Trend Analysis in Stulong Data

Jiří Kléma, Lenka Nováková, Filip Karel, Olga Štěpánková

Gerstner Laboratory, Department of Cybernetics,
Czech Technical University, Technicka 2,
166 27 Prague, Czech Republic
{klema,novakova,step}@labe.felk.cvut.cz

Abstract. The ECML/PKDD data mining challenge concerns a dataset describing the data collected during a longitudinal study of atherosclerosis prevention on around 1400 middle-aged men (Stulong study). The data challenge entry from the Czech Technical University in Prague focuses on trend analysis in this dataset. Firstly, it proposes and verifies a preprocessing method based on windowing. The suggested approach guarantees that the identified trend aggregates are generated without falling into the trap of introducing anachronistic attributes. Secondly, it applies the windowing method to the Stulong dataset. Finally, it studies influence of these trend aggregates on a possible future development of cardiovascular diseases (CVDs).

1 Introduction

The study Stulong [10] is a longitudinal primary preventive study of middle-aged men lasting twenty years. The study contains data resulting from observation of approximately 1400 men, the main intention of the project is to identify risk factors of atherosclerosis.

The study has been running at the 2^{nd} Department of Internal Medicine, 1^{st} Faculty of Medicine of Charles University and University Hospital, Prague 2, Czech Republic (head Prof. M. Aschermann, MD, SDr, FECS), under the supervision of Prof. F. Boudik, MD, SDr, with the collaboration of M. Tomeckova, MD, PhD, and Ass. Prof. J. Bultas, MD, PhD. Collected data was transferred to the electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences, Czech Republic (head Prof. RNDr. J.Zvarova, SDr). The data analysis is supported by the Project Nr LN 00B 107 of the Ministry of Education of the Czech Republic.

The data is inherently multi-relational, consisting of four separate tables. The table Entry describes data collected during the entry examinations of all patients, Control includes results of a series of long-term observations recording the development of risk factors and associated conditions, Letter provides complementary information collected by questionnaire filled-in by all the patients and records about death of some patients appear in the Death table. This paper is concerned with two tables only - Entry and Control.

2 Overview of our last year's challenge entry

Our team participated also in the last year Stulong discovery challenge [5]. In order to gain better understanding of individual attributes and their relation to the occurrence of a CVD, our analysis started with the Entry table. As the first subgoal, we tried to answer the following question (the 6th analytical question in terms of the discovery challenge tasks): *Are there any differences in the entry examination between men of the risk group, who came down with the observed cardiovascular diseases during the studied period and those who stayed healthy?*

Besides well-known risk factors—such as age, smoking, obesity, hypercholesterolemia or hypertension—we also focused on less known, perhaps even surprising dependencies. Probably the most surprising correlation regarded the Alcohol feature group. A significantly negative correlation of beer drinking with respect to cardiovascular diseases was revealed, namely *increased beer consumption correlated with decreased CVD rate*, what is more, regular drinkers consuming over 1 liter of beer a day showed the lowest rate of CVD! The studied data support even more general claim *occurrence of CVD correlates negatively (inversely) with increased alcohol consumption*. Contrary to wide-spread belief, we did not discover any specific influence of wine consumption on the CVD rate, though. Other identified correlations can be seen in [5].

Further we decided to focus on the temporal aspect of the data supplied and utilize the advantage of long-term observations. Our intention was to answer the question: *Is there any difference in the development of risk factors and other characteristics between men of the risk group who came down with the observed cardiovascular diseases and those who stayed healthy?*

Our approach required demanding data preparation. SumatraTT 2.0 ([8], [11]) provided an environment for rapid development of various preprocessing scripts and thus helped to form a set of well-defined data transformations. Considering the temporal examination data, the following aggregates were derived: mean, gradient, intercept, correlation and standard deviation. This set of aggregates was generated for a selected group of the temporal variables and served as a transformation towards attribute-value representation. The intuitive meaning of the individual aggregates is demonstrated in Figure 1.

An important feature of the Control table is that each of the patients appears several times: the value of ControlCount, i.e., the number of checkup examinations of a single patient, ranges from 1 to 20. There are even (about 60) patients in the Entry table, who do not appear in the Control table at all. In other words, multi-relational transformation of both the Entry and Control tables results in a new table, where data about a single patient can appear in upto 21 distinct records (depending on the value of ControlCount).

Regarding the trend aggregates, the key issue seems to be selection of a subset of the examinations they are based on. The most straightforward approach is to use all the available examinations for each patient at once. This method is referred to as the *global* approach. But there is a danger in this approach since the number of all controls can be an anachronistic attribute ([6]): we do not know in advance, how many times the new patient will come. As soon as the number

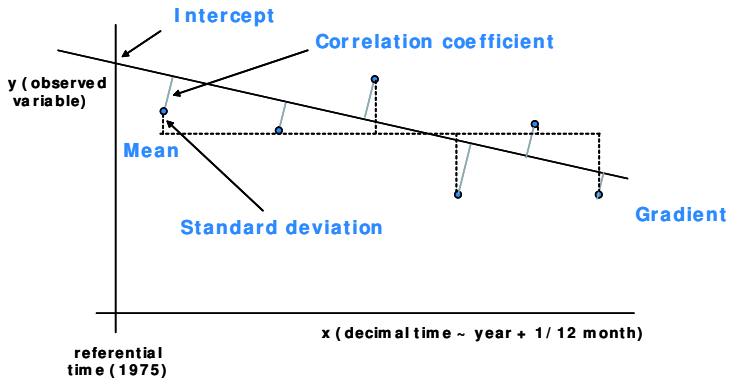

Fig. 1. The set of aggregates generated for selected temporal variables.

of all the patient controls correlates with the overall CVD rate, the global trend aggregates can never be used in the CVD risk analysis.

Our previous paper pointed out that the low number of ControlCount is by far the strongest risk factor in the study. We have argued why the derived trend attributes based on global approach should not be used for answering the question mentioned above and we have proposed an alternative *windowing* approach. It seemed that none of the window based attributes on its own is able to identify any knowledge which could play a role in preventing CVD. The future work was proposed to aim at the further refinement of the windowing approach and the analysis of the original data combined with the new window based aggregates.

3 Windowing

Windowing is a simple and often used method to transform data, see e.g., [1]. Two types of windowing can be distinguished. The sequence of data can be either decomposed into several disjunctive 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 behaviour. 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 sub-series. Both choices lead to a simplified representation of the input sequence, which can result e.g., in a more effective definition of a similarity level for clustering.

In our task we tackle to find a relation between individual risk factors 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 be applied to the time series consisting of at least l measurements, only. The 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 Figure 2.

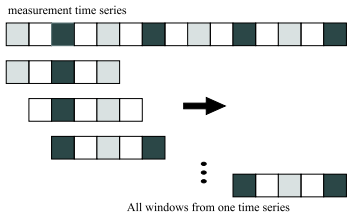

Fig. 2. Windowing a temporal data - a basic sliding window.

3.1 Windowing - new tasks

The windowing method proposed above 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, when one can expect such changes in time series, which allow to predict the result. Consequently, the first distinct task of this paper is to answer the question: *What is the optimal window length in the Stulong study?*

Another important issue applies to missing data. There is no doubt that data about some patient examinations can miss certain values of usually measured variables. Our second task is to propose such a windowing that *treats the missing data in a proper way*, i.e., it always keeps a chance to reference windows (and thus also aggregates) of different variables with the same time tag while losing minimum measurements that are present.

The third topic relates to introduction of a temporal definition of the target CVD attribute. Non-temporal definition of the CVD attribute is trivial. It is a boolean attribute being false for all the patients who develop no coronary disease during the control examinations and vice versa. *When using windowing method, one should introduce a more sophisticated distinction that considers a time shift between the sliding window and the possible CVD onset.*

3.2 Windowing and missing data

The windowing method with the fixed window length forms windows (sub-series) that can be concisely represented by the aggregate attributes introduced in section 2. Two principal approaches can be applied when considering time series with occasional missing values:

– the first one sticks to the fixed number of checkups — missing values are omitted and the next values in the time series are taken into account: the series of values is shifted,

– the second one insists on the fixed length of the time window — in the later case, reasonable substitutes for the missing values have to be found: the replacement approach is applied.

Replacement of the missing value by shifting the series. The direct shift skipping the missing value can cause a severe problem of synchronization. Let us consider an object, which is described by two time series corresponding to development of variables (attributes) A and B, see Figure 3.a.

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 the variable A, because there are no missing values in the 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 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.

Replacement of the missing value by a new value. The other treatment (Fig. 3.b.) recommends to denote the place with the missing value by some symbol and replace this symbol cautiously at the end of transformation process. When we consider attributes developing 'relatively slowly and smoothly', i.e., data exhibiting a certain 'time inertia', the mean calculated from the former and the future value can be a suitable replacement.

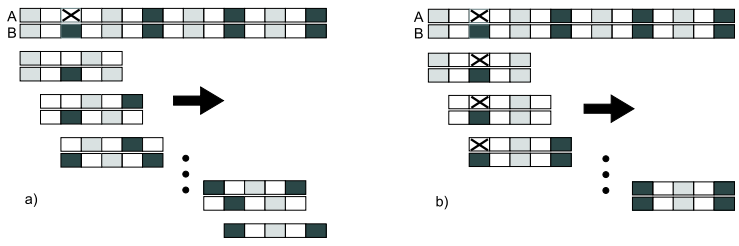

Fig. 3. Replacement of the missing value by (a) shifting and (b) a new value.

3.3 Temporal CVD definition

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 the attribute CVD must be related to the considered time of measurement.

In the last year's study, we have defined the binary attribute CVD-in-3yrs: its value at time t is 1 if the given patient comes down with a cardiovascular disease during next three years (i.e., in the year $t + 1$, $t + 2$ or $t + 3$) and it is 0 otherwise. A more profound approach applied in this paper introduces a derived attribute CVDi the value of which is equal to the distance (in years) from the actual moment to the time when the patient becomes ill (he gets CVD). When the patient remains healthy, the value of the attribute CVDi is set to a distinguished number which is ever interpreted as 'healthy'.

3.4 The optimal window length

Since the number of examinations in our case ranges from 1 to 21, the choice of the window length is a trade-off between the amount of data which have to be omitted and the length of the observable period. For the Stulong study, the window length of 5 measurements has been chosen first. The histogram and the pie chart (Fig. 4) show that taking this option, about one fifth of the patients is rejected due to the fact they have less than five records. If we insist on utilization of longer window of 8 or 10 measurements, the data of nearly a half of the patients would be rejected.

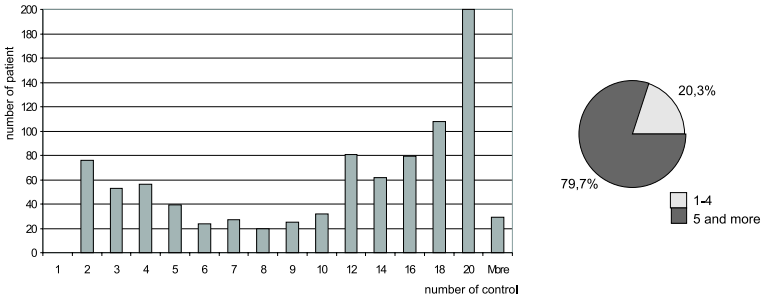

Fig. 4. Histogram of examinations in the Stulong study.

Let us focus on gradients (trends), one of the most interesting and natural aggregate types from the point of view of physicians and patients. The gradients of five important variables have been considered: the systolic and diastolic blood pressure (SYSTGrad, DIASTGrad), cholesterol (CHLSTMGGrad), triglyceride level (TRIGLMGGrad) and BMI (BMIGrad).

The gradients have been calculated for windows of length 5, 8 and 10 examinations (denoted as W5, W8 and W10). The windows have been defined by the constant number of examinations and not by the constant observation period. For a great majority of patients the examinations are taken regularly, e.g., the window of the length 8 mostly corresponds to time period lasting from 7 to 9 years, the influence of different time durations is not studied here. The need for time relativization of the CVD attribute has already been mentioned, see 3.3. For simplicity, a specific modification of the CVDi has been taken, namely the attribute CVD1. CVD1 is 'true' if and only if the cardiovascular disease appears

at the examination directly following the last windowed examination, i.e., the patient develops a cardiovascular disease in one year from the windowed period.

Table 1 shows the level of significance p of the χ^2 test of independence between CVD1 and the gradients defined above (the gradient variables were discretized into 10 distinct equi-depth intervals before testing). In the first column, there are results for the global approach, where the trends are computed from all available measurement. This approach suggests a strong dependency between all the gradients and CVD. When using windowing, only SYSTGrad, DIASTGrad and BMIGrad seem to correlate with CVD1. Moreover, this correlation can be observed for specific window lengths only (SYSTGrad and DIASTGrad when using W5, BMIGrad with W10).

Table 1. Dependencies between CVD and selected gradients, 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**

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

Although this correlation may have physiological explanation for some risk factors, the influence of the anachronistic number of examinations is very clear. The gradients tend to be extreme when the patients have fewer examinations. At the same time, the patients with fewer examinations are exactly those patients for which the suspicion for appearance of a cardiovascular disease is stronger as these patients are always removed from the study before its end (see Section 2).

The results obtained by the windowing suggest weaker influence of the studied gradients. For SYSTGrad and DIASTGrad (and W5) it holds that *the more blood pressure increases the more likely CVD is*. The same kind of dependency was observed for BMIGrad and W10. These dependencies make sense and this confirms that windowing is a reliable way for their retrieval. On the other hand, the experiments show 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 great short time 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 10 examinations at least.

On the contrary, blood pressure can exhibit different type of behaviour 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 cardiovascular disease during the study and for the healthy people - see Figure 5. The figure shows the 10-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 cardiovascular disease. There is a striking difference between both 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 behaviour can be observed at individual patients with CVD as well. But to prove this hypothesis, there have to be defined new aggregated attributes (qualitative shapes, time derivatives of the second order). Obviously, the linear gradients calculated for the longer windows cannot serve for this purpose since they tend to balance between decreasing and increasing trends in the CVD group what prevents them from discrimination between both the groups.

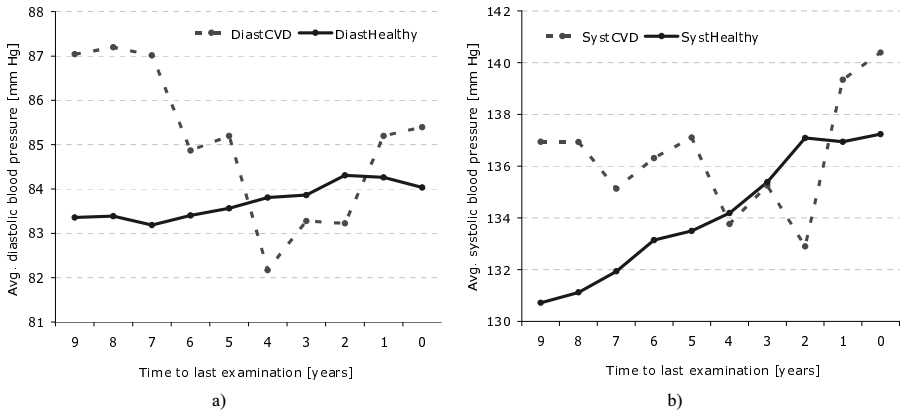

Fig. 5. 10-year course of blood pressures in the CVD and Healthy groups.

Let us notice that multivariate analysis may have proven a higher influence of the defined gradients. It is very likely that increasing BMI is more risky for a fat person than for a slim one which is not considered here. Although the study deals with more than 1400 of men, we can identify only 27 men who came down both with a CVD and the obesity risk. This group is not large enough to study a statistical influence of BMIGrad.

4 Windowed aggregates as risk factors

The next step becomes a search for *possible multivariate interactions among the generated gradient aggregates, interactions among the gradients and other risk factors and finally subgroups with the above-average CVD risk.*

As mentioned in the previous section, multivariate analysis in the Stulong study can often suffer from an insufficient number of patients belonging to potential target groups. The statistical analysis can hardly be carried out. That is why, the best way to point out potential multivariate dependencies seems to be association rule mining. Of course, the identified dependencies have to be understood as "heuristics" to be used by a domain expert.

In the rest of the contribution we stick to *the gradients and means generated using W5.* The other aggregate types are not considered in order to keep a reasonable number of attributes and to minimize random dependencies. The missing values are replaced by adjoining values. It has been verified that in the great majority of the patients the windowed values of different variables truly represent the same period.

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 last but one 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, the gradients are calculated from their 5 central examinations (year 8-12). The examinations are taken in such a way that the average age in both groups is the same. Let us remind that this approach guarantees independence of the considered instances/transactions.

4.1 Newly aggregated variables

In the previous section we have dealt with 5 selected variables only (SYST, DIAST, CHLSTMG, TRIGLMG and BMI). In this section, a couple of new variables has been added from original tables (HDL, LDL, POCCIG and MOC). The χ^2 independence test has been applied to study their relation to CVD. Two gradients proved to be exercise strong influence:

- A decreasing HDL cholesterol level clearly relates to the increasing risk of CVD (p=0.001). As HDL represents a 'good' cholesterol and it is well known that the low HDL level is associated with the increased risk of heart diseases in men. Our experiment states that even decreasing HDL increases CVD risk no matter what its immediate value is.
- A decreasing POCCIG (the average number of cigarettes smoked per day) clearly relates to the increasing risk of CVD (p=0.0001). It says that the rate of CVD increases if a patient tends to stop smoking! It is a confirmation of the statement presented in our earlier entry, however, in contradiction with the domain knowledge. 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.

4.2 Ordinal association rule mining

The main point of our interest in this paper are the derived aggregated attributes such as gradients and means representing quantitative variables. Traditional algorithms for association rule mining work with qualitative variables. To apply them, we have to transform the data correspondingly. The most natural solution of this task relies on an appropriate partitioning of the values of the considered variables. The adjacent partitions can then be combined as necessary into binary attributes [9], [7].

Another approach to mining of such association rules is presented in [2] and [3]. In this approach, the quantitative attributes still have to be discretized, however logical conjunctions of binary attributes are replaced by an addition of ordinal values. As soon as the variables in the rule antecedent and succedent are added, the resulting variables are tested for independence. If they prove to be dependent, the given ordinal association rule holds. The further two specialization steps infer specialized rules that approach the rules derived by the traditional approaches mentioned above.

In both approaches, the quantitative variables have to be discretized and thus transformed into the ordinal attributes first. In this work, the equi-depth discretization has been applied—the individuals are uniformly distributed in each variable among 5 groups. The resulting ordinal partitions are denoted 1..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'.

4.3 Ordinal 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 representative of the 'traditional' approach LispMiner [7]. OAR is a derivative of the algorithm briefly presented in the second paragraph of the previous section. It has been implemented by our team recently, the Stulong study gave a good chance to test it. A detailed description of the algorithm is out of scope of this problem-oriented paper.

In general, LispMiner 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 e.g., a combination of two neutral values.

The results are summarized in Table 2. The first group of rules (denoted a#) concerns the interactions among the individual derived gradient aggregates.

Within this group, there was a great prevalence of the rules joining together either blood pressures (DIASTGrad and SYSTGrad) or cholesterol attributes (HLDGrad, LDLGrad and CHLSTGrad). In our opinion, these rules agree with common sense reasoning and cannot be considered valuable. The table presents 'weaker' but more interesting rules between gradients of different types.

Table 2. Overview of the association rules found

nr.	rule	confidence	lift	support
a1	$chlstgrad = 3 \rightarrow diastgrad = 3..5$	0.73	3.70	0.14
a2	$hmotgrad = 1 \rightarrow diastgrad = 1..2$	0.57	1.42	0.12
a3	$diastgrad = 4 \rightarrow hmotgrad = 4..5$	0.56	1.38	0.11
a4	$chlstgrad = 1 \rightarrow triglgrad = 1$	0.4	2.01	0.08
b1	$hdlgrad = 1 \wedge hmotgrad = 4..5 \rightarrow CVD = 1$	0.69	1.90	0.06
b2	$hmotgrad = 5 \wedge chlstgrad = 2..3 \rightarrow CVD = 1$	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 second group of rules (denoted b#) concerns an influence of multiple gradients on CVD development. The strongest rule (nr. b1) can be interpreted as follows. The individuals whose HDL cholesterol level 'quickly decreases' (more than 0.034 mmol per liter and year) and whose weight 'somehow or quickly increases' (more than 0.13 kg/year) might suffer from the higher CVD risk. The increase in probability of CVD given the antecedent is 90%.

This rule represents an interesting hypothesis which has to be verified later since we deal with insufficient target groups. The given rule covers 6% transactions, which makes 26 individuals, i.e., instead of 10 prospective diseased patients we actually observe 19.

5 Conclusions

The presented windowing methodology is a general technique applicable to any domain with instances represented by an uneven number of observations possibly containing missing values. It also helps to escape the danger of anachronistic attributes. In the frame of the Stulong study, combination of the windowing approach, statistics and association rule mining drew attention to particular dependencies, namely:

- There were found relations between the gradients SYSTGrad, DIASTGrad, BMIGrad and CVD, which are highlighted in Table 1. The relations can be identified only for certain window lengths given by the physiological nature of the observed variables.
- Decreasing level of HDL cholesterol as well as decreasing level of POCCIG is related to increasing risk of CVD, see Section 4.1.

– The multivariate association rules based on the gradient attributes called attention to further possible dependencies reviewed in Table 2. These dependencies have to be verified by a domain expert.

The choice of optimal window length is influenced by number of factors, e.g., the physiological nature of the observed events. Of course, it can vary for different variables. When analyzing the data created by windowing, specific development patters proved to be important, e.g., the trend pattern which can be characterized as 'down-up' (see Fig. 5). We believe that gradients of the second order can bring a unified view of the curve shape of the observed variable. We plan to design some additional derived attributes for this purpose.

Due to low support it makes little sense to mine the full set of aggregate attributes 'blindly'. However, there remain a few derived attributes that could possibly play an important role, e.g., SYST-DIAST and its gradient.

Acknowledgments

This research work was supported within the Transdisciplinary Biomedical Engineering Research Programme MSM 210000012 funded by the Czech Ministry of Education.

References

1. Antunes C. M., Oliveira A. L.: Temporal Data Mining: An Overview. Workshop on Temporal Data Mining, 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'01), 2001.
2. Guillaume S.: Discovery of Ordinal Assocional Rules. In Proc. 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'02), pp. 322–327, Taipei, Taiwan, 2002.
3. Guillaume S.: Ordinal Association Rules towards Association Rules. In Proc. 5th Int. Conf. Data Warehousing and Knowledge Discovery (DaWaK'03), Prague, pp. 161–171, Czech Republic, 2003.
4. Miller R.J., Yang Y.: Association Rules over Interval Data. In Proc. ACM-SIGMOD Int. Conf. Management of Data, pp. 452–461, Tuscon, AZ, 1997.
5. Novakova L., Klema J., Jakob M., Stepankova O., Rawles S.: Trend Analysis and Risk Identification. In Workshop Proc. and Tutorial Notes ECML/PKDD 2003 [CD-ROM]. Stuttgart: IRB Verlag, pp. 95–107, 2003.
6. Pyle D.: Data Preparation for Data Mining, Morgan Kaufmann, California, 1999.
7. Rauch J., Simunek M.: Mining for 4ft Association Rules. In Discovery Science, Springer Verlag 2000. Eds. Arikawa S., Morishita S., pp. 268–272, 2000.
8. Stepankova O., Aubrecht P., Kouba Z., Miksovsky P.: Preprocessing for Data Mining and Decision Support. In Data Mining and Decision Support: Integration and Collaboration. Dordrecht: Kluwer Academic Publishers, pp. 107–117, 2003.
9. Srikant R., Agrawal R.: Mining Quantitative Association Rules in Large Relational Tables. In Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data, pp. 452–461, Montreal, Canada, 1996.
10. Stulong study, www page: http://euromise.vse.cz/stulong.
11. SumatraTT homepage: http://krizik.felk.cvut.cz/sumatra.