used.

Fig 2 shows a sample of data generated by HPLC-MS after eliminating noise. Mass spectrometry data from HLPC-MS has two dimensions: *Time* and *Mass*. *Time* describes the isolated time of the molecule and *Mass* represents the mass/charge ratios. *No.* is the index for each record and *Frog* is the name of the frog from which the data of the venom is collected.

No	*Time*	*Mass*	*Frog*
1	3.51	1114.55	Aur
2	3.64	609.0467	Aur
3	44.75	141.46	Aur
4	16.9382	176.934	Inf
5	37.8	234.14	Caer
6	15.52	617.963	Aur
7	15.1618	617.986	Caer
8	15.43	618.004	Inf
9	60.252	1998.44	Inf
10	41.68	1998.9	Inf

Figure 2. Mass Spectrometry Data

Mass spectrometry data has some unusual features. Only molecules which have similar *Time* and have similar *Mass* will be the same, and therefore in the same cluster. As shown in Figure 2, though molecule No.1 and molecule No.2 have similar *Time*, since their *Mass* values are quite different, they are not the same molecule, and cannot appear in the same cluster. For molecule No. 9 and No. 10, their *Mass* values are similar but their isolated *Time* values are quite different. Therefore they should not be in the same cluster. Molecules No. 6, No. 7 and No. 8 are similar in both *Time* and *Mass* and therefore they are deemed to be the same molecule even though they have been isolated from different frogs.

Another feature of the data is that the ranges of *Time* and *Mass* are large, and the intervals of *Time* and *Mass* between the data of two molecules are small. There is some regularity in the data and it is not necessary to consider the entire data space.

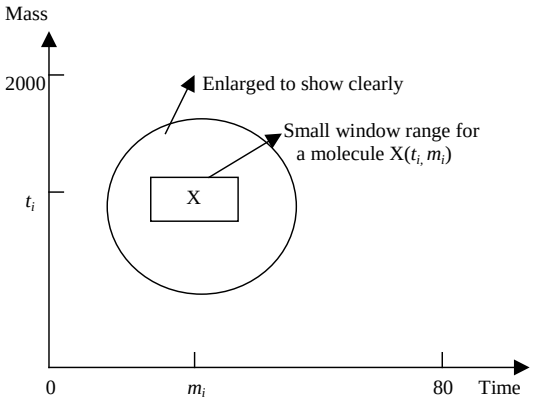


Figure 3. Data space of mass spectrometry data generated by HPLC-MS

Figure 3 shows an example of the data within the instance space. X (t_i ,m_i) is a molecule in the data space. Comparing to the entire data space, the range of the window is very small need to be enlarged. All of the data points within the window represent the same molecule X (t_i ,m_i), and all of the data points representing the same molecule X (t_i ,m_i) will fall in the widow. Therefore it is unnecessary to search the entire space for the same molecule.

All these features give us a heuristic to select our algorithm for clustering of mass spectrometry data.

3 CLUSTRING TECHNIQUE FOR INTELLIGENT DATA ANALYSIS

The various clustering concepts available can be grouped into two broad categories: hierarchical methods and nonhierarchical (or partitioning) methods [6].

Nonhierarchical (or partitioning) methods include those techniques in which a desired number of clusters are assumed at the starting point. Data points are reallocated among clusters so that a particular clustering criterion is optimized. A possible criterion is the minimization of the variability within clusters, as measured by the sum of the variance of each parameter that characterizes a point. Given a set of objects and a clustering criterion, nonhierarchical clustering obtains a partition of the objects into clusters such that the objects in a cluster are more similar to each other than to objects in different clusters. The basic idea of these algorithms is demonstrated by K-means and *k*-medoid methods [5]. For K-means, the centre of gravity of the cluster represents each cluster; for K-medoid, each cluster is represented by one of the objects of the cluster located near the center. A well known algorithm is CLARANS (Clustering Large Applications based on RANdomized Search) which uses a randomized and bounded strategy to improve the performance [8], and can be viewed as an extension of these methods for large databases.

The K-means method constructs a partition of a database of N objects into a set of K clusters. It requires as input the number of clusters, and starts with an initial partition, uses a minimizing of the distance of each point from the cluster certre as the search criterion and an iterative control strategy to optimize an objective function.

The *k*-medoid method is used by PAM (Partitioning Around Medoids) [5] to identify clusters. PAM selects *K* items arbitrarily as medoids and swaps with other items until *K* items qualify as medoids. Then an item is compared with the entire data space to obtain a medoid. It requires as input a value for input *K*. CLARNS is an efficient improvement of this *k*-medoid method. However, since it uses randomized search, it cannot be guaranteed to converge when the amount of data is large.

Andrew Moore [3] and Dan Pelleg [1] use the *kd*-tree data structure to reduce the large number of nearest-neighbor queries issued by the traditional K-means algorithm, and use the EM (Expectation Maximization) method for finding mixture models. They use a hyper-rectangle *h* as an additional parameter to determine the new centroids. The initial value of *h* is the hyper-rectangle with all of the input points in it. It updates its counters using the centre of mass and number of points that are stored in the *kd*-node corresponding to *h*. Otherwise it splits *h* by recursively calling itself with the children.

Hierarchical methods include those techniques where the input data are not partitioned into the desired number of classes in a single step. Instead, a series of successive fusions of data are performed until the final number of clusters is obtained. A hierarchical clustering is a nested sequence of partitions. Agglomerative hierarchical clustering starts by placing each object

in its own cluster and then merges these atomic clusters into larger and larger clusters until all objects are in a single cluster. Data partitioning based hierarchical clustering starts the process with all objects in a cluster and subdividing into smaller pieces. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) uses data partitioning according to the expected cluster structure called CF-tree (Cluster Feature Tree) which is a balanced tree for storing the clustering features [9]. STING (Statistical information grid-based method) is based on a quad-tree-like structure, and DBSCAN [2] relies on a density-based notion of cluster and uses an R*-tree to achieve better performance.

Hierarchical algorithms do not need *K* as an input parameter. This is an obvious advantage over the nonhierarchical algorithms, though they require a termination condition to be specified.

BIRCH uses a CF-tree for incrementally and dynamically clustering the incoming data points to produce a condensed representation of the data, and applies a separate cluster algorithm to the leaves of the CF-tree. It uses several heuristics to find the clusters and to distinguish the clusters from noise. It is one of the most efficient algorithms because it condensed data. However, BIRCH is sensitive to the order in which the data is input and so different cluster may result due to simply a change in ordering the data.

DBSCAN [2] defines clusters as density-connected sets. For each point, the neighborhood of a given radius has to contain a minimum number of points - the density in the neighborhood has to exceed some threshold. DBSCAN can separate the noise and discover clusters of arbitrary shape. STING [8] uses a quad-tree-like structure for condensing the data into grid cells. Hierarchical grid clustering algorithms organize the data, sort the block the blocks by their density, and then scan the blocks iteratively and merge blocks. The order of the merges forms a hierarchy. It is crucial to determine a proper criterion to merge grids and to terminate the clustering.

We propose DYWIN (DYnamic WINdow- based) clustering algorithm with full consideration of the complexities of the mass Spectrometry data to overcome the above disadvantages.

4 DYNAMIC WINDOW–BASED CLUSTRING

In drug development, it is difficult to predict the number of biologically active compounds. Also, it is difficult for us to predict the number of functional molecules in the samples. In other words, it is difficult to input the value of *K* for nonhierarchical clustering.

For density-based or grid-based hierarchical methods, we divide the entire data space into grids, use the density of each grid as a criterion to merge clusters, and obtain the dominant clusters. According to the features of our data, the points in a single cluster are within a range *W* [Δ*t*, Δ*m*] of the two dimensions of *Time* and *Mass*. When two grids are merged, the size of the new cluster should be within the range *W* also. Therefore, it is not easy to select the size of the grids. If the size of the grid is too small, performance of the algorithm will deteriorate, if too large, the merging process will fail.

Furthermore, in our case, we need not search the whole data space. Rather, we can use the features of the data outlined in section 2 to assist in clustering.

The algorithm presented here is based on work by Andrew Moore [3] and Dan Pelleg [1] on grid-based algorithms, but we use dynamic windows defined on the instance space instead of hyper-rectangles. We use the density of the window, but do not split the entire space into grids. The positions of windows are not fixed, so we call it *dynamic window-based clustering* (DYWIN).

Assume the number of frogs is *T*, input data $P_i = (t_i, m_i)$, *W* is a window with width Δt in the time dimension and height Δm in the mass dimension. D_i is the density of P_i in each W_i corresponding to cluster C_i. Figure 4 shows a simplified data space containing from three different frogs.

The algorithm DYWIN developed contains two main steps: to eliminate noise in the data from same frogs to get the real molecule data and to search the common functional molecules between different frogs that have the similar pharmacological significance.

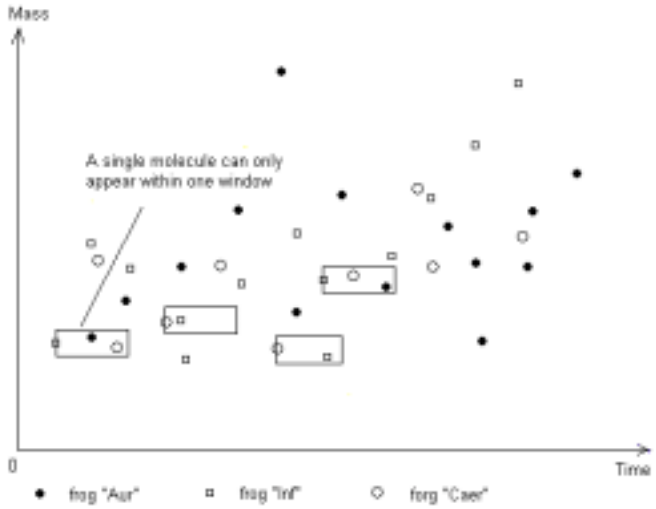


Figure 4. Clustering based on dynamic window

Step-1 Remove noise in data of same frog, the algorithm is shown below:

```
Cluster1( )
{
  Input Δt, Δm
  Sort Data X = {P1, P2, ..., Pn}
  Candidates = X
  While Candidates exist
  {
    pick next Candidates Pi = (ti, mi)
    define Window, Wj by the 4 vertices
      { (ti, mi + Δm/2), (ti + Δt, mi + Δm/2),
      (ti, mi − Δm/2), (ti + Δt, mi − Δm/2) }
    k = 1, Cj = 0
    while i + k is contained in Wj
    {
      k++
      Candidates = Candidates – { Pi+k }
```

$C_j = C_j \cup \{P_{i+k}\}$

}

}

}

Here, $\Delta t, \Delta m$ are empirical values provided by chemists.

After data of each frog is clustered to remove the noise, we can move to the next step.

Step-2 Search the function molecules between different frogs

Here, the input data is from different frogs within which the noise has been eliminated. A density threshold Td is an input value depends on what we expect to find in these different frogs. When we are searching for the common molecules that appear in all the frogs with same functions, Td is set to the number of the types of the frogs. If we want to query about the molecules which might play a special function appears in two frogs of all the frogs clustered by us, then is equal to 2. If the function is only appeared in one frog, the value of Td is set to 1.

```
Cluster2( )
{
  Input Td
  Input data {P1, P2, ..., Pn} collected from
                                different frogs
  Cluster1( )
  Calculate density Di for each Cluster Ci
  If Di ≥ Td then
    keep Ci
    output Ci
  end if
}
```

5 RESULT AND DISCUSSION

We have applied this algorithm to real mass spectrometry data generated from HPLC-MS and have successfully identified the molecule components in the different frogs.

To date, we have got data from 12 types of frogs. Three of them, “Aur”, “Inf” and “Caer” are different type of frogs from the same species, and have some of the same active functions.

After clustering the data to remove the noise in the data, “Aur” contains about 24500 molecules, “Inf” has about 50700 molecules and “Caer” contains about 40200 molecules. When using a value of 3 for Td , DYWIN output 42 clusters, in other words, 42 common molecules are distinguished in these three frogs. One of them is the molecule that we use to isolate the venom sample. It is definitely right to appear in the result. For Td = 2, we find 4192 clusters. Since the name of the frog is recorded during clustering, the result clusters show that there are about 1130 common molecules between “Inf” and “Caer” but not in “Aur”.

Figure 5 give a simple explanation to the relations of the frogs, functions and molecules.

Frogs	Functions	Molecules
Aur	F1,F3,F4,F5	ABCDEFG
Inf	F1,F2,F5,F6	B D HIS
Caer	F1,F2,F4,F7	AB E H PQ

Figure 5. Relations of frogs, functions and molecules

When Td = 3, we get the molecule “B” which exists in all of the three frogs, then, we can infer that molecule “B” play an important role in the function “F1” which is the common function in these three frogs. When Td = 2, we also get the molecules “D”, “E”, and “H”. According to the frog name from which we collect the data, we can also inform that “D” is important to function “F5” for frog “Aur” and “Inf”, “E” may affect “F4” which appears in frog “Aur” and “Caer”, and so on. While it is easy to manually decipher the SAR from the data in Fig 5, this is generally difficult to achieve in real data, as the number of molecules is very large. Thus we will, in the future, employ a classification technique for this purpose.

This work is an initial step in the complex process of drug development. We will get more data from the different types in same species with similar function to isolate and identify the functional molecules. Further more, we can establish the database from the results to distinguish the species and the types of different individual frogs by the functional molecule clusters.

The other point we are going to discuss is that, though the benefits of applying data specific clustering techniques are obvious, we still need to do further work on it.

From the experiments, we find the selection of the empirical value of $\Delta t, \Delta m$ may affect the result. We find a case show as Figure 6.

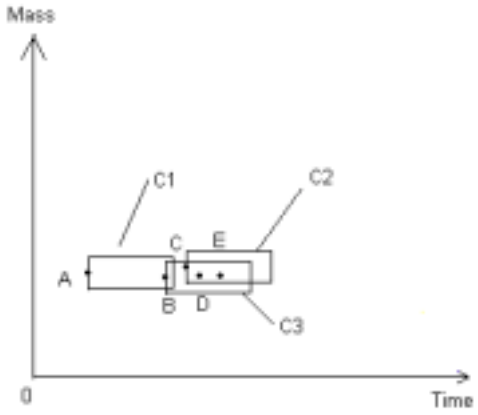


Figure 6. A case need to be discussed

Dots A, B, C, D and E represent five points in the data space. According to DYWIN, molecule A and B are in the same cluster C1, and C, D and E are clustered to C2, though, it seems more reasonable to cluster B, C, D, and E into the same cluster C3. The problem is caused by the size of the window which is determined by $\Delta t, \Delta m$. We are considering applying adaptive values of $\Delta t, \Delta m$ to determine the size of the windows. However this adjustment needs the supports from the chemists.

ACKNOWLEDGEMENT

We would like to thank Professor Chris Shaw and Dr. Stephen McClean for providing the experimental data and giving helpful suggestions for analyzing the mass spectrometry data.

REFERENCE

[1] Dan Pelleg, Andrew Moore (99) Accelerating exact *k*-means algorithms with geometric reasoning, *KDD'99*, pp 277- 281

[2] Ester M, Kriegel H, Sander J, Xu X(96) A density-Based Algorithm for discovering clusters in large spatial databases with noise, *Proceedings of 2nd international conference on KDD*

[3] Andrew Moore (98) Very fast EM-based mixture model clustering using multi-resolution kd – tree, *Neural information processing system conference*,1998

[4] Graham L. Patrick(1997), *An introduction to medical chemistry*, Oxford, pp 82-89

[5] Kaufman L, Rousseeuw PJ (1990) *Finding Groups in data: an Introduction to Cluster Analysis.* John Wiley & Sons, Chichester

[6] Rakesh Agrawal, Johannes Ges Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan (1998) Automatic subspace clustering of high dimensional data for data mining applications, *Proc. of the ACM SIGMOD Int'l Conference on Management of Data*, Seattle, Washington, June 1998

[7] Stephen Mcclean(1999), PhD Thesis: *An Investigation of Modern Analytical Techniques for the Identification and Determination of Selected Drugs and Pollutants, their egradation Products and Metabolites,* University of Ulster, U.K

[8] Wang W, Yang J. Muntz R (1997) STING: A Statistical Information Grid Approach to Spatial Data Ming, *Proceedings of the 23rd VLDB conference,* Athens, Greece, pp 186-195

[9] Zhang T, Ramakrishnan R. Livny M (1996) BIRCH: An Efficient Data Clustering Method for Very Large Database. In *Proceedings of the 1996 ACM SIGMOD International Conference on Mangement of Data,* Montreal, Canada, pp 103 – 114

[10] Gholamhosein S,Surojit C and Aidong Z(2000) WaveCluster: a wavelet-based clustering approach for spatial data in very large databases, *VLDB Journal*(2000) 8:289-304

Mining a database of Fungi for Pharmacological Use via Minimum Message Length Encoding

Robert Zimmer[1] and **Andrew Barraclough**[2]

Abstract. This paper concerns the use of fungi in pharmaceutical design. More specifically, this research involves mining a database of fungi to determine which ones have waste products that are unusual in their spectral fingerprints, and therefore worth being tested for medicinal properties. The technique described in this paper involves Minimum Message Length encoding. Minimum Message Length (sometimes called Minimum Description Length) encoding is a method of choosing a binary coding for a set of data. The method's goal is to use the frequency of occurrence of each data point to ensure that frequently occurring data are given short codes. Minimum Message Length encoding provides a solution that is optimal in the sense that if the entire data set is employed in the encoding, then the code generated will have the property that no other unambiguous prefix code will provide a shorter encoded version of the entire set. In this paper, the process is turned on its head.

The problem that is addressed is: given a large database, how can we pick out the elements that are quite different from the others. The first step in our solution involves using the Minimum Message Length algorithm to generate a compact code for all, or a representative learning section, of the data. The data that require long descriptions in this code are likely to be the ones that possess unusual features. In this paper, we describe this process in some detail, and explain the application of it to a database of fungi.

1 Introduction to the Problem and Technique

The pharmaceutical industry uses fungus excretion to produce drugs. As this process has been used for quite a long time, a large number of chemicals have already been produced and tested. The problem addressed by the techniques described in this paper is the following: given a large database of unexplored fungi how can one decide which fungi should be chosen to explore next. The first step in the answer to this question is to choose a fungus the spectral sequence of which seems to be as different from the rest as possible. One way to do this is by using clustering techniques to group the data and then to choose a fungus that falls out of any of the groups. In this paper we take a different tack. We use a coding technique that is designed to make typical elements have short encodings. Then the fungi with long encodings are likely to be the least typical and, therefore, ripe for further study. Given a large database, the encoding can be generated from a small sample and then all the fungi (and new ones that come along) can be encoded to see how typical they are.

In the next section the coding technique that underlies the system is explained

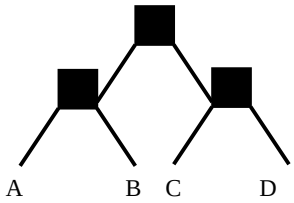
2 Minimum Message Length (MML) Encoding

The Minimum Message Length encoding scheme begins by applying the Huffman coding method to a set of characters. Consider, for example, the question of encoding, as a string of 0's and 1's, the following message:

AABABCAADABAABC

Without taking account of the frequency of the characters, the best thing is to treat the distribution as uniform, and code each of the four letters as a two-bit string. For example: A → 00, B → 01, C →10, D → 11.

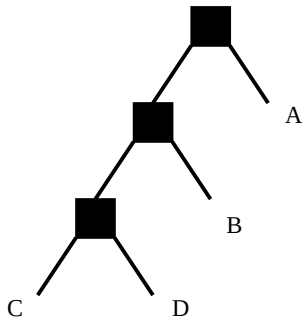
And the decoding tree for the coding is the following balanced tree:



With the convention that reading "0" means go left. And the this encoded message is the 30-bit long word:

000001000110000011000100000110

This is one of the extrema that could be employed in the coding. The other extremum is to make a near linear tree as follows:



[1] Department of Computer Science, The Open University, Milton Keynes MK6, UK, email: R.M.Zimmer@open.ac.uk

[2] Information Systems Group, UNISYS UK, The Octagon, Slough, UK.

The coding then is A → 1, B → 01, C → 001, and D → 0001. And the message is encoded as:

11011010011100011011101001

which has only 26 bits. This example was designed to make the linear code effective. Indeed, with four characters there is nothing between the balanced and linear trees. In general the optimal solution is between the two extremes. In the next section the technique, Huffman Encoding, of generating the trees is described.

2.1 Huffman Encoding

The principal of Huffman encoding can be summarized as follows:

(i) Work out the frequency of each data value, Pi

$$Pi = Ni / N$$

Where:

Pi is the probability of a particular data value

Ni is the number if instances of that particular data value in the data

N is the total number of items in the data

(2) Take the two least frequently occurring letters, treat them as one term whose frequency is the sum of the two individual frequencies. The two letters are then thought of as a node of a binary tree with two subtrees, each of which is a leaf.

(3) Perform (2) iteratively until you reach a single tree.

Consider the following example. We wish to code the string ABRACADABRARCRC with the shortest possible binary string.

The first step is to count the frequency of each character:

A occurs 5 times.
B occurs 2 times.
R occurs 4 times.
C occurs 3 times.
D occurs 1 time.

The two least frequent letters are B and D. They are joined together and jointly appear 3 times.

A occurs 5 times.
[B,D] occurs 3 times.
R occurs 4 times.
C occurs 3 times.

The pair now ties with C as the least frequent set.

The three are then treated as one data point. Leading to the new frequency table:

A occurs 5 times.
[[B, D], C] occurs 6 times.
R occurs 4 times.

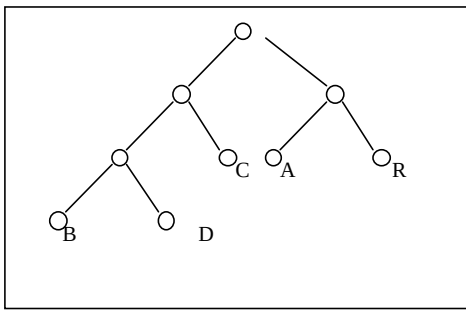
Now A and R are the two least frequent items so get so get paired as:

[[B, D], C] occurs 6 times.
[A, R] occurs 9 times.

The whole thing is then the structure:

[[[B, D], C], [A, R]].

This structure gets turned into a binary tree such that every pairing represents a node with the two elements of the pair as subtrees. The tree generated by this process is given below:



This process clearly has the effect of making the least frequently occurring letters have the longest paths. This is optimal in the sense that:

> "The length of the encoded message is equal to the weighted external path length of the Huffman frequency tree…. "No tree with the same frequencies in external nodes has lower weighted external path length than the Huffman tree"
> (SEDGEWICK. (1988) page 326-7).

That this is the minimal message length of all possible message lengths can be proven by induction. The length in bits of each codeword (i.e. the binary encoding for a specific data value) is:

$$L_i = - \log_2 (P_i)$$

Where L_i is the codeword length required for datatype i and P_i represents the probability of that datatype occurring.

In the next section we show how to apply this technique to a database of fungi.

3 Mining Fungi Data

The problem under consideration is that of analysing data representing the chemical structure of different types of fungi. A number of such fungi have been collected and each

one broken down into 5000 attributes representing its chemical structure. Each one of these attributes is represented as a numerical value.

The goal of the data mining exercise is to identify those fungi that contain an unusual chemical structure. This entails identifying patterns among the 5000 attributes common amongst the fungi and those fungi that stand out as having unusual patterns. This presents an unusual knowledge discovery task in so far as the goal is clearly defined and the data is presented in a consistent numerical format allowing the data mining stage of the process to be immediately embarked upon. However, as will be shown this does not mean that data processing or modelling will not be required.

4 Using MML to Find Outliers

The overview of the technique is:

1. Take a Sample set of Fungi, and use them to generate the Huffman Code
2. Encode the rest of the database using the Huffman Code
3. Choose the fungi with long descriptions as possible outliers, and test these

4.1 Implementation

The fungi data is represented as 5000 numerical attributes for each individual fungus.

	Attributes										
	1										5000
F	*F1*	2	5	33	11	88	12	44	6	122	11.....
U	*F2*	3	6	12	77	2	9	21	5	44	2.....
N	⋮										
G	⋮										
I	*F n*	.....									

Each attribute from 1 to 5000 has its own set of values consisting of one value per fungus. Using MML encoding a Huffman tree for each of the 5000 attributes based upon the relative frequencies of values of that particular attribute was constructed.

Having created an encoding tree for each of the attributes, these trees can then be used to encode each individual fungus as a binary string. This means that each binary value encoding will be shortest where that particular value is frequent. The fungi with the overall shortest binary encoding will therefore have the overall most 'common' attribute values. Those fungi with the longest overall binary encoding will represent the fungi with the 'least typical' values. These are the unusual fungi that are required to be identified.

4.2 Possible Pitfalls

In order for this method to be effective the range of values for each attribute was considered individually. It turns out that for some attributes, no particular value occurs terribly frequently. For these attributes the values were binned: so that values within a certain region were treated as identical. Care must be taken not to over emphasize any particular attributes relative to others. Statistical techniques have been used to analyse the attributes and their range and distribution of values.

Given the fact that any processing performed upon the attributes must retain the relative frequency distribution amongst attributes it is probable that some attributes will be significant in the classification process whilst others may well show little or no grouping of values. Such attributes will be identifiable as their encoding trees will be much deeper than those where a high frequency of values leads to fewer values needing to be encoded.

Once identified these attributes need to be excluded from the classification process as they may distort the final results should they be included. This means that once the attribute encoding trees have been constructed their depths can be compared and a maximum depth determined above which attributes are not included in the fungi encoding process.

4.3 *Encoding the data*

Once the encoding trees have been built for each of the attributes the original data can be encoded using these trees. A simple encoding can be performed where all of the attributes are included in the encoding process. However, some of these attributes do not show any particular grouping of values and as such will not be relevant to the encoding process. Furthermore, if these attributes were included in the encoding process the final results would be distorted. Therefore the relevance of attributes to the classification of cases is necessary to provide a refined encoding where irrelevant attributes are excluded. We've done this by calculating the standard deviation of frequency values for each of the attributes. Those attributes with a high standard deviation will have a higher grouping together of values than those with a lower standard deviation. A threshold is used to determine which attributes are not included in the encoding process. Two alternative thresholds were tried—one based on the average and the other on the median of the standard deviations of the attributes. The efficacy of the two methods was then compared.

5 Evaluation

Evaluation showed that the binning function was very susceptible to producing misleading results dependent upon the shape of the data. Where the range and shape of data fit the number of bins set in the program attributes are grouped together correctly. However, where this is not the case attributes can be separated where they logically belong to the same grouping and vice versa. The grouping can be particularly affected by extreme values as shown in Test 9 (e.g. one very high value leading to a very high bin size) which will cause values to be grouped falsely together. This suggests some method of determining optimal bin sizes for each attribute is necessary for this part of the program to

function efficiently. The tests showed the limitations of a program based upon a fixed number of bins.

Classification depends upon the encoding length given to values by method traversal. This method was checked to show that the encoding lengths it assigned to values were correctly based upon the values position in the relevant coding tree. This position in turn being based upon the frequency assigned the attribute value. This was verified to be the case. The actual encoding values assigned however can be the same for values of different frequencies where they occur at similar depths within the tree. This means those values of similar but not exactly the same frequencies carry the same encoding weight. Whilst a correct implementation of Huffman encoding it must be noted that the encoding process will differentiate most between large variations in frequency.

5.1 System Testing

The system has been tested on various example databases, each of which is contrived to have one or more interesting or potentially misleading attribute. In this section a discussion of a few of the test cases is given.

Test Sys1
The first data set consisted of 15 cases and 5 attributes. Each of the attributes had clearly defined values. The purpose of this test was to check the program could identify those cases (rows) with the most common values and assign identical encoding to identical rows. To this end one row was replicated five times, one row replicated four times, one row replicated three times, one row replicated two times and one row left unduplicated. Thus, five groups were created, no two groups shared common values and the values were such that values were allocated separate bins. This enabled expected classification of the rows to be easily predicted and the system performed as expected giving identical rows identical encoding values. The most frequent rows (five occurrences) received the lowest encoding and the single row together with the row with two occurrences received the highest encoding. It was noted that no differentiation occurred between the single occurrence and the double occurrence. This was expected as their values shared the bottom level of the encoding tree.

Also as expected refined encoding produced the same results as all the attributes were equally significant in the encoding process.

Test Sys2
The second data set used was a refined version of the data used in Sys1. The same format was applied but in this case attribute values in the identical rows were changed slightly. These changes were made so that each value would still fall within the same bin.

For example, the first row in the table was changed in the following way.

1.0	1.0	1.0	1.0	15.0	15.0	25.0	25.0
1.1	1.2	1.3	1.4	15.1	15.2	25.1	25.2

25.0	35.0	40.0	40.0	40.0	40.0	40.0	40.0
25.3	30.5	40.1	40.2	40.3	40.4	40.5	40.6

Changed values are still grouped to the same bins.

In this way the program was run again testing that the values were grouped together correctly and the same encoding trees built as in the previous test. The frequency counting and tree building in both examples should produce the same encoding results and this was proven to be the case.

Again, as expected refined encoding produced the same results as all the attributes were equally significant in the encoding process This test does not assess the validity of the actual binning process in grouping data values but shows that the tree building and encoding process functions correctly with the given binned values.

Test Sys3
In this test the table used in Sys2 was further modified by incrementally changing the attribute values of case 9, which originally had no common attribute values. It was given first one common value and then progressively more common values until it became a member of the most common class. The addition of one common value immediately raised it above cases 4 and 5 the two rows with least common attributes. As more common values were added the encoding value was reduced accordingly. The test showed the more common attributes a case has the lower its encoding value and that this is affected by changes in individual attributes.

Test Sys4
In this test the effect of bin size combined with range of data was examined to see the effect on classification. The two attributes from the fungi data having maximum and minimum range were taken and the binning process applied to them. Where the range was small groupings were best detected with a smaller number of bins – grouping almost disappearing when 100 bins were used. However where the range was large 10 bins grouped together many attributes of widely differing values, the grouping becoming better defined with 50 bins. When encoding was applied to 39 cases using these two attributes the encoding results were different but a general trend of classification remained. In the case of only 10 bins the number of values in the encoding trees was smaller reducing the difference in encoding values of frequent and infrequent values.

6 RESULTS OF CLASSIFICATION OF FUNGI DATA

Before classification was attempted some exploratory analysis of the data was performed to try to ascertain the ranges of the 9960 different attributes. It was found that these varied immensely. For example, one attribute has a range of 99948.36, while another has a range of 15.16. The average range of the attributes was 1348.8712921686774

The number of cases used in the demonstration was 39. Tests were first run with different numbers of bins (10, 50, 100) on these two extreme attributes and the results of binning compared. The tests with 10 bins showed that although the attribute with the minimum range (attribute 36) was grouped appropriately, with attribute 1832 large groupings of values occurred often-encompassing widely differing values. As the number of bins was increased less values occurred in each bin but by 100 bins grouping had almost disappeared for attribute 36 and was severely reduced for attribute 1832. A choice of 50 bins seems to provide the best compromise. In general, the less grouping is detected in

those attributes with a smaller range emphasising the significance of attributes with a larger range in the classification process. However, a small number of bins leads to false grouping of attributes of greater range.

The tests were then run on the full evaluation set again using 10, 50 and 100 bins. The overall classification results proved to be consistent whatever the bin size chosen the eight most unusual fungi identified being the same in each case indicating that the classification of cases is similar whichever attributes are emphasised.

6.1 Comparison with other Methods

We have also implemented several clustering algorithms to solve the same problem. The results obtained by this method are broadly similar to those obtained by clustering. However, once the learning phase is done, the process of determining whether a given new fungus is unusual is much faster in this system: it is only a matter of coding the data, as opposed to computing the distances from all the clusters. Moreover, more different things are taken into account: so we are hoping it will prove more robust for large, poorly structured database.

7 CONCLUSIONS

This data mining tool has been constructed with the goal of classifying numerical data and identifying unusual cases within the data. An early version of the tool without any binning of values showed that where comparisons can be made using exact values, the classification process proves quite effective, testing showing that cases with many 'common' values were clearly identified with outliers showing up as having an unusually high encoding value. This requires that all attributes that group do so around exact values. Where this occurs cases can be identified as having similar encoding values. This similarity represents the fact that they have overall the same 'score' of common values. Specific patterns of values within the attributes will not necessarily be detected by this method nor will like cases necessarily share the same common attributes but outliers can clearly be identified as having the largest encoding value indicating the least number of common values.

However, the fungi data the tool has been constructed to analyse does not allow comparison of exact values and this is true of much data requiring such classification. This necessitates the binning together of values and the calculation of group frequencies using these groups to construct the encoding trees. Currently the tool uses a fixed number of bins consistent over all the attributes in an attempt not to emphasise one attribute over another. This approach seems to be flawed, often resulting in false grouping making the final encoding results unreliable, and will be replaced in later versions. In fact the combination of attribute range and number of bins results in some attributes becoming more significant in the encoding process where the range of attribute values varies.

What is needed to make this tool function correctly is a way of making the binning stage of the process more reliable. The approach of a uniform binning of all attributes would need to be replaced by some way of identifying optimal bins for each of the attributes. This could be done by performing some form of clustering on each of the attributes in turn, effectively identifying classes within a particular attribute and use those classes to bin an attributes values together. This would result in an optimal binning for each of the attributes based upon each attributes particular range and distribution of values.

The Huffman trees could then be built upon the frequencies identified for these optimal bins resulting in a more reliable classification. Minimum Message Length encoding is the underlying theory used to produce this classification tool. The idea being that the more frequently a value occurs the more 'common' it is and the minimum encoding length of a set of values can be calculated using this property. This is widely used in data compression applications where Huffman encoding is used to encode data in the shortest way possible giving the most common values the shortest encoding and resulting in a message of minimal length.

Using this technique to identify cases in data sets with few common values proves successful where data consists of specific discreet values. Where this is not the case, the problems associated with binning the data prevent the easy application of this method to many data sets. However, refinements are being implemented that will produce a more practical version of this tool applicable to a wider range of data.

8 BIBLIOGRAPHY

Adriaans, P. Zantinge, D. (1996). Data Mining. London: Addison Wesley Longman Ltd.

Agrawal, R. Faloutsos, C. Swami, A. (1993) Efficient Similarity Search In Sequence Databases: In Proc Of The 4th Intl Conf. On Foundations Of Data Organisation And Algorithms (Fodo'93) 1993.

Bollobas, B. Das, G. Gunopulos, D. Manilla, H.(1997). Time-Series Similarity Problems And Well-Separated Geometric Sets: Proc. 13th Acm Symp. Computational Geometry, Scg, Pp. 454-456, Acm Press, 4-6 June 1997.

Boulton,D,M. Wallace,C,S.(1968). A Program For Numerical Classification. The Computer Journal Vol 13. No 1. February 1970 (Pages 63 – 69)

Breiman, L. Freidman, J, H. Olshen, R, A. Stone, C, J. (1984) Classification And Regression Trees: Belmont California. Wadsworth.

Chan, P, K. Matheus, C, J. Piatetsky-Shapiro, G. (1993). Systems For Knowledge Discovery In Databases: Ieee Transactions On Knowledge And Data Engineering. Vol 5. No 6. December 1993.

Fayyad,U.M.(Ed.) Piatetsky-Shapiro,G.(Ed) Smyth,P.(Ed) And Uthurusamy,R(Ed). (1996a). Advances In Knowledge Discovery And Data Mining. Menlo Park, California: The Aaai Press/Mit Press.

Hou,W,C.(1996). Extraction And Application Of Statistical Relationships In Relational Databases. IEEE Transactions On Knowledge And Data Engineering, , Vol 8, No 6, December 1966

Jain, A, K. Dubes, R,C. (1988). Algorithms For Clustering Data: New Jersey: Prentice-Hall, Inc.

Keim, D, A. Kriegel, H, P.(1993). Using Visualization To Support Data Mining Of Large Existing Databases: In Proc. IEEE

Oliver, J. Hand, D.(1994). Introduction To Minimum Encoding Inference. Tech Report 4-94, Dept Of Statistocs, The Open Universtiy, Walton Hall, Milton Keynes Mk7 6aa

Piatetsky-Shapiro, G. (Ed.) Frawley, W.J.(Ed). (1991). Knowledge Discovery In Databases. Menlo Park, California: The Aaai Press/Mit Press.

Quinlan ,J.R. (1982).Semi-Autonomous Acquisition Of Pattern-Based Knowledge: New York, Gordon And Breach.

Quinlan ,J.R. (1990). Induction Of Decision Trees: San Francisco: Morgan Kaufman Publishers Inc.

Quinlan ,J.R. (1993). C4.5: Programs For Machine Learning: San Francisco: Morgan Kaufman Publishers Inc.

Quinlan, J, R.(1993). C4.5: Programs For Machine Learning: San Francisco: Morgan Kaufman Publishers Inc

Quinlan, J. R. And. Rivest, R. L (1987) Inferring Decision Trees Using The Minimum Description Length Principle: Technical Memo, Massachusetts Institute Of Technology, Laboratory For Computer Science, Number Mit/Lcs/Tm-339, P. 22, September 1987.

Reese, G. (1997). Database Programming With Jdbc And Java. Sebastopol Ca: O'reilly & Assc

Risasanen, J. (1987). Stochastic Complexity. Journal of the Royal Statistical Society (Series B), 49:223-239

Sedgewick, R. (1988). Algorithms. Menlo Park California. Addison-Wesley Publishing Inc

Sedgewick, S.(1988). Algorithms: Menlo Park, California. Addison-Wesley Publishing Company Inc.

Shavlic, J,W. Mooney, R, J. Towell, G.G. (1991). Symbolic And Neural Learning Algorithms: An Experimental Comparison. Machine Learning, Vol 6, No 2, Pages 111-143, 1991.

Siddalingaiah,M. Lockwood, S,D.(1997). Java Api For Dummies-A Quick Reference. Foster City Ca: Idg Books Worldwide.

Silberschatz,A. Tuzhilin,A(1996).What Makes Patterns Interesting In Knowledge Discovery Systems. IEEE Transactions On Knowledge And Data Engineering, Vol 8, No 6, December 1966

Wallace, C. S. and Freeman P.R. (1987) Estimation and Inference by Compact Coding, Journal of the Royal Statistical Society (Series B), 49:240-252.

Wallace,C.(1990). Classification By Minimum-Message-Length Inference. Advances In Computing And Information - Icci '90 Springer-Verlag S Lncs V 468p 72-81m May D 1990

Weiss, S,M. Indurkhya,A,N.(1998). Predictive Data Mining. San Francisco: Morgan Kaufman Publishers Inc.

Westphal, C. And Blaxton, T. (1998) Data Mining Solutions. New York: John Wiley And Sons Inc.

Zupan, J.(1982). Clustering Of Large Data Sets: London: John Wiley & Sons Ltd.

Zytgow, J,M. Baker, J.(1991). Interactive Mining Of Regularities In Databases. Knowledge Discovery In Databases, (Pages 31-53), Menlo Park, California: The Aaai Press/Mit Press.