On Temporal Validity Analysis of Association Rules

Bert Arnrich[1], Jörg Walter[1], Alexander Albert[2]
[1]*Institute of Neuroinformatics, University of Bielefeld, Germany*
[2]*Clinic for Cardiothoracic Surgery, Heart Institute Lahr/Baden, Germany*

Association rule mining [1] is a prominent data-mining method used in many domains. Despite the fact that most large datasets are collected over longer time spans, the considered systems are in most cases assumed stationary, which leads to complete ignorance of temporal effects.

In this contribution we present statistical and discretization techniques of partitioning the data recording time into intervals where the considered association rules remain homogeneous with respect to their support and confidence. In contrast with previous work where the considered time intervals are fixed they are determined in a data driven manner, which introduces a problem of optimal time granularity [2].

Furthermore, we demonstrate applicability of the risk-adjusted quality assessment in medical domain, specifically as it relates to heart surgery. For example, in comparison with the Euroscore risk system [3] the outcome prediction models for duration of intensive care or mortality can be significantly enhanced. Interesting pattern changes can be identified and assigned to systematical and organizational modifications of the considered system.

References

[1] Agrawal C and Srikant R. Fast Algorithms for Mining Association Rules. VLDB-94. 1994

[2] Liu B, Ma Y and Lee R. Analyzing the Interestingness of Association Rules from the Temporal Dimension. Proc. of the 1st IEEE Int'l Conf. on Data Mining. 2001

[3] Roques F, Nashef SA, Michel P, Gauducheau E, de Vincentiis C, Baudet E, Cortina J, David M, Faichney A, Gabrielle F, Gams E, Harjula A, Jones MT, Pintor PP, Salamon R, Thulin L. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. Eur J Cardiothorac Surg. 1999 Jun;15(6):816-22

Introduction

The objective of data mining is to find interesting and useful knowledge hidden in databases. A prominent technique is association rule mining which aims at finding comprehensible rules to describe regularities. In the past few years a large number of algorithms were proposed and tested in various domains. Despite the fact, that most large datasets are collected over a long time period the stationary of the considered system is assumed and temporal effects on the behavior are rarely captured, see e.g. [2]. In this paper we introduce a combination of statistical, discretization and association rule techniques in order to systematically analyze temporal effects.

Standard association rule mining is commonly stated as follows [1]: Let $I=\{i_i, ..., i_n\}$ be as set of items, and D be a set of records. Each record consists of a subset of items in I. An association rule is an implication of the form $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$. The set X is often termed antecedents and Y consequence. The rule $X \rightarrow Y$ holds in D with confidence c if $c\%$ of records in D that support X also support Y. The rule has support s in D if $s\%$ of records in D contains $X \cup Y$. Given a set of records D, the problem of mining association rules is to discover all rules that have support and confidences greater than the user-specified minimum support and minimum confidence.

In previously suggested techniques (e.g. in [2]) the dataset is first partitioned into sub-datasets D_i corresponding to manually chosen time periods in which they were collected (e.g. month, years, etc.) and the rules are mined. For rules that appear not in all D_i the missing support and confidence information is obtained in certain time periods. A rule is classified as a stable rule if none of its confidences (or supports) in the time periods is below the minimum confidence (or the minimum support) and the confidences (or supports) over time do not, i.e., they are homogeneous respective a Chi-Square test (see Table 1). The aim of rule reduction is achieved by presenting only stable rules which can be trusted in the future to the user. This method however, has two major shortcomings: the granularity of the time periods has to be manually specified depending on the application domain and for classification purposes potentially useful unstable rules are discarded.

	time period T_1	time period T_2	Row total
satisfy $X \wedge Y$	N_{11}	N_{12}	$N_{11} + N_{12}$
satisfy $X \wedge \neg Y$	N_{21}	N_{22}	$N_{21} + N_{22}$
Column total	$N_{11} + N_{21}$	$N_{12} + N_{22}$	

Table 1: confidence stability tests based on a contingency table for the rule $X \rightarrow Y$ in the two time periods T_1 and T_2

In this paper we argue, that an important improvement for rule classification systems based on historically grown datasets can be achieved, if the rule variability over time is systematically estimated and employed in the classification model.

Our experimental results are based on a research database in heart surgery. Heart operations are the most frequent realized surgeries in Europe and North America. The main surgery indication is summarized as cardiovascular disease (CVD) which includes coronary heart disease, high blood pressure, atherosclerosis, and stroke. CVD is an important research topic since it causes nearly half of all deaths in Europe [12].

The Clinic for Cardiothoracic Surgery of the Heart Institute Lahr is a highly specialized hospital and performs about 2000 open heart operations per year. For historical reasons the Heart Institute choose to operate independent clinical information systems (HIS) since its beginning in 1995. A data mart system integrates all relevant current and historical data from several disconnected HIS operated by autonomous departments. The transformation from HIS to target data base is done with carefully designed rules in tight cooperation with domain experts. For example the attribute *left ventricle ejection fraction* is gained from 7 different source values . Due to the historical changes in the source data base structures (caused by software updates, report form changes, variation in the physician team, etc) the explanatory power of certain parameters differs over the time they were collected. For example the correlation of a certain risk factor with postoperative mortality is higher in some time periods and significant lower in others. Other common sources of variation over time in the medical domain are the learning effect, i.e., the experience of the health professionals [4], and the enhancements of techniques in diagnosis and therapy [5]. Up to now, data from more than 13.000 heart operations with 277 pre-, intra- or postoperative attributes per case are available for multiple purposes.

In the following we will describe the used methods, report the significant performance improvements for the classification of postoperative outcomes in the result section and discuss the retrospective assignment of changes in organization and staff structure based on the temporal analysis.

Methods

In general the estimation of time periods can be traded as a binary discretization problem of a continuous attribute. Here a set of records D with k classes $C_1, ..., C_k$ has to be partitioned into the subsets D_1 and D_2 according to a threshold value T of a continuous-valued attribute A. Previously suggested discretization methods can be classified into global vs. local, supervised vs. unsupervised, and static vs. dynamic approaches; see [6] for a good overview. In this work we used global, supervised and static discretization algorithms for a two class problem: all records which accomplish the rule (satisfy $X \wedge Y$) belongs to class one and the others where

only the antecedent and not the consequence is present (satisfy X $\wedge \neg$Y) fall in class two. With this approach we can employ the well known evaluation functions (reported compactly below) for the discretization of continuous-valued attributes.

Independent of the algorithmic strategy that is used for partitioning it is important to ensure that obviously "bad" partitions, i.e., a sequence of records belonging to a single class should not be broken apart, are not selected by the evaluation function [7]. Fayyad and Irani [8] defined the concept of a boundary point: A value T in the range A is a boundary point if in the sequence of records sorted by the value of A, there exist two records r_1, $r_2 \in D$, having different classes, such that $A(r_1) < T < A(r_2)$; and there exists no other example $r' \in D$ such that $A(r_1) < A(r') < A(r_2)$. In other words, a threshold value that separates two successive records that all belong to the same class is not a boundary point.

In the following we briefly introduce five evaluation functions for finding optimal split points. Of course only boundary points are candidate values. Note that simple blockwise dataset partitioning does not ensure selection of meaningful boundary points.

1. The **class information entropy** ***E(A,T;D)*** of a set D partitioned into two sets D_1 and D_2 induced by threshold value T of the continuous-valued attribute A is the weighted average of their resulting class entropies:

$$E(A,T;D) = \frac{|D_1|}{|D|} Ent(D_1) + \frac{|D_2|}{|D|} Ent(D_2)$$

where $Ent(D_j)$ measured the amount of information needed, in bits, to specify the k classes in D_j and $P(C_i, D_j)$ is the proportion of records in D_j that have class C_i:

$$Ent(D_j) = -\sum_{i=1}^{k} P(C_i, D_j) \log(P(C_i, D_j))$$

A binary partition for D is determined by selecting the boundary point T_A for which $E(A,T_A;D)$ is minimal along all the boundary points. Based on the Minimum Description Length Principle a partition induced by a boundary point is accepted or rejected otherwise; for details see [8].

2. The **gain ratio criterion** [9] assesses the desirability of a partition as the ratio of its information gain $Gain(D, T)$ to its split information $Split(D, T)$:

$$Gain(D,T) = Ent(D) - \sum_{j=1}^{2} \frac{|D_j|}{|D|} Ent(D_j)$$

$$Split(D,T) = -\sum_{j=1}^{2} \frac{|D_j|}{|D|} \log\left(\frac{|D_j|}{|D|}\right)$$

3. The **Contrast-Entropy criterion *CE*** [10] favors partitions with high contrast by maximizing of the Euclidean distance between two partitions and minimizing the distance of the elements within each of them and low entropy. With the mean value m_i for the continuous-valued attribute A in the partition D_i it is defined by:

$$CE(D_1, D_2, T) = \frac{Contrast(D_1, D_2, T)}{Ent(D)}$$

$$Contrast(D_1, D_2, T) = \frac{|D_1| \cdot |D_2|}{|D|}(m_1 - m_2)^2$$

4. At the **Chi-Square method** a contingency table (see Table 1) is built for each boundary point. The partitioning with the highest significance level (tested with Chi-Square) is chosen.

5. The **Kolmogorov-Smirnov statistic** is principally used to compare two cumulative distribution functions based on the maximal absolute difference D_{max} between them. A significance level (null hypothesis imply that the distributions are the same) can be computed. At our two-class problem ($X \wedge Y$ vs. $X \wedge \neg Y$) we induce a partition at D_{max}.

All proposed methods can be used in a recursive manner: beginning with all records where $X \wedge Y$ or $X \wedge \neg Y$ is present the best binary partitioning according to the value of the evaluation function is chosen. Subsequently the resulting partitions are analyzed until a stopping criterion (e.g. minimum significance level or number of partitions) is reached.

The resulting partitioning of a rule $X \rightarrow Y$ can be used for the retrospective improvement of the classification performance on Y. This is crucial if like in our case surgical quality assessments (e.g. postoperative mortality as consequence) are done. In general association rules may be used for classification purposes in the following way: for each antecedent X an attribute which satisfy X is generated and entered in the classification model. To introduce the partitioning in the model each attribute is split into new attributes according to the found partitions, i.e., for each partition a new attribute is generated. The area under the receiver operating characteristic (ROC) curve (see [11] for a detailed description) was used to measure the discrimination power of the underlying quality assessment model.

Assessing the quality of cardiac surgical care through inter-hospital and inter-surgeons comparison of mortality rates and complications after cardiac surgery is of increasing importance. One of the established risk score systems for postoperative mortality in Europe is the European System for Cardiac Operative Risk Evaluation (EuroSCORE) [3]. The data mart database has enabled us to apply the EuroSCORE retrospectively in the Heart Institute in 75% of all cases even though most of the parameters are never collected in a way that comply the exact definitions in [3]. In this study we could analyze

8758 cases, where 181 died and 777 had a prolonged stay in the intensive care unit (ICU). To get comparable classification results with the original study in [3] we used the stepwise logistic regression as risk model. Beside postoperative mortality as outcome we examined also the classification of an prolonged stay in ICU.

To inspect the temporal variation of the correlation between the risk factors with postoperative outcomes we focussed on rules with one antecedent and one consequence, i.e., rules in the from EuroSCORE parameter $\rightarrow$ postoperative outcome.

Results

The resulting partitioning of each risk parameter were introduced in the model as described in the method section. As shown in Table 2 the best discrimination power of the different models were achieved by using the Kolmogorov-Smirnov statistic. For both outcomes the classification results are clearly better with the additionally usage of the found partitions in the particular risk model. A significant improvement was achieved for the prolonged stay in ICU. Figure 1 visualize found time intervals where rule confidences differ.

Model	Area under ROC with 95% confidence interval
Mortality	0.771 (0.736 - 0.806)
Extended model for Mortality	0.782 (0.746 - 0.817)
ICU stay > 7d	0.738 (0.720 - 0.757)
Extended model for ICU stay > 7d	0.760 (0.743 - 0.788)

Table 2: Discrimination power of the logistic regression models.

Furthermore the analysis of time variations in association rules allows the identification of temporal irregularities in data collection. Figure 2 show, that significant fewer recent myocardial infarcts were diagnosed in the Heart Institute since November 1997. A subsequent review revealed that a reporting procedure was modified at that time which results in an imprecise recording of this attribute. Actions will be taken in the Heart Institute to get this value from other sources.

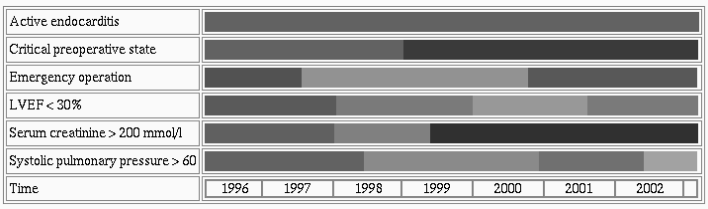

Figure 1: Resulting time segments for six attributes which correlate with an ICU stay longer than 7 days. The color coding indicates the confidences of the particular rules in different time segments such that as darker a time interval appears, the stronger the association. Note that the segments obviously do not correspond to year segments and our approach avoid the granularity problem of finer segments.

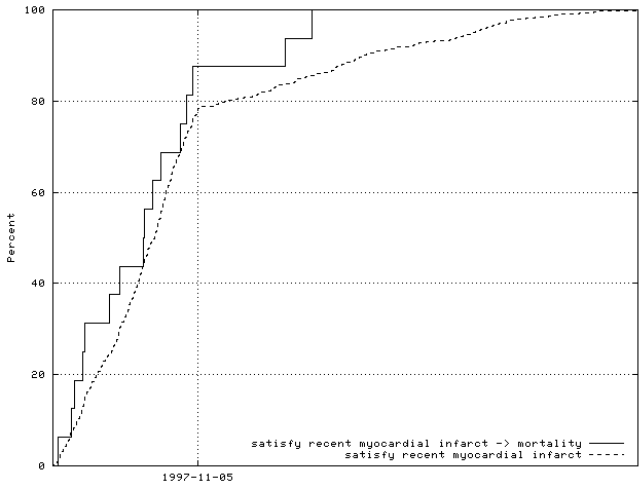

Figure 2: Cumulative sum of high risk cases which are marked having recent myocardial infarct and cases which additionally died within 30 days after operation.

Discussion

In medicine well known factors which influence clinical practice and contributing to outcome are experience of the individual staff members, work environment, organization and management factors [13]. Beside the improvement of quality assessment it is valuable to identify causes of temporal performance changes. On the basis of the temporal analysis of the

EuroSCORE risk factors we were able to assign changes in organization and staff structure to postoperative outcomes. The following two examples for postoperative mortality (pulmonary pressure and ejection fraction) shall illustrate the procedure.

Severely elevated systolic pulmonary pressure usually developed in patients with long existing heart failure caused in most cases by heart valve disease over long years. The postoperative care on these high risk patients on ICU demands an extraordinary experience of the health professionals because slight treatment failures can cause fatal events. A similar situation concerns patients with reduced ejection fraction where the ability of the heart muscle to eject blood is impaired. In the beginning of 2002 the cardiosurgical department of the heart institute, especially the ICU team, experienced significant changes in staff and management. As shown in Figure 3 the conditional mortality rates for reduced ejection fraction is increasing since the ICU reorganization. Actions were taken including the ICU staff (e.g. the engagement of two new specialists for intensive care medicine), organizational and management changes and resulted in regaining good performance since beginning of 2003.

The elevated mortality rates on reduced ejection fraction and high pulmonary pressure since the end of March in 2000 and middle of June in 2000 (see Figure 3 and 4) correlates with the discharge of an experienced surgeon who was specialized on high risk patients, i.e., from this time the other partly less experienced surgeons took over these cases.

Since end of 1997 the operations performed per day increased, but the capacity of the ICU was extended not until the end of 2000. As shown in Figure 1 this leads to an early discharge between 1998 and 2000 in other clinics even in high risk patients (those with emergency operation or elevated systolic pulmonary pressure). Since end of 2000 a longer medical care of severe illness patients in the Heart Institute is possible again.

These three examples show, that a retrospective temporal analysis can gain new insights in correlation between organizational aspects and postoperative outcomes which were not recognized before. Although there were presumptions about temporal performance variations (e.g. consequences results from the discharge of the most experienced surgeon) but for the first time the changes were analyzed systematically. Indeed for some observed correlation a plausible cause could not be identified yet.

Comparative studies with more historical grown real-life datasets and synthetic data to investigate the specific behavior of evaluation functions for partitioning the time dimension are planned.

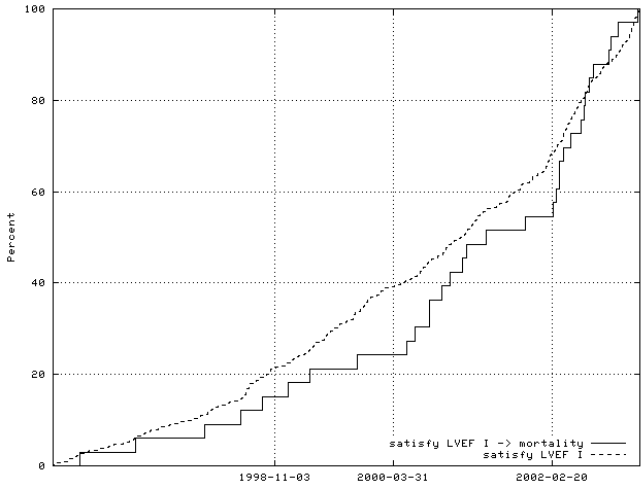

Figure 3: Left ventricle ejection fraction and conditional mortality over time. The elevated mortality rate after ICU reorganization in the beginning of 2002 is conspicuous identifiable. The increase since April 2000 correlates with the discharge of an experienced surgeon.

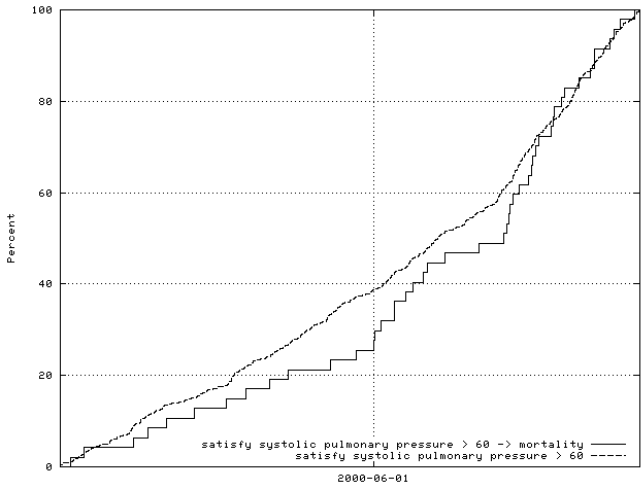

Figure 4: Elevated systolic pulmonary pressure and conditional mortality over time. Like above the increase in June 2000 correlates with the discharge of an experienced surgeon.

Additional References

[4] Altman DG. Hidden effect of time. Statistics in Medicine. 1988. 7:629-637

[5] Kirklin JW, Barratt-Boyes BG, Kouchoukos NT, Blackstone EH, Hanley FL, Doty DB, Karp RB. Cardiac Surgery. Elsevier Science. 2003

[6] Dougherty J, Kohavi R and Sahami M. Supervised and Unsupervised Discretization of Continuous Features. International Conference on Machine Learning. 1995. 194-202

[7] Elomaa T and Rousu J. On the Well-Behavedness of Important Attribute Evaluation Functions. Scandinavian Conference on AI. 1997. 95-106

[8] Fayyad UM and Irani KB. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. Proceedings of the 13th International Joint Conference of Artificial Intelligence (IJCAI). 1993. 1:1022-1027

[9] Quinlan JR. Improved Use of Continuous Attributes in C4.5. Journal of Artificial Intelligence Research. 1996. 4:77-90

[10] Van de Merckt T. Decision trees in numerical attribute spaces. Proceedings of the 13th International Joint Conference of Artificial Intelligence (IJCAI). 1993. 1:1016-1021

[11] Hanley JA and McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology. 1982. 143:29-36

[12] Rayner M and Petersen S. European cardiovascular disease statistics 2000 edition. British Heart Foundation. 2000

[13] Vincent C. Understanding and responding to adverse events. New England Journal of Medicine. 2003. 348(11):1051-1056.