Sequential Data Mining: A Comparative Case Study in Development of Atherosclerosis Risk Factors

Jiří Kléma, Lenka Nováková, Filip Karel, Olga Štěpánková, and Filip Železný

***Abstract*—Sequential data represent an important source of potentially new medical knowledge. However, this type of data is rarely provided in a format suitable for immediate application of conventional mining algorithms. This paper summarizes and compares three different sequential mining approaches based, respectively, on windowing, episode rules, and inductive logic programming. Windowing is one of the essential methods of data preprocessing. Episode rules represent general sequential mining, while inductive logic programming extracts first-order features whose structure is determined by background knowledge. The three approaches are demonstrated and evaluated in terms of a case study STULONG. It is a longitudinal preventive study of atherosclerosis where the data consist of a series of long-term observations recording the development of risk factors and associated conditions. The intention is to identify frequent sequential/temporal patterns. Possible relations between the patterns and an onset of any of the observed cardiovascular diseases are also studied.**

***Index Terms*—Anachronism, episode rules, inductive logic programming, temporal pattern, trend analysis, windowing.**

I. Introduction

MEDICAL databases have accumulated large quantities of information about patients and their clinical conditions. Relationships and patterns hidden in this data can provide new medical knowledge as has been proved in a number of medical data mining applications. However, the data are rarely provided in a format suitable for immediate application of conventional attribute-valued learning (AVL). In some tasks, a domain-independent preprocessing methodology (e.g., feature selection) is sufficient. In other tasks, domain-specific preprocessing proves vital and may strongly increase mining performance. But, the domain-specific algorithms are frequently applied in a trial-and-error manner, which is often time consuming and demands both an experienced researcher and a medical expert.

The present paper focuses on mining temporal and sequential medical data, which usually asks for complex preprocessing. We deal with sequences of events where each event is described by a numeric or symbolic value and a time stamp. Event types can also be distinguished. The dataset can either be a single sequence or it can be composed of a number of sequences. The ultimate goal is to identify strong sequential patterns, i.e., such event chains (subsequences) that appear frequently in the dataset, and optionally study their interaction with the target event. The event chains may also be transformed into an alternative representation (i.e., trends, generalizations, etc.). The typical target event in a medical application can be a disease manifestation or a change of the state of health.

Mining and learning in sequential data represents an important and popular research field. The problem of mining sequential patterns is introduced in [1], which also presents three Apriori-based algorithms to solve it. The authors confine to a transactional dataset with distinct symbolic items. The problem of sequential supervised learning is thoroughly discussed in [2]. The main goal is to construct a classifier that correctly predicts future events, given their history. Although we go beyond this definition as our main goal is to introduce and interpret meaningful patterns rather than to construct a classifier, the paper stresses the common issues in sequential data mining: construction of features capturing long-distance or complex interactions and adjoined computational costs. The paper presents the sliding window method as an important technique of sequential data preprocessing. Lesh *et al*. [3] adapt data mining techniques to act as a preprocessor to select features for standard classification algorithms. The papers [4]–[6] introduce the term "episode" and define the standard episode rule mining problem—to find all episode rules satisfying the given frequency and confidence constraints. They also propose two algorithms: 1) the Winepi algorithm based on the occurrences of the patterns in a sliding window along the sequence and 2) the Minepi algorithm that relies on the notion of minimal occurrence of patterns. Meger and Rigotti [7] present a general tool WinMiner allowing to search for episode rules while the smallest window size that corresponds to a local maximum of rule confidence is found.

Regarding medical applications, [8] proposes a first-order rule discovery method for handling irregular medical time-series data. A vast effort was made to analyze a temporal hepatitis dataset [9], [10]. The latter paper proposes a hybridization of phase-constraint multiscale matching and rough clustering to find similar patterns in temporal sequences.

The main contribution of this paper lies in parallel application and comparison of three different sequential mining techniques. As regards novelty of the individual methods themselves, the application of inductive logic programming (ILP), i.e., the tool relational subgroup discovery (RSD) represents an innovative approach to sequential data mining. Although the windowing technique is not new, the paper deals with related issues such as missing values, anachronism, or window length optimization.

Manuscript received October 9, 2006; revised February 1, 2007 and May 25, 2007. This work was supported in part by Czech Ministry of Education under the Programme MSM 6840770012 Transdisciplinary Biomedical Engineering Research II, and in part by Czech Academy of Sciences under Grant 1ET101210513. This paper was recommended by Guest Editors Z. Wang and X. Liu.

The authors are with the Department of Cybernetics, Czech Technical University in Prague, Prague 16627, Czech Republic (e-mail: klema@labe.felk.cvut.cz; novakova@labe.felk.cvut.cz; karelf1@fel.cvut.cz; step@labe.felk.cvut.cz; zelezny@labe.felk.cvut.cz).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCC.2007.906055

Episode rule mining, addressed in Section VI, is inspired by an approach suggested in [11], which we enhance here. Lastly, we identified numerous sequential patterns whose value was acknowledged by the domain expert.

This paper is organized as follows. The STULONG study is introduced in Section II, together with basic preprocessing issues such as introduction of prospective target variable(s) and the risk of anachronism, i.e., utilization of variables containing information that is unavailable when a decision/prediction is called for. Windowing is discussed in Section III. A special emphasis is put on missing value treatment and the optimal setting of the window length for various variables. Multivariate interactions among the windowed aggregates in the form of ordinal association rules are presented in Section IV. Section V deals with relational subgroup discovery within the first-order logical framework. The main outcome lies in conjunctive rules based on temporal features. The rules identify interesting subgroups of the total population. Section VI briefly introduces sequential episode rules and exemplifies their contribution to understanding of the STULONG risk factors (RFs). The conclusions in Section VII compare the individual methods and summarizes the lessons learnt.

II. STULONG Study

The study STULONG [12] is a longitudinal 20 years lasting primary preventive study of middle-aged men. The study aims to identify prevalence of atherosclerosis RFs in a population generally considered to be the most endangered by possible atherosclerosis complications, i.e., middle-aged men. It follows the development of these RFs and tries to discover their impact on the health of the examined men, especially with respect to atherosclerotic cardiovascular diseases (for further details, see acknowledgment and the web site [13]).

The study contains data resulting from 20 years of observation of approximately 1400 men in the middle age. The study asks both for nonsequential and sequential data mining as it concurrently follows substantially different goals. The intention of the project is not only to identify possibly new RFs, but also to study the influence of their time changes. It is generally known that overweight contributes to atherosclerosis, but it is not so obvious whether its changes (e.g., short-term weight losses and gains) play an additional role. This paper applies sequential methods to answer this type of questions and complement the more straightforward nonsequential analysis.

The data consist of four relational tables. This paper is concerned with two of them: table *Entry* describes the data collected during the entry examinations of patients (exactly one record per patient), table *Control* includes the results of a series of long-term observations recording the development of RFs and associated conditions (0 to 21 records per patient). By merging these two tables, it is possible to obtain a multivariate time series of examinations for each patient. The data have a nonuniform character. Although the same attributes are observed during repeated checkups for all patients, there is a significant difference in the number of examinations between individual patients. Some conducted measurements are missing in the data. In order to prepare these data for application of any AVL method, it is necessary to create a uniform attribute structure based on the available time-series data.

The first idea is to substitute the time series of separate measurements by one or more aggregated values, e.g., by the mean or extreme values. This technique is often used in medical domains (see, e.g., [14]). The trends of the observed variables are considered to be one type of RFs. The trends can be obtained as parameters of the regression line calculated from all available checkups for the given patient. We denote this approach as *global* (or en bloc) and treat it as a counterpart of windowing introduced later. But there is a pitfall in this approach. It can add anachronistic attributes [15] to the data because the equation for calculating trends (details in Section II-B) uses n, the total number of checkups. Certainly, the number of all checkups is an anachronistic attribute, because we do not know in advance, how many times the new patient will come. On the other hand, we do not need all patient's measurements to estimate the development trend of a certain attribute: the parameters of the regression line can do the job, and they can be obtained from the first two values already. Is it reasonable to compare regression parameters calculated from the training data with those of new patients calculated from their existing checkups? We will try to answer this question at the end of Section II-B.

One of the main goals of STULONG is predictive diagnostics of cardiovascular diseases (CVD). In this paper, the CVD label corresponds to occurrence of any of the following diseases: angina pectoris, (silent) myocardial infarction, ischaemic heart disease, silent myocardial ischaemia, and claudications. Prediction can be achieved by creating a model or patterns, which differentiate between a patient with CVD and a patient without CVD (non-CVD). The binary attribute CVD is used as the target function (a possible disease always appears during the last checkup). A precise temporal definition of CVD is given in Section III-A. This paper searches for derived attributes, which seem to relate to CVD; they are described and analyzed in Section II-B.

A. Verifying Relation Between the Number of Checkups and the Occurrence of a Cardiovascular Disease

Various statistical tests or visualization tools can be applied to verify dependency between the number of checkups and CVD. We have used the criterion "area under the ROC curve." A receiver operating characteristic (ROC) [16] curve is a graphical representation of the tradeoff between the true positive (TP) and false positive (FP) rates for every possible cutoff of the independent variable. In other words, the ROC curve is the representation of the tradeoffs between sensitivity and specificity. The plot shows the FP rate on the x-axis (1-specificity) and the TP rate on the y-axis (sensitivity). The area under curve (AUC) quantifies the ability of a model to separate two classes under different conditions. A random model shows AUC $= 0.5$, the perfect model shows AUC $= 1$, and the inverse perfect model has AUC $= 0$.

Let us consider a trivial model of CVD based on the number of checkups only. This model does nothing but orders the patients according to the number of their checkups. The model supposes that the more checkups the patient has, the more likely he is to

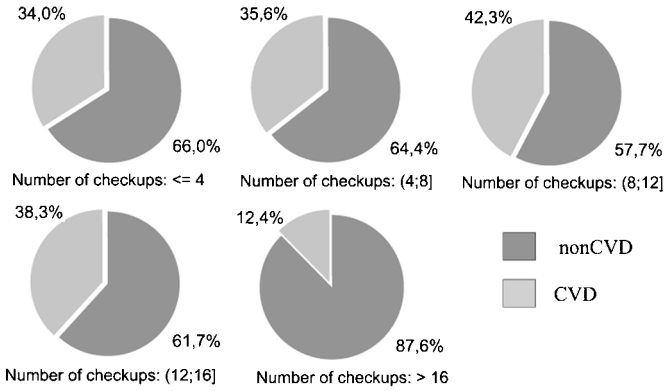

Fig. 1. Relation between the number of checkups and the frequency of cardiovascular diseases—pie chart.

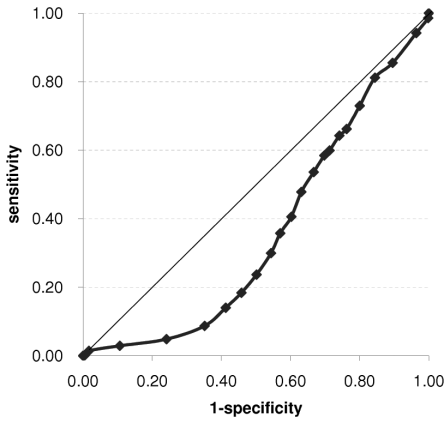

Fig. 2. Relation between the number of checkups and the frequency of cardiovascular diseases—ROC curve

be struck by CVD. It takes a random checkup threshold, which distinguishes between CVD and non-CVD patients (the checkup numbers below and above the selected threshold).

For every possible threshold value of the checkup number, the AUC (model_variable, target) expresses the probability that a patient will be correctly classified into one of the target groups, in our case CVD or non-CVD. By means of the nonparametrical Wilcoxon statistics, we have estimated the AUC (the number of checkups, CVD) $= 0.38$ (the 95% confidence interval is $[0.33, 0.45]$). In other words, the trivial model based only on the checkup number assigns patients to the correct class with probability 0.38. It is evident that the frequency of CVD significantly falls with rising number of checkups (AUC > 0.5). The corresponding ROC curve is shown in Fig. 1. AUC can also be used to relate the strength of the studied dependency to the influence of the other variables that are well-known to affect CVD occurrence—for example, body mass index (BMI), age, or cholesterol—AUC(BMI, CVD) $= 0.55$, AUC(age, CVD) $=$ 0.60, AUC(cholesterol, CVD) $= 0.58$. The comparison suggests that the checkup number makes probably the strongest RF in the study.

Similarly, it is possible to reject the independence hypothesis between the number of checkups and CVD by χ^2 test (on the level of significance $p = 0.005$). The relation can also be visualized, e.g., by the categorization pie chart (see Fig. 2). The graph shows the reduced frequency of disease in the patient group with more than 16 checkups.

We can explain the observed relation between CVD and number of checkups as a consequence of the methodology applied when monitoring the patients. Monitoring continued for up to 20 years with some patients. During the project, measurements were taken regularly each year. But monitoring was sometimes stopped earlier because the patient fell to CVD and his monitoring was cut. This patient was deleted from the monitored group and transferred to the clinical program of secondary cardiovascular disease prevention. Due to this procedure, a patient with fewer checkups is more likely to be struck by a CVD than a patient with many checkups.

Despite the commented causal relation between the occurrence of a cardiovascular disease of a patient and the total number of his checkups, the total number of checkups cannot be used to build the predictive model since it is not known in advance—it is an anachronistic attribute.

B. Using Aggregation Value for Repeated Measurements

The characteristic values (mean, standard deviation) or parameters of the regression line (curve) are often used as a means for simplification of time-series data. How are these values calculated? Let us consider, for example, the expression for estimation of the parameters k, q of the regression line $y = kx + q$, where y stands for values of the observed variable and x for time of measurement, or alternatively its sequence number

$$k = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$q = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

The regression equations refer to basic statistics. It is to be noted that the number of checkups n reappears in these expressions. Just as this attribute is anachronistic, so can be the derived attribute. Experiments on STULONG data prove that this is indeed the case. There is a strong dependency between the values of derived attributes calculated as the regression coefficients and the number of checkups used for their calculation (as experimentally proven in Section III-C). Consequently, the observation that CVD classification is closely related to the values of the regression coefficients can be due to their dependency on the number of checkups n. That is why these aggregation attributes should not be used for building of the CVD predictive model. The global approach as such is unsound—the regression parameters of series of different length cannot be matched, and

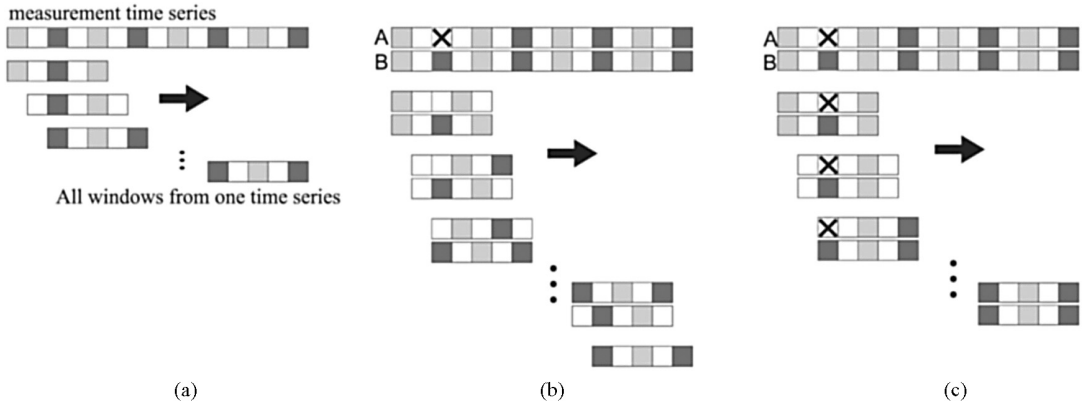

Fig. 3. Windowing a temporal data. (a) Basic sliding window. (b) Replacement of the missing value by a new value. (c) Shifting.

the new patient necessarily misses knowledge of the future total check-up number.

We have to search for a more sophisticated transformation of the input data in order to ensure that the predictive result is not influenced by the number of checkups. Only under such conditions can the time-series data be replaced by some aggregation values. We will apply windowing for this purpose.

III. Windowing

Windowing is a simple and often used method to transform data (see, e.g., [17]). Two types of windowing can be distinguished. The sequence of data can be either decomposed into several disjoint windows or a sliding window approach can be applied. The first choice is applied whenever we decide for a "per partes" linear approximation of the original data. Then, the windows correspond to intervals in which data exhibit "similar," i.e., close to linear behavior. These intervals can differ in length. On the other hand, the sliding window has a fixed length. It "slides" over the original series in regular steps, generating overlapping subseries. Both choices lead to a simplified representation of the input sequence. The sequential representation changes into the classical attribute-valued representation—the aggregate variables corresponding to a window are mapped into an individual output value.

In our task, we tackle to find a relation between individual RFs expressed as time trends and a possible development of CVD. For this purpose, the method of the fixed length sliding window seems most suitable. Generally, the sliding window method transforms a time series with n consecutive measurements into a new set of time series. It consists of items with a constant number of measurements, denoted as l. Of course, this approach can only be applied to time series consisting of at least l measurements. All shorter series with less than l measurements have to be neglected. The results of this transformation can be safely used in further analyses, as the resulting trends are not influenced by the number of controls, and the number of considered measurements is constant in the new set of data. The elementary transformation process of a temporal data is illustrated in Fig. 3(a).

A. Windowing—New Tasks

The proposed windowing method is parametrical. The choice of the window length can significantly influence the result, e.g., the fidelity, reliability, or predictive ability of a future model. Unfortunately, there is no universal recommendation to choose the appropriate length of the window *a priori*. This decision is task-dependent, most often based on the estimate of the minimal time period allowing to predict the result. We, thus, need to determine *the optimal window length in the STULONG study*. Another important issue applies to the missing data. As patient examinations miss certain values of usually measured variables, our second task is to enhance windowing *to deal reasonably with missing data* so that it aligns the measurements of different variables with the same time tag while losing minimum measurements that are present.

When using any windowing method, one should also introduce a more sophisticated CVD distinction. Certainly, there is a significant difference between a patient who will show some slight signs of CVD after 20 years and the patient who will prove a strong CVD onset during the next examination. The value of attribute CVD must be related to the considered time of measurement. We introduce a derived attribute CVD_i whose value is equal to the distance in years from the present moment to the time when the patient gets CVD. When the patient remains healthy, the value of the attribute CVD_i is set to a distinguished number, which is ever interpreted as "healthy."

B. Windowing and Missing Data

The windowing method with the fixed window length forms windows (subseries) that can be concisely represented by aggregate attributes. Two principal approaches can be applied when considering time series with occasional missing values.

1) The first one insists on the fixed length of the time window—reasonable substitutes for the missing values have to be found: the replacement approach is applied.
2) The second one adheres to the fixed number of checkups—missing values are omitted and the next values in the series are used: the series of values is shifted.

Replacement of the missing value by a new value [Fig. 3(b)] has to guarantee pertinence of the new value. For attributes

developing relatively slowly and smoothly, the mean calculated from the former and the future value can be a suitable replacement. This option was applied in STULONG. It has been verified that in majority of the patients, the windowed values of different variables truly represent the same period. In other words, there are only a few variables and time periods that exhibit long series of missing values, causing multiple propagation of the same and possibly out-of-date value. Replacement of the missing value by shifting the series [Fig. 3(c)] can cause a severe problem of synchronization. Let us consider an object described by two time series corresponding to development of variables (attributes) A and B. The missing value in variable A is marked by a cross. When we transform the data, we replace this missing value by the next value. We shift only variable A, because there are no missing values in variable B—the windows are not synchronized any longer. We have fewer windows for variable A than for variable B and the windows are mutually shifted. The corresponding aggregates are created from the measurements of different time stamps.

The other possibility is to omit those measurements whenever the value of one variable (A or B) is missing. This solves the time-shift problem as the window remains identical in time, but we are losing a part of valuable measurements. In our case, it means that we would have used only those checkups, which are complete (all measurements have been taken for the considered patient). This approach is reasonable only for problems where only few values are missing.

In spite of its theoretical simplicity, windowing calls for excessive data preprocessing; in our case, it was significantly eased by a general data preprocessing tool SumatraTT [18]. It also served for all the generic preprocessing operations.

C. Optimal Window Length

Since the number of examinations in our case ranges from 1 to 21, the choice of the window length is a tradeoff between the amount of data that have to be omitted and the length of the observable period. For the STULONG study, the window length of five measurements has been chosen first. The histogram and the pie charts (Fig. 4) show that using this option, data concerning about one-fifth of the patients have to be rejected due to the fact they have less than five records. If we insist on utilization of a longer window of eight or ten measurements, this would happen with data of more than a third of the patients.

Let us now focus on trends (mathematically the regression slope k), one of the most interesting and natural aggregate types from the point of view of physicians and patients. The trends of five important variables have been considered: the systolic and diastolic blood pressure (SYSTGrad, DIASTGrad), cholesterol (CHLSTMGGrad), triglyceride level (TRIGLMGGrad), and BMI (BMIGrad). The trends have been calculated for windows of length five, eight, and ten examinations (denoted as W5, W8, and W10). For simplicity, a specific modification of the CVD_i introduced in Section III-A has been taken, i.e., the attribute CVD_1. CVD_1 is "true" if and only if the CVD appears at the examination directly following the last windowed examination, i.e., the patient develops a CVD in one year from the windowed period.

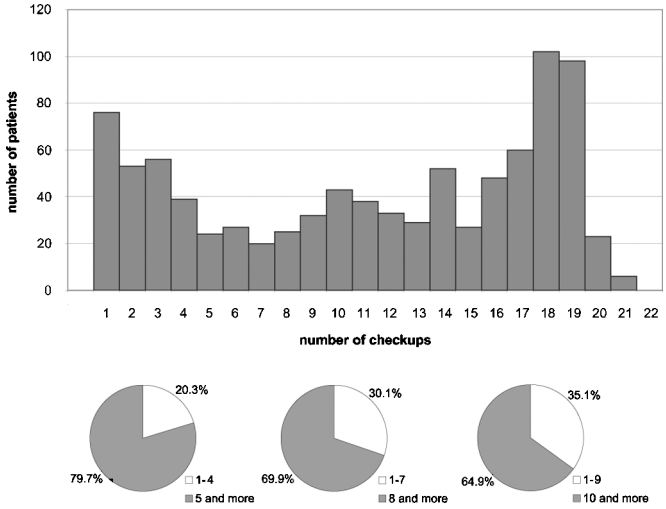

Fig. 4. Histogram of examinations in the STULONG study.

TABLE I
DEPENDENCIES BETWEEN CVD AND SELECTED TREND, χ^2 TEST

10 intervals	Global	W5	W8	W10
SYSTGrad	0.065	**0.005**	0.703	0.571
DIASTGrad	0.078	**0.072**	0.114	0.683
CHLSTMGGrad	0.005	0.497	0.487	0.950
TRIGLMGGrad	0.002	0.321	0.183	0.624
BMIGrad	0.007	0.804	0.746	**0.061**

Table I shows the level of significance p of the χ^2 test of independence between CVD_1 and the trends defined earlier (the trend variables were discretized into ten distinct equidepth intervals before testing). In the first column, there are results for the global approach where the trends are computed from all available measurements. This approach suggests a strong dependency between all the trends and CVD. When using windowing, only SYSTGrad, DIASTGrad, and BMIGrad seem to correlate with CVD_1. Moreover, this correlation can be observed for specific window lengths only (SYSTGrad and DIASTGrad when using W5, BMIGrad with W10).

We have further analyzed the obtained dependencies. As for the global approach, the trends of all the concerned RFs correlate with occurrence of the disease in the same way: when the trends are strongly increasing or decreasing, CVD is more likely to develop. Conversely, if the trends are stable, CVD is less likely to appear.

The results obtained by windowing suggest weaker influence of the studied trends. For SYSTGrad and DIASTGrad (and W5), it holds that *the more the blood pressure increases, the more likely is the CVD*. The same kind of dependency was observed for BMIGrad and W10. These dependencies are plausible and confirm that windowing is a reliable way for their retrieval. On the other hand, our experience shows that especially when dependencies searched for are weak, experiments with various window lengths should be accomplished. For example, BMI tends to oscillate between the examinations. Some patients show

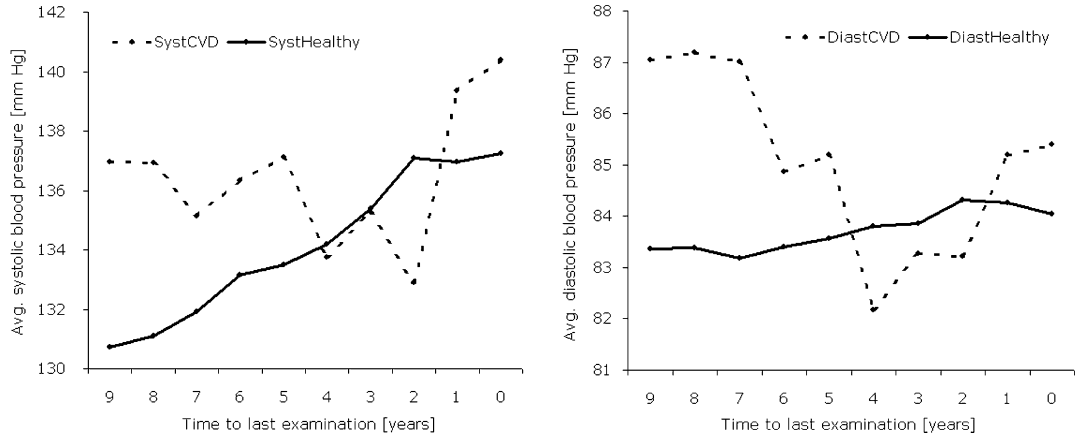

Fig. 5. Ten-year course of average diastolic and systolic blood pressures in the CVD and healthy groups.

abrupt changes of their weight. These changes do not seem to influence their health as much as slow but long-term weight gain. A long time window is needed to distinguish between these types of changes, and therefore, BMI has to be followed for ten examinations at least.

On the contrary, blood pressure can exhibit different types of behavior in time. The preliminary attempt to identify some pattern in time development of the blood pressure value can be based on comparison of its average values for the persons who came down with a CVD during the study and for the healthy people (see Fig. 5). Fig. 5 shows the ten-year development of average diastolic and systolic blood pressures for two patient groups: the first group (solid lines) represents individuals who remained healthy, the second group (dashed lines) corresponds to the patients whose last examination identified a CVD. There is a striking difference between both the groups: while the healthy patients exhibit nearly linear development, average blood pressure of patients with CVD tends to decrease first and then increase back or even higher. It seems probable that this type of behavior can be observed in individual patients with CVD as well. But to prove this hypothesis, new aggregated attributes would have to be introduced (qualitative shapes, time derivatives of the second order). Obviously, the linear trends calculated for the longer windows cannot serve this purpose since they tend to balance between decreasing and increasing trends in the CVD group, which makes them unable to discriminate between the groups.

IV. Windowed Aggregates as Risk Factors

The next step is a search for possible multivariate interactions among the generated trend aggregates, interactions among the trends and other RFs, and finally, subgroups with the above-average CVD risk.

Multivariate analysis in the STULONG study can often suffer from an insufficient number of patients belonging to the potential target groups. As a result of the increased data dimensionality incurred by multivariate analysis, statistical techniques typically do not yield conclusions with acceptable confidence intervals. Furthermore, due to the vast amount of possible hypotheses in the multivariate case, their manual formation becomes infeasible and we, thus, resort to the automated technique of association rule mining.

In the rest of the paper, we constrain ourselves to *the trends and means generated using W5*. The other aggregate types are not considered here in order to keep a reasonable number of attributes and to minimize random dependencies. The dataset is further modified so that each patient is considered only once. Each patient who develops CVD during the study represents a positive example—the data from his penultimate examination are taken into account, and the aggregates are calculated from the last five examinations prior to the onset of CVD. The patients who did not come down with a CVD represent negative instances. Their trends are calculated from their five central examinations; the examinations are taken in such a way that the average age in both the groups (positive and negative) is the same. The goal is to minimize the influence of patient age, the preliminary experiments resulted in the selection of examinations 8–12 in the negative group. It is to remember that this approach guarantees independence of the considered instances/transactions.

A. Newly Aggregated Variables

In Section III, we have dealt with five selected variables only (SYST, DIAST, CHLSTMG, TRIGLMG, and BMI). In this section, a couple of new variables have been added from original tables—HDL and LDL stand for high- and low-density lipoprotein cholesterol carriers, while POCCIG gives the average number of cigarettes smoked per day. The χ^2 independence test has been applied to study their relation to CVD. Two trends proved to have strong influence.

1) A decreasing HDL level clearly relates to the increasing risk of CVD ($p = 0.001$). HDL represents the "good" cholesterol, and it is well known that the low HDL level is associated with the increased risk of heart diseases. Our experiment states that *decreasing* HDL also increases CVD risk no matter what its immediate value is.
2) A decreasing POCCIG clearly relates to the increasing risk of CVD ($p = 0.0001$). It states that the rate of CVD increases if a patient tends to stop smoking! This observation seems to contradict the domain knowledge.

TABLE II
OVERVIEW OF THE ASSOCIATION RULES FOUND

nr.	rule	confidence	lift	support
b1	$hdlgrad = 1 \wedge bmigrad = 4..5 \rightarrow CVD = 1$	0.69	1.90	0.06
b2	$bmigrad = 5 \wedge chlstgrad = 2..3 \rightarrow CVD = 1$	0.61	1.65	0.05
b3	$diastgrad = 2 \wedge triglgrad = 1..2 \rightarrow CVD = 1$	0.60	1.64	0.05
b4	$triglgrad = 1..2 \wedge hdlgrad = 1..2 \rightarrow CVD = 1$	0.56	1.53	0.08

The last observation suggests that it might be interesting to compare the development of smoking habits in the CVD and the healthy group. Surprisingly, the former one has much higher percentage of those who gradually give up smoking—there are 46% of men who smoke less in the CVD group whereas only 19% of men smoke less in the healthy group. Perhaps, a plausible explanation is that patients stop smoking because their health condition becomes bad, but it is already too late to stop the oncoming disease.

On the other hand, none of the introduced derived attributes is strong enough to point to the fact that CVD is likely to appear. That is why we have decided to concentrate on the search for valid association rules in the considered dataset.

B. Ordinal and Quantitative Association Rule Mining

Traditional algorithms for association rule mining have exhibited an excellent ability to identify interesting association relationships among a set of binary attributes. Although the rules can be relatively easily generalized to other variable types, the generalization can result in a computationally expensive algorithm generating a prohibitive number of redundant rules of little significance. This danger especially applies to quantitative and ordinal variables present in the STULONG dataset. We tried to verify an alternative approach inspired by [19] and [20] to the quantitative and ordinal association rule (OAR) mining. In this approach, quantitative or ordinal variables are not immediately transformed into a set of binary variables. Instead, it applies simple arithmetic operations in order to construct the cedents (antecedents or succedents), and searches for areas of increased association, which are finally decomposed into conjunctions of literals. This scenario outputs rules syntactically equivalent to classical association rules.

A detailed description of the applied OAR algorithm is out of scope of this problem-oriented paper. In most approaches, the quantitative variables have to be discretized and, thus, transformed into the ordinal attributes first. In this paper, the equi-depth discretization has been applied for each variable separately—the individuals are equally distributed among five bins of possibly different width. The resulting ordinal partitions are denoted $1, \ldots, 5$. Physiological tags as "somehow increasing" or "quickly decreasing" can be attached to these partitions; however, they do not necessarily have an analogous mapping in different variables. The partitions can sometimes be interpreted as $1 =$ "quickly decreasing," $2 =$ "somehow decreasing," $3 =$ "steady," $4 =$ "somehow increasing," and $5 =$ "quickly increasing." In other variables, a better mapping can be $1 =$ "steady" and $2 =$ "slowly," $3 =$ "somehow," $4 =$ "quickly," and $5 =$ "extremely increasing."

C. Ordinal and Quantitative Association Rules Found in STULONG Study

In order to search the space of association rules in the STULONG study, we have applied the OAR algorithm as well as the AR mining procedure 4ft-Miner [21]. 4ft-Miner is based on efficient bit string representation of analysed data. When its search is complete, it provides an interface that facilitates the processing of quantitative variables. OAR is a derivative of the algorithm presented in [22].

In general, 4ft-Miner carries out more verifications (in order of several magnitudes), which results in a more time-consuming run and a higher number of more specific rules. OAR is faster, searching for more general rules. On the other hand, it shows to miss a certain portion of rules that can be considered as interesting. The typical example is a rule with two antecedent attributes, where a range of low values of one attribute interacts with a range of high values of the other attribute. For the sake of addition, this interaction cannot be differentiated from, for example, a combination of two neutral values.

The results of OAR are outlined in Table II. The rules concern an influence of multiple trends on CVD development. The strongest rule (nr. b1) can be interpreted as follows. The individuals whose HDL cholesterol level "quickly decreases" and whose weight "somehow or quickly increases" might suffer from the higher CVD risk. The increase in probability of CVD, given the antecedent, is 90%.

This rule represents an interesting and plausible hypothesis. However, it has to be verified later since it deals with insufficient target groups. The given analysis was run with 423 transactions (fewer than 1400 men because some of them do not have enough examinations to fill the window, contain too many missing values in critical RFs, or developed disease other than CVD), out of which 155 transactions were CVD positive (37%). The given rule covers 6% transactions, which makes around 26 individuals. An average group of 26 individuals contains ten prospective diseased patients, the rule with the lift of 1.9 actually addresses 19.

V. RSD

Relational rule learning is typically used in solving classification and prediction tasks. The former research within the STULONG domain [23] has proven that the discovered patterns (and undoubtedly hidden ones too) do not show the strength to reliably distinguish between diseased and healthy individuals *a priori*. The task should rather be defined as subgroup discovery. The input is a population of individuals (middle-aged men) and a property of those individuals we are interested in (a CVD onset), and the output includes population subgroups that are statistically "most interesting": are as large as possible, have the

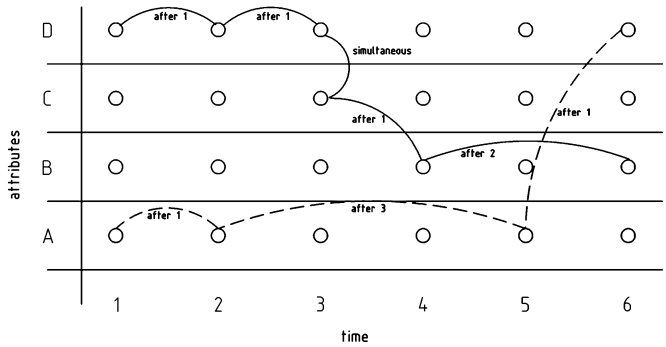

Fig. 6. Two examples of intertransactional sequences.

most unusual statistical (distributional) characteristics with respect to the property of interest, and are sufficiently distinct for detecting most of the target population. The definition of subgroups arises out of the sequential patterns reflecting temporal development of RFs.

Relational rule learning can be adapted also to subgroup discovery—RSD has been devised [24]. It is based on principles that employ the following main ingredients: exhaustive first-order feature construction, elimination of irrelevant features, implementation of a relational rule learner, use of the weighted covering algorithm, and incorporation of example weights into the weighted relative accuracy heuristic.

The whole process can be simplified as follows. The system first tries to construct features, i.e., conjunctions of literals available in the domain. Their essential property is their potential to form subgroups as defined in the previous paragraph. Then, the features are grouped into rules, whose most important characteristic is very similar. The rules only stress the coverage issue, i.e., they try to cover as many target individuals as possible that have not been covered yet (for details, see [24] and [25]).

A. Feasibility, Complexity, Resolution

When mining the STULONG data, the most general and natural approach seems to be to allow arbitrary sequential features. Those features capture sequences of arbitrary length, and they are inherently *intertransactional*, i.e., each sequence may contain *events* from different *risk factors*. The sequences made from events of a single RF are referred to as *single-transactional*. Two examples of such sequences/features that emphasize time relations are shown in Fig. 6. Time relations are modeled by binary predicates $\text{after}_1, \text{after}_2, \ldots, \text{after}_n$—meaning that the second event occurred 1, 2, or n checkups after the first event—and simultaneous—meaning that the events occurred in the same checkup. Of course, there could also have been defined various generalizations of after predicate, e.g., the second event occurred at an arbitrary checkup following the checkup of the first event.

Although the variability of candidate sequences is desirable from the point of view of the final practical knowledge, it can hinder feasibility of the search in the sequence space. The number of generated features can become exceedingly high and, thus, unable to generate the rules in reasonable time. Suppose we have a number of attributes a, a number of values of each attribute v, and a length of a sequence l. Then, the amount of possible single-transactional sequences is $O(n_s) = av^l$ bounded, while the number of intertransactional sequences is $O(n_i) = (av)^l$. The amount of sequences grows exponentially with the maximal allowed sequence length. Computation is even more cumbersome when considering features. As exemplified, the feature length multiplication exceeds the sequence length, while computational burden grows exponentially with the maximal allowed feature length again. The sequence consists of events, the feature consists of predicates, each event corresponds to more predicates, and the events have to be mutually organized (the predicates after). In some sense, the feature space exceeds the originally intended sequence space since the system cannot distinguish between meaningful and pointless features (corresponding to no sequence).[1]

Therefore, the feature and, consequently, the sequence length as well as the number of values being distinguished as events and the number of RFs have to be limited. As a result, sequences that are long and consist of many attributes with many different values cannot be generated.

Searching for intertransactional sequences is computationally demanding. Let us estimate the number of candidate sequences in STULONG. The number of checkups varies from 1 to 21, around 80% of men were measured for five and more times—it seems reasonable to allow for the sequences limited by five events. The group of the most significant variables consists of five RFs: SYST, DIAST, CHLSTMG, TRIGLMG, and BMI; there were tens of different values measured. Obviously, there are tens of billions of candidate sequences.

Consequently, the number of attributes has to be cut down, which causes loss of the information about the relationships between attributes. At the same time, the length of sequences needs to be cut down, which reduces resolution in the time domain. The number of possible attribute values has to be lowered (a reasonable amount of discretization categories can only be used), which reduces the resolution in the data domain. The intertransactional nature of sequences may, therefore, be seen by some as a problem rather than a feature, but we have to keep in mind that albeit its computational intensity, it is a new way of handling information and, as proposed in [26], new and more effective algorithms of intertransactional rules processing are being developed. We have spent some time trying to find an equilibrium between the number of attributes and the length of sequences, and then we decided to take kind of a "third way" and divided the data to three disjunct windows as described in Section V-B.

B. Data Preprocessing: The Final Setup

One of the first tasks we have to cope with in order to use RSD effectively is preprocessing. RSD loads and interprets language declarations and data in a predicate logic format [25].

The main preprocessing tasks are: 1) to generate the Prolog code for feature template construction; 2) to carry out

[1]RSD by no means generates arbitrary features, i.e., arbitrary conjunctions of literals. The feature space is implicitly reduced as every variable has to be used as the input variable at least once, features cannot be decomposable, predicates can be defined as antisymmetric, etc. The real computation also depends on background knowledge design that can introduce high-level predicates further reducing the feature space.

attribute discretization; and 3) to perform trend construction. While the first task is necessary formatting, the other two tasks address effectiveness. Immediate utilization of Prolog predicates for preprocessing turned out to be quite ineffective, because an extra predicate is needed for each discretization made, which effectively doubles the length of the features and decreases computational effectiveness of feature generation. Thus, the features were discretized in advance in terms of preprocessing. The following discretized attributes were generated: NORMBMI, NORMSYST (NORMDIAST), NORMCHLSTMG, NORMTRIGLMG. All these attributes are derived from appropriate STULONG risk factors BMI, SYST, DIAST, CHLSTMG, and TRIGLMG mentioned earlier. The discretized attributes were transformed from the original attributes by equidistant discretization into three intervals referred to as "low," "medium," and "high."[2]

Another way that simplifies feature construction and makes it more effective is introduction of short-time trends. The attributes TRENDBMI, TRENDSYST (TRENDDIAST), TRENDCHLSTMG, and TRENDTRIGLMG represent transformations of original sequences, which are reflecting the speed of change of the attribute value in time. Possible values of the "trend" attributes are "down2," "down," "flat," "up," and "up2," meaning "big decrease," "decrease," "no change," "increase," and "big increase" of the attribute value, respectively.

Proposed preprocessing reasonably reduces the feature length while preserving the complexity of the underlying sequence. To finish the final setup, proximity of CVD onset has to be also quantified. The target (class) attribute CVD is a binary attribute signalling the presence of a CVD at the end sequence corresponding to the given individual (0: nondiseased, 1: diseased). The length of the original sequences varies from 1 to 21 checkups, the average is around 8. The individual sequences (SYST, BMI, etc.) were divided into three disjunct windows called *begin, middle, end*, where *end* covers last four events, *middle* covers another four *events* before the *end*, and *begin* covers the rest—all the events from the beginning of the sequence to the middle window. Each generated feature is located in one of those windows, and it may contain one sequence of a maximum length of 2. The time predicates after_i were replaced by the binary predicates after_beg, after_mid, and after_end defining that the second event occurred at an arbitrary time after the first event, and both the events are located in the same window (*beg* stands for the beginning window etc.).

When combining features into *rules*, each *rule* consists of a maximum of three *features*. This restriction allows to describe a sequence up to the length of 6. These constraints may, of course, vary in future applications, but the principle will be essentially the same. Examples of final rules can be seen in the following section.

[2]There are many alternate ways to discretize—a finer partitioning, equidepth discretization, or local approaches defining interval boundaries for every single patient separately could have also been applied.

C. Results

In this section, selected generated rules and their interpretations are presented. Let us take a look at the following rule:

$$\texttt{class} : 0, \texttt{conf} : 0.968, \texttt{cov} : 0.156, \texttt{lift} : 1.308$$

$$\texttt{f}\ (7369,\ \texttt{A}) : -\texttt{checkup}(\texttt{A},\ \texttt{B}),\ \texttt{normsyst}\ (\texttt{B},\ \texttt{medium}),$$

$$\texttt{trendbmi}(\texttt{B}, \texttt{flat}), \texttt{trendsyst}(\texttt{B}, \texttt{up}).$$

$$\texttt{f}(3068, \texttt{A}) : -\texttt{checkup}(\texttt{A}, \texttt{B}), \texttt{checkup}(\texttt{A}, \texttt{C}),$$

$$\texttt{after\_mid}(\texttt{C}, \texttt{B}), \texttt{trendbmi}(\texttt{C}, \texttt{flat}).$$

$$\texttt{f}(1158, \texttt{A}) : -\texttt{checkup}(\texttt{A}, \texttt{B}), \texttt{checkup}(\texttt{A}, \texttt{C}),$$

$$\texttt{after\_beg}(\texttt{C}, \texttt{B}), \texttt{normtriglmg}(\texttt{B}, \texttt{low}),$$

$$\texttt{trendtriglmg}(\texttt{C}, \texttt{up2}).$$

The rules have the same form as the classical decision rules Cond $\Rightarrow$ Class, where Cond (premise) is "object satisfies all the listed features" and Class (result) is "object is assigned the listed class" [for the particular rule: IF f(7369,A) AND f(3068,A) AND f(1158,A) THEN class(A,0)]. However, the rules are not used to classify the individuals but to distinguish interesting subgroups. Thus, they can also or rather better be viewed and treated as association rules Ant $\Rightarrow$ Suc. As a matter of fact, classical association rule characteristics serve the purpose of their evaluation—they can be viewed at the first row of the rule. Class 0 suggests that the rule concerns nondiseased individuals. Coverage $\text{cov} = n(\text{Ant})/n$, where $n(\text{Ant})$ is the number of instances covered by the rule's antecedent, and n is the number of all patients. Coverage is the fraction of patients covered by the rule. Rules with low coverage (5% or less) are usually considered useless. Confidence $\text{conf} = n(\text{Ant} \cap \text{Suc})/n(\text{Ant})$ is the accuracy of the rule. It expresses the number of instances that satisfy the premise as well as the result. Lift is defined as $\text{lift} = \text{conf}/p_a$, where $p_a = n(\text{Suc})/n$ is the prior probability of the rule's class. It conveys how much better is the rule's performance compared to a trivial classifier, which assigns all instances into one class, and its performance is the same as the prior class probability.

The remaining rows present the antecedent, i.e., three features, which have to be satisfied simultaneously. The meaning of the first feature is that the patient had an examination, in which he had medium SYST, a steady trend of BMI, and a rising trend of SYST. The meaning of the second feature is that the patient had two examinations (B and C) in the middle of the time sequence, and examination C happened before examination B. In examination C, he had a steady trend of BMI. The third feature is that the patient had two examinations (B and C) in the beginning of the time sequence, and examination C happened before examination B. In examination C, he had a steeply rising trend of TRIGLMG, and in examination B, he had a low level of TRIGLMG. To summarize all the three features: Our patient, long time ago, had a steep rise of TRIGLMG followed by a low level of TRIGLMG. Short time ago, he had a steady trend of BMI. At any time in history, he had a medium level of SYST, steady trend of BMI, and a rising trend of SYST. The patient who satisfies these conditions has 30.8% more chance of not having a CVD than the average. The percentage is taken

from the lift characteristics, $p = (\text{lift} - 1) \times 100\%$. Let us take a look at another rule:

$\texttt{class} : 1, \texttt{conf} : 0.615, \texttt{cov} : 0.049, \texttt{lift} : 2.367$

$\texttt{f}(4380, \texttt{A}) : -\texttt{checkup}(\texttt{A}, \texttt{B}), \texttt{checkup}(\texttt{A}, \texttt{C}),$

$\texttt{after\_end}(\texttt{C}, \texttt{B}), \texttt{normsyst}(\texttt{B}, \texttt{high}), \texttt{trendbmi}(\texttt{C}, \texttt{flat}).$

$\texttt{f}(4124, \texttt{A}) : -\texttt{checkup}(\texttt{A}, \texttt{B}), \texttt{checkup}(\texttt{A}, \texttt{C}), \texttt{after\_end}(\texttt{C}, \texttt{B}),$

$\texttt{normbmi}(\texttt{B}, \texttt{medium}), \texttt{trendchlstmg}(\texttt{C}, \texttt{up2}).$

$\texttt{f}(4439, \texttt{A}) : -\texttt{checkup}(\texttt{A}, \texttt{B}), \texttt{checkup}(\texttt{A}, \texttt{C}),$

$\texttt{after\_end}(\texttt{C}, \texttt{B}), \texttt{normsyst}(\texttt{B}, \texttt{high}), \texttt{trendchlstmg}(\texttt{C}, \texttt{up2}).$

This rule has a very good lift, but its coverage is on the edge of usefulness. So the rule is very strong, but valid only for a small fraction of instances. All the events are happening at the end of the sequence, very short time before the CVD was found. This patient had a flat trend of BMI followed by high SYST, steeply rising trend of cholesterol level followed by medium level of BMI and high SYST. To summarize a bit again, those features mean that the patient had normal BMI with steady trend, and after that he had a steep rise of cholesterol level followed by high SYST. The patients who satisfy these conditions have 137% more risk of CVD than the average.

RSD can also be utilized for nonsequential data. In such a case, the application is still more straightforward resulting in rules as follows:

$\texttt{class} : 0, \texttt{conf} : 0.910, \texttt{cov} : 0.084, \texttt{lift} : 1.230$

$\texttt{f}(9745, \texttt{A}) : -\texttt{liquors}(\texttt{A}, \texttt{none}).$

$\texttt{f}(9737, \texttt{A}) : -\texttt{beer}(\texttt{A}, \texttt{more\_than\_1\_liter}).$

This rule means that strong beer drinkers who do not drink liquors are 23% less likely to have a CVD. Sequential and nonsequential predicates/features can be naturally combined as demonstrated in the following rule:

$\texttt{class} : 1 \texttt{conf} : 0.568, \texttt{cov} : 0.055, \texttt{lift} : 2.185$

$\texttt{f}(9738, \texttt{A}) : -\texttt{beer}(\texttt{A}, \texttt{occasionally}).$

$\texttt{f}(8453, \texttt{A}) : -\texttt{checkup}(\texttt{A}, \texttt{B}), \texttt{normchlstmg}(\texttt{B}, \texttt{medium}),$

$\texttt{trendchlstmg}(\texttt{B}, \texttt{flat}).$

$\texttt{f}(3787, \texttt{A}) : -\texttt{checkup}(\texttt{A}, \texttt{B}), \texttt{checkup}(\texttt{A}, \texttt{C}), \texttt{after\_mid}(\texttt{C}, \texttt{B}),$

$\texttt{trendtriglmg}(\texttt{B}, \texttt{down2}), \texttt{trendtriglmg}(\texttt{C}, \texttt{flat}).$

The rule can be interpreted such that occasional beer drinkers with a normal cholesterol level with a steep drop of TRIGLMG level in blood have a 118% more chance of developing CVD. Of course, the coverage characteristic has to be considered again. When putting these two rules together, one might infer that a good prevention of CVD is to stop drinking liquors, and interestingly, it also helps to drink beer, except for people with dropping TRIGLMG in blood.

D. Discussion

The generated rules are able to describe detailed interconnections between attributes in time, and are quite immune to errors coming from having too many different sequences because of minor changes in attribute values or time placement. On the other hand, the proposed method is not effective when used on systems, where those minor changes may have a major influence on the property of interest. The time axis is abruptly split while physiological nature of the modeled phenomenon ask for smooth treatment. The method is also better suited for finding local patterns than global models. When used directly for classification (i.e., intended for prediction), its performance is the same or worse than standard learning techniques (i.e., decision table, J48 decision tree, Bayesian network, etc.).

On the other hand, the proposed relational method is able to find patterns, which might be omitted by standard association rule learning algorithms and systems for mining nonintertransactional episode rules from sequential data. The method can be considered as fully general, though its real performance is highly dependent on the data mining goal, the nature of the dataset, and subsequent design of preprocessing and/or background knowledge.

VI. Episode Rule Mining

Episode rule mining, the general tool WinMiner and its application to STULONG data were described in [7] and [11]. Informally, an episode is a sequential pattern composed of events. The standard episode mining problem is then to find all the episodes satisfying a given minimal support constraint. The way the support is established depends on the dataset type. In STULONG dataset dealing with distinct patients and, thus, representing a base of sequences, the support of an episode corresponds to the number of patient sequences in which this pattern occurs at least once. Episode rules reflect how often and how strong a particular group of events tends to appear close to another event group. WinMiner checks all the episode rules that satisfy frequency and confidence thresholds, and outputs those episode rules for which there exists a time interval maximizing the confidence measure [first local maximum (FLM) rules].

The main preprocessing task in episode rule mining is to establish a reasonable set of events. The individual events have also to be made distinct—the events are coded in such a way that each value and type of the original measurement represents a single event. A proper discretization has to be proposed in order to deal with a representative set of events; intertransactional patterns can be searched for if and only if the event types are distinguished also for distinct RFs. During runtime, the total number of events has to be limited in order to cope with the exponential growth of the search space. The maximum time gap between the first and the last prospective event also influences the task complexity. As WinMiner mines a single event sequence only, the time scale of the individual patient sequences has to be coded in such a way that the events of different patients can never be considered inside a single episode (the time span between the patients is larger than the maximum allowed time gap between events).

Six FLM rules relating to different RFs were reported in [11]. The authors did not target CVD solely, and none of the rules concerned them. We have redesigned the preprocessing mentioned

in the previous paragraph in order to be better able to compare the reached results with the methods mentioned earlier. The main difference lies in the introduction of the general CVD event such as defined in Section II. It results in stronger support of the event that represents the onset of a disease and helps to find meaningful episode rules considering these diseases. We have also conducted several distinct experiments with various limited sets of RFs that are likely to interact [27]. This limitation enables to search for longer episode rules. There were tens of FLM rules dealing with CVD found for various definitions of the events. We present here four rules, the first two of them can be considered as expected/typical while the rest were classified as surprising/interesting (by the domain expert). For each rule, the first line gives an original WinMiner output. The episode rule consists of the prefix (presumption) made of more events (separated by dashes) and the suffix (consequence) having a single event. The prefix and suffix are separated by a star. The rules are given with their window length (w), confidence (cw), and support (sw), as well as the verbal interpretation.

CHPHACT − CHSMOKE − PHACT_high ∗ nonCVD

w = 79 : cw = 0.86 : sw = 153

The men who change their physical activity to be classified as high at least once after the change and also change their smoking habits remain healthy for 79 months with 86% probability.

TRIGL_grows-CHLST_dec ∗ CVD : w = 46 : cw = 0.38 : sw = 55

The men whose triglyceride level remarkably grows and their cholesterol level decreases tend to develop CVD with 38% probability.

ANAMN-DIAST_grows ∗ nonCVD : w = 52 : cw = 0.80 : sw = 92

The men with positive family anamnesis and increased diastolic blood pressure tend to remain healthy with probability higher than 80% at least for 52 months.

CIGNUM_dec − TRIGL_grows ∗ CVD : w = 48 : cw = 0.37 : sw = 58

The men who remarkably cut down smoking (at least by ten cigarettes per day) but their triglyceride level remarkably grows (at least by 180 mg) tend to develop CVD with 37% probability.

The last rule is in line with the earlier observations that cutting down smoking is rather an effect of failing health than a future cause of its improvement (see Section IV-A). The other RF interactions confirm a real positive effect of stopping smoking: The men who decrease smoking at least twice in three consecutive checkups and then keep it constant (presumably zero) make their blood pressure steady in 11 years.

The detailed overview of the identified episode rules can be found in [27]. To sum up, the optimal window size delivered with FLM rules is valuable additional information. Moreover, the notion of the FLM reasonably reduces the number of rules (thousands of rules satisfying confidence and support thresholds versus tens of final FLM rules). WinMiner has five critical parameters whose setting critically influences quantity and quality of the output rules. Their setting may prove time consuming but enables a certain degree of focus. In this particular domain, the length of episodes is strongly limited and cannot be more than 5 for very limited sets of two RFs. If a target event such as CVD is the only interest, postprocessing is needed.

VII. Conclusion

Let us mutually compare the presented methods. Windowing is a simple and often used method to transform sequential data. The sequence of data can either be decomposed into several disjunctive windows, or a sliding window approach can be applied. In both cases, the windows are subsequently replaced by aggregate attributes (linear trends are mostly used) and analyzed by traditional AVL. Although this method brought very good results in our study, it has to be tailored to the analyzed domain. Questions such as "What is the optimal window length?" or "Is the linearization a proper generalization of the prospective patterns?" have always to be considered.

Episode rule mining and inductive logic programming represent general solutions to sequential pattern mining problem. Both the methods ask for a certain amount of data preprocessing and tuning, which vastly influences the final outcome. In principle, RSD allows the user to better prespecify the searched patterns via background knowledge and, thus, restrict the searched space, which may, on the other hand, turn out to be time consuming and ask for expert knowledge in relational learning. Both the methods allow to search for more general patterns than linear trends used in windowing. This arbitrariness is repaid by limitations related to the length of the prospective patterns and their complexity regarding the number of different RFs. To sum up, sequential mining in nontrivial data faces a problem of combinatorial explosion. The presented methods differ in a way to cope with it, representing a broad diversity of tactics. Windowing very early focuses on very specific patterns. The search space is pruned, and long-term or multivariable (intertransactional) patterns can be found. Redefinition often equals complete readjustment of the process. ILP gives the user great flexibility in task definition and reformulation. As soon as the user introduces the logical predicates corresponding to the basic building blocks of patterns, the task can be quickly redefined for very different constraints. Episode rule mining deals with free set of patterns and limited set of parameters only. The approach is general and user friendly, but the user often has to confine himself to patterns with a short gap or a restricted event set.

Regarding meaning and practical importance of the concrete sequential patterns, none of the methods clearly outperforms the others.

The presented paper summarizes long-term research carried out in close cooperation with physicians. Their guidance helped to focus on particularly important or interesting RFs and their

interactions. The medical expertise of the great majority of the generated patterns was given, and it often inspired the subsequent progress. Even though soundness of the complex sequential patterns cannot be immediately statistically confirmed, they became an integral part of the STULONG study evaluation. The study had already been closed before its mining started. That is why the patterns and their expertise could not have reciprocally influenced the study parameters or duration, which would have made an ideal feedback cycle.

There are still open questions to handle. An interesting issue is related to the discovery of the higher order dependencies for individual patients as opposed to those averaged ones visualized in Fig. 5. There are also questions that cannot be answered because they lack the necessary data support. Although the study has a large extent (more than thousand of patients followed for 20 years), the main strength of sequential mining methods in identification of truly complex long-term patterns dealing with two and more RFs cannot be fully exploited. For example, it is definitely interesting to combine the trends with the absolute values. In particular, knowing that the long-term increase of BMI increases CVD risk, it is likely that an increasing BMI is more risky for a fat person than for a slim one. Out of more than 1400 men, we can identify only 27 who developed both CVD and obesity. This group is not large enough to distinguish the influence of BMIGrad. This example indicates that multivariate analysis, which is fully in the scope of the presented methodology, may have proven a higher influence of the defined trends in a larger study. It also illustrates that a general exploratory study, which does not aim only at a few specific questions since the beginning, must have been extensive in order to search for complex and *a priori* unanticipated sequential patterns. Likewise, the consecutive sequential analysis of such a large dataset could not have been applied blindly—the queries such as "give me all the interesting episode rules up to length of 10 must fail. The analysis needs to be gradual, employ a wider scale of methods and has to take place in close cooperation with the expert.

ACKNOWLEDGMENT

The STULONG study has been conducted at the 2nd Department of Internal Medicine, 1st Faculty of Medicine of Charles University, and University Hospital in Prague (Head Prof. M. Aschermann), under the supervision of Prof. F. Boudik, with the collaboration of M. Tomečková, and J. Bultas. The collected data were transferred to the electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Czech Academy of Sciences (Head Prof. J. Zvárová). The authors thank M. Tomečková for her systematic assistance and medical expertise of the generated patterns.

REFERENCES

[1] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. 11th Int. Conf. Data Eng.*, 1995, pp. 3–14.

[2] T. G. Dietterich, "Machine learning for sequential data: A review," in *Proc. Joint IAPR Int. Workshop Struct., Syntactic, Statist. Pattern Recog.*, 2002, pp. 15–30.

[3] N. Lesh, M. J. Zaki, and M. Ogihara, "Scalable feature mining for sequential data," *IEEE Intell. Syst.*, vol. 15, no. 2, pp. 48–56, Mar./Apr. 2000.

[4] H. Mannila, H. Toivonen, and I. Verkamo, "Discovering frequent episodes in sequences," in *Proc. 1st Int. Conf. KDD 1995*, pp. 210–215.

[5] H. Mannila and H. Toivonen, "Discovery of generalized episodes using minimal occurrences," in *Proc. 2nd Int. Conf. KDD 1996*, pp. 146–151.

[6] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of frequent episodes in event sequences," *Data Min. Knowl. Discovery*, vol. 1, no. 3, pp. 259–298, 1997.

[7] N. Meger and C. Rigotti, "Constraint-based mining of episode rules and optimal window sizes," in *Proc. 8th Eur. Conf. Princ. Pract. Knowl. Discovery Databases*, New York, 2004, pp. 313–324.

[8] R. Ichise and M. Numao, "First-order rule mining by using graphs created from temporal medical data," in *Lecture Notes in Artificial Intelligence,* Berlin, Germany: Springer-Verlag, vol. 3430, pp. 115–128, 2005.

[9] M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi, "A rule discovery support system for sequential medical data in the case study of a chronic hepatitis dataset," in *Proc. ECML/PKDD-2003 Discovery Challenge Workshop*, Croatia, pp. 154–165.

[10] S. Hirano and S. Tsumoto, "Mining similar temporal patterns in long time-series data and its application to medicine," in *IEEE Int. Conf. Data Min.*, 2002, pp. 219–226.

[11] N. Meger, C. Leschi, and C. Rigotti, "Mining episode rules in STULONG dataset," in *Proc ECML/PKDD 2004 Discovery Challenge—Collaborat. Effort Knowl. Discovery*, Pisa, Italy, Sep. 2004, pp. 1–12.

[12] M. Tomeckova, J. Rauch, and P. Berka, STULONG—Data from a longitudinal study of atherosclerosis risk factors. [Online]. Available: http://lisp.vse.cz/challenge/ecmlpkdd2002/

[13] STULONG study website. (2002). [Online]. Available: http://euromise.vse.cz/stulong-en

[14] R. A. Baxter, G. Williams, and H. He, "Feature selection for temporal health records," in *50th Asia-Pacific Conf. Knowl. Discovery Data Min. Lecture Notes in Computer Science,* Berlin, Germany: Springer-Verlag, vol. 2035, pp. 198–209, 2001.

[15] D. Pyle, *Data Preparation for Data Mining*. San Mateo, CA: Morgan Kaufmann, 1999.

[16] C. J. Lloyd, "Using smoothed operating characteristic curves to summarize and compare diagnostic systems," *J. Amer. Statist. Assoc.*, vol. 93, no. 444, pp. 1356–1364, 1998.

[17] C. M. Antunes and A. L. Oliveira, "Temporal data mining: An overview," in *Proc. Temporal Data Mining Workshop, 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 1–13.

[18] P. Aubrecht and Z. Kouba, "A universal data pre-processing system," in *Proc. DATAKON 2003*, pp. 173–184.

[19] S. Guillaume, "Discovery of ordinal assocional rules," in *Proc. 6th Pacific-Asia Conf. Knowl. Discovery Data Mining*, 2002, pp. 322–327.

[20] S. Guillaume, "Ordinal association rules towards association rules," in *Proc. 5th Int. Conf. Data Warehousing Knowl. Discovery*, 2003, pp. 161–171.

[21] J. Rauch and M. Simunek, "Mining for 4ft Association Rules," in *Proc. 3rd Int. Conf. Discovery Sci.*, 2000, pp. 268–272.

[22] F. Karel, "Quantitative association rules mining," in *Proc. 10th Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, 2006, pp. 195–202.

[23] L. Novakova, J. Klema, M. Jakob, O. Stepankova, and S. Rawles, "Trend analysis and risk identification," in *Proc. Workshop ECML/PKDD 2003*, pp. 95–107.

[24] N. Lavrac, F. Zelezny, and P. Flach, "RSD: Relational subgroup discovery through first-order feature construction," in *Proc. 12th Int. Conf. Inductive Logic Program.*, Jul. 2002, pp. 149–165.

[25] F. Zelezny. RSD User's manual. [Online]. Available: http://labe.felk.cvut.czzelezny/rsd/

[26] F. Guil, A. Bosch, and R. Marin, "TSET: Algorithm for mining frequent temporal patterns," in *Proc. ECML/PKDD Workshop Knowl. Discovery Data Streams—Collaborat. Effort Knowl. Discovery*, pp. 65–74.

[27] V. Bláha, "Ateroskleróza v sekvenčních lékařských datech," Master's thesis, Dept. Cybern., FEE, Czech Tech. Univ., Prague, Czech Republic, 2006.

Jiří Kléma received the Ph.D. degree in artificial intelligence and biocybernetics from the Czech Technical University (CTU), Prague, Czech Republic, in 2002.

He is an Assistant Professor in artificial intelligence with the Department of Cybernetics, CTU. He was a Postdoctocral Fellow with the University of Caen, France. His main research interests include data mining and its applications in industry, medicine, and bioinformatics.

Lenka Nováková is currently a Research Fellow with the Gerstner Laboratory, Czech Technical University, Prague, Czech Republic. Her main research interests include visualization techniques for data mining.

Filip Karel is currently working toward the Ph.D. degree in quantitative associated rule mining at the Department of Cybernetics, Czech Technical University, Prague, Czech Republic.

His main research interests include association rules, medical data mining, and usage of e-learning.

Olga Štěpánková received the Graduate degree in mathematics from Charles University, Prague, Czech Republic, in 1972.

She is the Vice-Head in the Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University, Prague, Czech Republic, where she was appointed a Full Professor in 1999. Her current research interests include problems related to use of formal reasoning methods in artificial intelligence, data mining, and applications of ICT in medicine. She is the author or coauthor of eight books and more than 100 papers published in several journals and conference proceedings.

Filip Železný received the Ph.D. degree in artificial intelligence and biocybernetics from the Czech Technical University (CTU), Prague, Czech Republic, in 2003.

He is the Head of the Intelligent Data Analysis Research Group, Gerstner Laboratory, CTU. He was a Postdoctoral Fellow with the University of Wisconsin, Madison. He was a Visiting Professor with the State University of New York, Binghamton. Currently, he is a grantee of the Czech Academy of Sciences and the European Commission. His main research interests include relational machine learning and its applications in bioinformatics.