Computing and Informatics, Vol. 26, 2007, 329–344

LESSONS LEARNED FROM THE ECML/PKDD DISCOVERY CHALLENGE ON THE ATHEROSCLEROSIS RISK FACTORS DATA

Petr Berka, Jan Rauch

Department of Information and Knowledge Engineering
Faculty of Informatics and Statistics
University of Economics
W. Churchill Sq. 4
130 67 Prague, Czech Republic
&
Centre of Biomedical Informatics
Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
e-mail: {berka, rauch}@vse.cz

Marie Tomečková

Centre of Biomedical Informatics
Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
e-mail: tomeckova@euromise.cz

Revised manuscript received 8 December 2006

Abstract. It becomes a good habit to organize a data mining cup, a competition or a challenge at machine learning or data mining conferences. The main idea of the Discovery Challenge organized at the European Conferences on Principles and Practice of Knowledge Discovery in Databases since 1999 was to encourage a collaborative research effort rather than a competition between data miners. Different

data sets have been used for the Discovery Challenge workshops during the seven years. The paper summarizes our experience gained when organizing and evaluating the Discovery Challenge on the atherosclerosis risk factor data.

Keywords: Atherosclerosis risk, data mining, discovery challenge

1 INTRODUCTION

It becomes a good habit to organize a data mining cup, a competition or a challenge at machine learning or data mining conferences. Such events serve several purposes: they can be used for comparison of various approaches and algorithms, they give the participants a possibility to access and analyze real-world data, and they can result in a knowledge interesting for the domain experts who provided the data.

Cups and competitions are usually organized around a clearly specified classification problem. The participants are provided with pre-classified training data and a set of examples to be classified. The goal is to build a model that will perform well on the evaluation data. The models are then ranked according to their performance and the winners (sometimes also the losers) are announced. Thus the first purpose is stressed. Let us mention here e.g. the COIL2000 competition, the EUNITE 2001 or 2002 competition and, of course, the KDD cups held since 1997. Challenges have a less competitive nature. The aim here is to prepare conditions of a real/realistic data mining problem (classification or description) and to find a solution. The results are then discussed with the domain experts. This kind of events is organized e.g. at the European or Pacific-Asian KDD conferences.

2 ECML/PKDD DISCOVERY CHALLENGES

The main idea of the Discovery Challenges organized at the European Conferences on Principles and Practice of Knowledge Discovery in Databases since 1999 was to encourage a collaborative research effort, a broad and unified view of knowledge and methods of discovery, and emphasis on business problems and solutions to those problems. During the seven Discovery Challenges we organized so far, different data sets from business (banking, e-commerce) as well as from medicine(thrombosis, atherosclerosis, hepatitis, gene expressions) have been used. See Table 1 for a summary of number of submissions related to different used data sets – this table clearly shows the growing interest in Discovery Challenges. Although the data came from very different domains, they shared some common features. The participants were faced with multi-relational problem with a mixture of static data (characteristics of clients or patients) and dynamic data (transactions or laboratory tests and examinations).

During a single year, the participants could choose any of the available datasets, formulate an interesting problem (from the users point of view) and perform corre-

year	location	data(no. papers)	sum
1999	Prague	Financial7), Thrombosis(3)	10
2000	Lyon	Financial(3), Thrombosis(2)	5
2001	Freiburg	Thrombosis(5)	5
2002	Helsinki	Atherosclerosis(5), Hepatitis(5)	10
2003	Dubrovnik	Atherosclerosis(9), Hepatitis(3)	12
2004	Pisa	Atherosclerosis(11), Hepatitis(5), Genes (2)	18
2005	Porto	Hepatitis(5), Genes(5), Click-stream(7)	17

Table 1. Summary of Discovery Challenge submissions

sponding analysis. The main idea of the Challenge was to follow as much closely as possible a real KDD process. An ideal contribution thus included

- the proposed business objectives (goals that may be of interest to database users),
- a brief summary of data mining effort; this summary may include the data preprocessing tasks like data extraction, sampling, data integration and homogenization, data cleaning, data transformation, the data mining step as well as the evaluation criteria approved,
- presentation of the discovered knowledge, and
- an explanation for database users how they can apply the discovered knowledge.

Our aim was to follow as closely as possible a real KDD process, nevertheless the Challenge conditions differ from conditions of a real KDD project in two main points: (1) the time for analysis was rather short (about two or three months) and (2) the participants have only indirect (if any) possibility to communicate with domain experts during the analyzes.

The rest of the paper summarizes our experience gained when organizing and evaluating the ECML/PKDD Discovery Challenge on atherosclerosis risk factor data. In our study we will follow the widely adopted CRISP-DM methodology, that defines a data mining process as shown in Figure 1. CRISP-DM addresses the needs of all levels of users in deploying data mining technology to solve real-world problems [9]. Its aim is to define and validate a data mining process that is generally applicable in diverse application areas.

3 THE ATHEROSCLEROSIS DOMAIN

Atherosclerosis is a total complicated disease of the vessels in all organisms. It is a dynamic process that begins in childhood and adolescence and continues for the whole life. The experts' opinions on the origin and progress of the disease are developing. Interaction and influence of genetic predisposition and exterior environment as well as of so-called risk factors is considered. On the other hand there are some so-called protective factors.

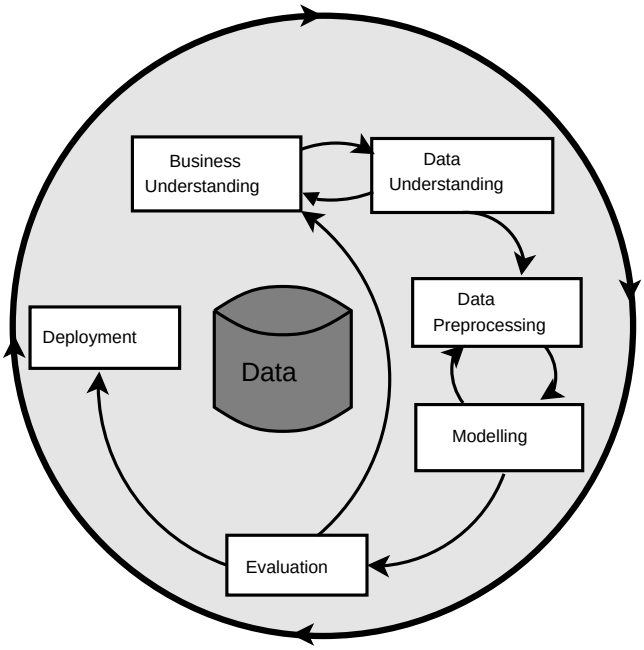

Fig. 1. Structure of the CRISP-DM process

Among the non-affectable risk factors, sex, age and family history are crucial. The affectable risk factors are factors of life style (e.g. physical activity, smoking, reaction on stress), blood pressure, metabolic factors (level of lipids and glucose), and many more (coagulopathies, infections, inflammation, factors changing the function of endothelium, social and psychological factors).

In the early seventies of the twentieth century, a project of extensive epidemiological study of atherosclerosis primary prevention was developed under the name National Preventive Multi-factor Study of Heart Attacks and Strokes in the former Czechoslovakia. The aims of the study were:

1. Identify atherosclerosis risk factors prevalence in a population generally considered to be the most endangered by possible atherosclerosis complications, i.e. middle aged men.
2. Follow the development of these risk factors and their impact on the examined men's health, especially with respect to atherosclerotic cardiovascular diseases.
3. Study the impact of complex risk factors intervention on their development and cardiovascular morbidity and mortality.
4. 10–12 years into the study, compare risk factors profile and health of the selected men, who originally did not show any atherosclerosis risk factors with a group of men showing risk factors from the beginning of the study.

The following risk factors were defined at the beginning of the study: arterial hypertension (BP $\geq$ 160/95 mm Hg), cholesterol (level $\geq$ 260 mg %), triglycerides (level $\geq$ 200 mg %), smoking ($\geq$ 15 cig./day), overweight (Brocka index $>$ 115 %), positive family case history. Later, further laboratory examinations were included: blood sugar level, HDL cholesterol, LDL cholesterol and uric acid.

The study included data of more than 1 400 men born between 1926–1937 and living in Prague 2. The men were divided according to presence of risk factors (RF), overall health conditions and ECG result into the following three groups: normal (a group of men showing no RF defined above), risk (group of men with at least one RF defined above – the prevalence of risk factors for this group is shown in Table 2) and pathological (group of men with a manifested cardio-vascular disease). Long-term observation of patients was based on following the men from normal group and risk group (randomly divided into intervened risk group – RGI and control risk group – RGC). The men from the pathological group were excluded from further observation.

Risk factor	n	%
hypercholesterolemia	290	34.2
hypertension	287	34.0
smoking	543	63.3
obesity	196	23.0
positive family history	216	25.3

Table 2. Prevalence of risk factors in the risk group

Intervention was the key problem of the study. We tried to optimize and modify influenceable RF. Intervention was based on non-pharmacological influence. Pharmacological intervention may be mostly used only in the last years:

1. *Non-pharmacological intervention*: interviews on lifestyle, i.e. diet, physical activity, suitability or necessity to stop smoking and reduce weight. The interviews were repeated during each stay and except for general instructions, they focused also around specific RF of a given man.

2. *Pharmacological intervention*: treatment of arterial hypertension and hyperlipoproteinemy was very limited in the initial stages of the study. Pharmacological therapy was recommended with respect to the overall risk of a given man and his possible other diseases.

Regular visits of a doctor themselves could represent an intervention, provided the patient knew the reason of the visit, parameters to be followed and desirable parameter values.

4 THE STULONG DATA AND PROBLEM DESCRIPTION

STULONG is the data set concerning the twenty years lasting longitudinal study of the risk factors of atherosclerosis in the population of 1 417 middle aged men. For the Discovery Challenges, four data files have been used:

- The file ENTRY contains values of 64 attributes obtained from entry examinations; these attributes are either codes or results of measurements of different variables or results of transformations of the rest of the 244 attributes actually surveyed for each patient.
- Risk factors and clinical demonstration of atherosclerosis have been followed during the control examination for 20 years. The file CONTROL contains results of observation of 66 attributes recorded during these control examinations (10 572 records).
- Additional information about health status of 403 men was collected by the postal questionnaire. Resulting values of 62 attributes are stored in the file LETTER.
- There are 5 attributes concerning death of 389 patients. Values of these attributes are stored in the file DEATH.

The STULONG data were analyzed using some statistical methods: descriptive statistics, logistic regression and survival analysis. The domain experts were curious about applying data mining methods to this data. Therefore they asked some questions concerning some uncovered relations hidden in the data. The listed analytic questions (possible tasks), which have not been subjected to study yet, can be divided into four groups:

- analytic questions related to the entry examination (what are the relations between social factors, or physical activity, or alcohol consumption and the risk factors),
- analytic questions related to the long-term observation (are there any differences between men of the two risk subgroups RGI, RGC, who came down with the observed cardiovascular diseases in the course of 20 years and those who stayed healthy),
- analytic questions concerning postal questionnaire,
- analytic questions concerning entry examination, long-term observation and death.

5 PREPROCESSING AND MODELING

The analytic questions given above predetermined both the modeling and preprocessing steps of the analysis. Preprocessing is the most difficult and most time consuming step in the whole KDD process. The aim of preprocessing is twofold:

1. to extract or construct attributes relevant for the data mining task,
2. to transform the data into formalism suitable for the data mining algorithms to be used.

When working with the ENTRY table only, the preprocessing was rather simple: to solve problems with numerical attributes and with missing values. When including the CONTROL table into the experiments, the preprocessing becomes more complex, as the CONTROL table contains results of laboratory tests taken from one patient over time. One possibility how to process the data was just looking if some value occurs in the list of examinations of one patient [1, 7, 15, 25], the other possibility was to process the subsequent results of one particular test as time series and to work with trends [7, 21, 23, 24].

5.1 Descriptive Tasks

The descriptive tasks – associations or segmentation (subgroup discovery), are used if the main purpose of the data mining is to find some relation between attributes or examples. As the analytic questions suggest to focus on mining for descriptive models, different forms of association rules were the results of most analyses. Beside "classical" association rules [16], rules and exceptions [14], hypotheses in the sense of GUHA method [8, 12, 21], fuzzy rules [5], sequential rules [7], episode rules [23] or first order rules [6, 29] have been used as well. The way of generating the association rules ranged from a systematic exhaustive way done by rule specialization to random search using genetic algorithms.

The mining for association rules was used to solve the questions concerning relations between characteristics of the patients taken from different data tables. The most simple analyses deal with the ENTRY table only; these analyses could answer questions concerning relations between different characteristics collected during the entry examination of patients, e.g. (from [16]):

```
beer(up to 1 liter) ==> vine(daily):  conf(0.98).
```

More complicated analyses included the tables ENTRY and DEATH; here, the DEATH table was usually used to define a new group of patients to be described using the entry examination, e.g. (from [8]):

```
education(university) & height[176-180] ==> death_cause(tumor): conf(0.62)
```

Most complicated analyses included the tables ENTRY and CONTROL. Due to the temporal character of the second table, the resulting associations can express statements like "IF the patient regularly consumed alcohol when he entered the study AND his physical activity after job decreased for several control examinations, THEN his cholesterol rate will increase at a control examination taken X months after that period" [7], "IF BMI is quickly decreasing THEN diastolic pressure is decreasing" [21], or "IF the patient has no hypercholesterolemia AND he sometimes

follows his diet, THEN the patient will have no hypercholesterolemia within next 40 months" [23].

A step towards better understanding of the resulting associations using automated transformation into natural language sentences was shown in [27]. So e.g. an association rule in the form

```
Physical activity after a job(great) &
Physical activity in a job (mainly sits) ==>  BMI(normal)
```

can be automatically transformed into sentences like "X patients confirm this dependence: if a patient has great activity after job and a sedentary job, then s/he has a normal value of BMI" or "a combination of great activity after job and a sedentary job implies a normal value of BMI. This fact is confirmed by X patients".

Interesting concept of emerging patterns has been used in [11, 24]. Emerging patterns are patterns whose frequency increases significantly form one data set to another. This basic idea was used to differentiate between healthy patients and patients with atherosclerosis (the classes were defined using data from the DEATH or CONTROL tables).

Some papers describe results of clustering based on various patients' characteristics. The clusters can correspond to the original groups (e.g. [13]) or can be defined using the data. Subgroup discovery based on the idea of identification of groups of patients showing significantly different rates of cardiovascular disease (CVD) in comparison with the overall CVD rate is shown in [24], the presence of CVD in this analysis was derived from the table CONTROL. Similar definition of clusters has been used also in [28], the aim here was to analyze social characteristics of healthy patients and patients with atherosclerosis.

5.2 Classification Tasks

Classification (and regression) tasks are used if the main purpose of the data mining is to build a model that can be used for decision or decision support. The classification tasks performed on STULONG data deal either with classifying patients into the three predefined groups normal, risk or pathological or with classifying them into classes derived form the data. Different approaches have been used also for this type of analyses. [18] used rough set approach to build a set of decision rules to classify patients into classes defined using the tables CONTROL and DEATH (no disease during follow-up, heart disease during follow-up or as the death cause, other disease during follow-up or as the death cause). In [29], the application of ILP to learn classification rules from the CONTROL table is presented; the goal was to predict whether a person in the risk group comes down with a cardiovascular disease or not. Papers [21, 24] (subsequent results of one research group) describe a model that predicts whether a patient will suffer from CVD in some future based on the values of control examinations taken before.

The classification problem can be eventually turned into a regression problem. A real-valued risk estimate for an individual is proposed in [17] and [1] (the subse-

quent results of one research group). In the first paper, the global risk estimate is defined as the sum on linear combination of variables related to the personal case history (a static part of the cardiovascular risk) and an exponential function variables related to the personal case history (taking into account cumulative effects of risk factors); the parameters of the function were tuned manually. In the second paper presented next year, genetic algorithm has been used to find optimal parameters for the risk function.

Another example of using genetic algorithm to evolve a discrimination function can be found in [30]. The resulting discrimination function was able to model 8 426 (71 %) records correctly; anyway the formula was too complicated to be evaluated and interpreted by the domain experts (see Figure 2).

```
(+ (+ (+ (+ (- (- (- alcohol vzdelani) (- (* (* (+ moc chlst) (+
kysmoc (+ (+ (* (- dusnost pivo12) (- alcohol kysmoc)) (- syst1 (-
hypll HTD))) (* ldl glykemie)))) (+ (* -3.33355 (* glykemie HT)) (+
(+ (+ (+ imtrv (* -3.33355 (* glykemie HT))) (+ (* glykemie HT) (+
(+ (- hypll HTD) (* ldl glykemie)) (* ldl glykemie)))) (- alcohol
vzdelani)) (+ (* ldl glykemie) glykemie)))) (+ (+ (- ICT vinomn)
(+ (+ (- (* -3.33355 byvkurak) HT) (* -3.33355 (* glykemie HT)))
hypll)) (* (- vyska HTD) (+ dusnost alcohol))))) HT) (* -3.33355 (*
glykemie HT))) dobakour) (+ (* (+ imtrv (* -3.33355 (* glykemie HT)))
byvkurak) syst2)) (+ (+ (+ vzdelani (+ (* vinomn byvkurak) smoking))
(* (- dusnost pivo12) (- dusnost pivo12))) (+ (+ (+ glykemie (*
glykemie HT)) (- hypll HTD)) (* ldl glykemie))))
```

Fig. 2. Discrimination function, taken from [30]

An approach based on co-occurrence matrix and principal component analysis (PCA) is described in [25]. Co-occurrence matrix describes (for one patient) the strength of relations between pairs (rows and columns) of characteristics. The set of these matrices is then used to create (using PCA) so-called feature vectors. The feature vector can then be presented in a network model. Figure 3 shows a network model that describes unhealthy behaviors that healthy patients shared in common, but did not develop any cardiovascular diseases at the end of the study.

6 EVALUATION AND DEPLOYMENT

6.1 The Data Mining Expert Point of View

It is worth mentioning that during the three subsequent Discovery Challenges 2002–2004 that used the STULONG data the analyses have become more elaborated and more complex. In the first year, all papers focused on the analysis of the ENTRY table only, and used mainly association rules to find relations between different characteristics of the patients. In the next Challenges, data miners used

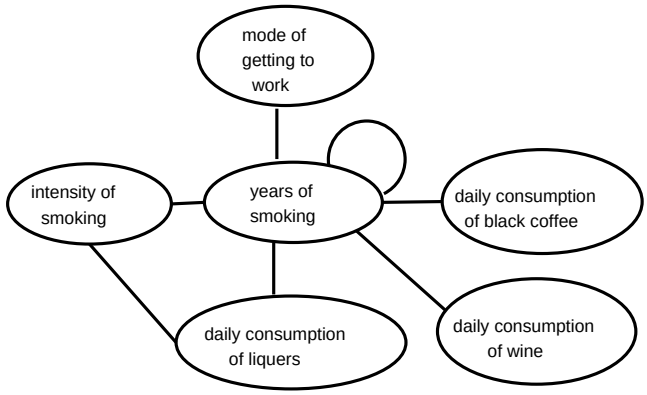

Fig. 3. Network model of unhealthy behaviors of healthy group, taken from [25]

a broader range of data mining methods including rules, exceptional cases, clusters, classification, emerging patterns or linear algebra. Several contributions focused on mining temporal data (e.g., sequential or episode rules, trend analysis, temporal abstraction, clustering time-series). Some contributions combined several methods and new approaches were proposed. Let us mention e.g. the transformation of association rules into expressions in natural language [27], emerging patterns [11], trend analysis [21], or eigen co-occurrence matrix algorithm [25]. So the Discovery Challenge contributed not only to a specific medical domain but also to data mining and machine learning in general.

6.2 The Domain Expert Point of View

The data providers gained from the Challenge deeper insight into the data. The experts preferred the results of description tasks in favor of the results of classification tasks. Even if the found associations were often no surprise, they were better accepted than the less understandable classification models. The need of understandability of the resulting models can be demonstrated on the results shown in Figures 3 and 4. Although in both cases rather complicated methods (genetic algorithms, linear algebra) have been used, the graphical representation of the latter was well acknowledged.

Some interesting, sometimes counter-intuitive results obtained in description tasks have been further analyzed and interpreted. As an example let us mention here the positive impact of drinking beer on the health status of the patient (not observed for wine) or the positive influence of number of visits during the follow-up examinations. The former result correlates with the known positive impact of alcohol consumption (Czech men are rather beer drinkers than wine drinkers), the latter result can be interpreted by the fact that during the visits the doctor educates the patient about healthy life style and also checks the impact of pharmacological

intervention on his health status. An interesting useful result was the correlation of Body Mass Index with the skin folds – a very good discrimination of the three basic groups of men (see Figure 4), or the negative relation between the level of education and the smoking habits.

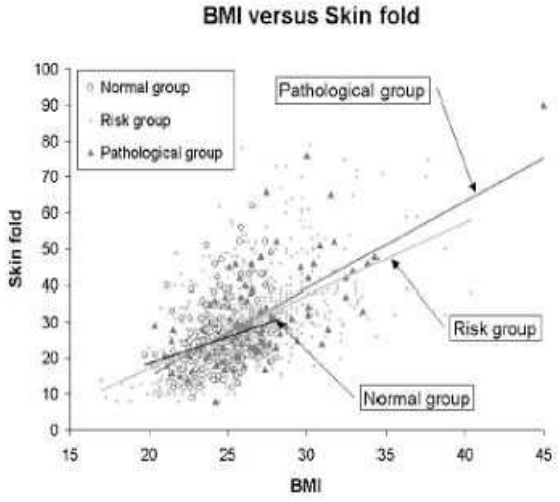

Fig. 4. Correlation between skin fold and BMI for different groups, taken from [26]

7 THE DISCOVERY CHALLENGE EXPERIENCE

Different groups of people involved in the Challenge gained different benefits. The data providers gained deeper insight into the data. The experts preferred the results of description task in favor of the results of classification tasks. Even if the found associations were often not surprising, they were better accepted than the less understandable classification models. The experts took regularly part in the Challenge workshops and evaluated the results. From their point of view, the Challenge was very useful.

The Challenge participants had the opportunity to analyze large real-world data and to test and present their approach. They gained a hands-on experience with realistic data mining projects. Such experience can motivate further research. The tasks formulated and solved in the Challenge can be reused in similar domains.

The reward for the organizers was the interest of the machine learning and data mining community in the Challenge tasks and Challenge workshops. Of course, lot of work had to be done for the Challenge success. The main organization work concentrates around the Challenge announcement and the Challenge deadline. The crucial point was to prepare the data for the Challenge; this included not only data

collection but also translation of data description (sometimes difficult for the domain experts). The Challenge deadline was chosen to be as late as possible with respect to deadline for printing the proceedings; this left less time for a review process (only obviously irrelevant papers were rejected) but gave the participants the possibility to work until the very end. The registration and password based access to the data was used mainly to keep track of the interested persons. So we could inform all participants about eventual changes. From this list we also found out that the response rate (the fraction of people downloaded the data who submitted a paper) was rather low, about 10 %. This was probably because of less competitive nature of the Challenge and less clear description of the tasks to be solved.

To substitute the discussions with domain experts during the data understanding step, we maintained a list of (participants') questions and (experts') answers on the Challenge web page.

The same data and same problems were used in three subsequent Challenges. Since the papers were available on the Challenge web page, we expected that the participants would use the previously published results. This did not happen. With the exception of two groups that participated in two Challenges and used their own results from the year before, we did not observe any expected synergy effect.

8 LESSONS LEARNED

Let us summarize the lessons we have learned from the Discovery Challenge results. We will give here some conditions we believe that must be fulfilled in a successful data mining project:[1]

Cooperate with domain experts: The well known problem with expert systems is the so called knowledge acquisition bottleneck. This is the nickname for the necessity to involve the domain expert in the time consuming and tedious process of knowledge elicitation. Machine learning methods were understood as a way how to eliminate this problem. Nevertheless, in real-world data mining, we can observe similar problem we can call "data acquisition bottleneck". The most difficult steps of the KDD process are data understanding and data preparation. In real-world problems, we need experts who help with understanding the domain, with understanding the problem, with understanding the data.

Use external data if possible: There are many external factors that are not directly collected for the data mining task, but can have a large impact on the data analysis.

Use powerful preprocessing methods: Typical preprocessing actions are joining tables, aggregation, discretization and grouping, handling missing values, creating new attributes. Very often, these operations are performed as domain

[1] Similar lessons have been drawn e.g. by R. Kohavi in his paper "Lessons and Challenges from Mining Retail e-commerce Data" [22].

independent. More background knowledge should be used in these transformations.

Look for simple models first: One of the common sources of misunderstanding between domain experts and data mining experts is that data mining experts are interested in applying their sophisticated algorithms and domain experts are interested in simple results. Sometimes even "simple" reporting and summarization gives acceptable results.

Make the results understandable and/or acceptable: The crucial point for success of a data mining application on real-world problem is the acceptance of results by the domain experts and potential users. The best solution is worthless if it is not used. Understandability of results is the keyword for this lesson. Domain experts are not interested in tables showing improvement of accuracy of 2.47 % or in lists of thousands of rules. They want to know the strengths and limitations of the classifiers or insight into found patterns. So explanation of the results, postprocessing or visualization is of great importance.

Show some preliminary results soon: To convince the domain experts (and the managers as well) about the usefulness of data mining methods, some preliminary results should be delivered in the early stage of the project. Even an initial data exploration can be very appreciated.

Assess the ROI of the models: The experts and users are interested in the benefit, the models bring when applying and deploying them. This can be better expressed as return of investment (ROI) rather then as classification accuracy.

9 CONCLUSIONS

The reusability of successful data mining solutions can help in new data mining projects. This fact has been recognized by the machine learning and data mining communities. An example for this is the EU research project MiningMart aiming at collecting solutions of business problems with emphasis on preprocessing (e.g. [20]). Discovery Challenges can provide a workbench for finding such prototype solutions of realistic problems. Atherosclerosis risk factors data are a good example or a real tractable data usable for such a purpose.

The Discovery Challenge participants had the opportunity to analyze large real-world data and to test and present their approach. They gained a hands-on experience with realistic data mining projects. Such experience can motivate further research – this was especially the case of groups that participated in more than one Discovery Challenge workshop. All papers presented at the Challenge Workshops are listed in the references and are available at the Discovery Challenge home-page `http://lisp.vse.cz/challenge`.

The main lesson learned from the series of Discovery Challenge workshops is (quoting the words of R. Kohavi from his KDD Cup 2000 talk) "Don't expect to press a button and get enlightenment. Data mining process is like peeling the onion."

Knowledge discovery is a long, tedious process, that requires cooperation of people from different areas. A success of this process depends on a number of factors, the most important being mentioned in this paper.

Acknowledgment

We are grateful to Jan Zytkow who came with the idea of Discovery Challenge as a collaborative approach and who initiated the first Discovery Challenge workshop held in 1999 in Prague during the PKDD conference.

The STULONG study was realized at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2 (head prof. MUDr. M. Aschermann, Dr. Sc.), under the supervision of prof. MUDr. F. Boudík, Dr. Sc., with collaboration of MUDr. M. Tomečková, C. Sc. and doc. MUDr. J. Bultas, C. Sc. The data were transferred to the electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences (head. prof. RNDr. J. Zvárová, Dr. Sc.).

The work is partially supported by the project 1M06014 of Ministry of Education of the Czech Rrepublic, the project 201/05/0325 of the Grant Agency of the Czech Science Foundation and the project 25/05 of the Grant Agency of University of Economics, Prague.

REFERENCES

[1] Azé, J.—Lucas, N.—Sebag, M.: A New Medical Test for Atherosclerosis Detection GeNo. In [3].

[2] Berka, P.—Cremilleux, B. (eds.): Discovery Challenge Workshop Notes. ECML/PKDD 2004, Pisa, 2004.

[3] Berka, P. (ed.): Discovery Challenge Workshop Notes. ECML/PKDD 2003, Cavtat-Dubrovnik, 2003.

[4] Berka, P. (ed.): Discovery Challenge Workshop Notes. ECML/PKDD 2002, Helsinki University, 2002.

[5] Berzal, F.—Cubero, J. C.—Sanchez, D.—Serrano, J. M.—Vila, M. A.: Finding Fuzzy Approximate Dependencies within STULONG Data. In [3].

[6] Blaťák, J.: Mining First-Order Frequent Patterns in the STULONG Database. In [2].

[7] Brisson, L.—Pasquier, N.—Collard, M.—Hébert, C.: HASAR: Mining Sequential Association Rules for Atherosclerosis Factor Analysis. In [2].

[8] Burian, J.—Rauch, J.: Analysis of Death Causes in the STULONG Data Set. In [3].

[9] Chapman, P.—Clinton, J.—Kerber, R.—Khabaza, T.—Reinartz, T.—Shearer, C.—Wirth, R.: CRISP-DM 1.0 Step-by-Step Data Mining Guide. SPSS Inc. 2000.

[10] Couturier, O.—Delalin, H.—Fu, H.—Kouamou, G. E.—Mephu-Nguifo, E.: A Three-Step Approach for STULONG Database Analysis: Characterization of Patients' Groups. In [2].

[11] Cremilleux, B.—Soulet, A.—Rioult, F.: Mining the Strongest Emerging Patterns Characterizing Patients Affected by Diseases Due to Atherosclerosis. In [3].

[12] Dolejší, P.—Lín, V.—Rauch, J.—Šebek, M.: System of KDD Tasks and Results within the STULONG Project. In [4].

[13] Durand, N.—Cleuziou, G.—Soulet, A.: Discovery of Overlapping Clusters to Detect Atherosclerosis Risk Factors. In [2].

[14] Goncalves, E. C.—Plastino, A.: Mining Strong Associations and Exceptions in the STULONG Data Set. In [2].

[15] Lachiche, N.—Fuchs, S.—Gancarski, P.—Derivaux, S.: Discriminative Power of Related Controls in the STULONG Project. In [2].

[16] Lin, F.: Longitudial Study of Atherosclerosis Risk Factors: Relation Factors to BMI and Finding in Entry Examination. In [4].

[17] Lucas, N.—Azé, J.—Sebag, M.: Atherosclerosis Risk Identification and Visual Analysis. In [4].

[18] Hoa, N. S.—Son, N. H.: Analysis of STULONG Data by Rough Set Exploration System (RSES). In [3].

[19] Karban, T.: SDS-Rules and Classification on PKDD2003 Discovery Challenge. In [3].

[20] Kietz, J. U.—Zücker, R.—Vaduva, A.: Mining Mart: Combining Case-Based-Reasoning and Multi-Strategy Learning into a Framework to reuse KDD-Application. In: Proc. 5th Int. Workshop on Multistrategy Learning MSL2000.

[21] Kléma, J.—Nováková, L.—Karel, F.—Štěpánková, O.: Trend Analysis in Stulong Data. In [2].

[22] Kohavi, R.—Mason, L.—Parekh, R.—Zheng, Z.: Lessons and Challenges from Mining Retail E-Commerce Data. Machine Learning, Vol. 57, 2004, No. 1/2, pp. 83–114.

[23] Meger, N.—Leschi, C.—Lucas, N.—Rigotti, C.: Mining Episode Rules in STULONG Dataset. In [2].

[24] Nováková, L.—Kléma, J.—Jakob, M.—Rawles, S.—Štěpánková, O.: Trend Analysis and Risk Identification. In [3].

[25] Oka, M.—Koiso, T.—Meng, E.—Kato, K.: Extracting Features of Patients Using the Eigen Co-Occurrence Matrix Algorithm. In [2].

[26] Salleb, A.—Turmeaux, T.—Vrain, C.—Nortet, C.: Mining Quantitative Association Rules in a Atherosclerosis Dataset. In [2].

[27] Strossa, P.—Rauch, J.: Association Rules in STULONG and Natural Language. In [4].

[28] Soulet, A.—Hébert, C.: Using Emerging Patterns from Clusters to Characterize Social Subgroups of Patients Affected by Atherosclerosis. In [2].

[29] van Assche, A.—Verbaeten, S.—Krzywania, D.—Struyf, J.—Blockeel, H.: Attribute-Value and First Order Data Mining within the STULONG Project. In [3].

[30] Werner, J. C.—Kalganova, T.: Risk Evaluation using Evolvable Discriminate Function. In [3].

[31] Wettschereck, D.: Educational Data Preprocessing. In [4].

Petr Berka is a full professor at the Department of Information and Knowledge Engineering, University of Economics, and also works in the Centre of Biomedical Informatics, Institute of Computer Science, Academy of Sciences of the Czech Republic. His main research interests are machine learning, data mining and knowledge-based systems.

Jan Rauch is an associate professor at the Department of Information and Knowledge Engineering, University of Economics, and also works in the Centre of Biomedical Informatics, Institute of Computer Science, Academy of Sciences of the Czech Republic. His main research interest is data mining.

Marie Tomečková is a physician working at the Department of Medical Informatics and in the Centre of Biomedical Informatics, Institute of Computer Science, Academy of Sciences of the Czech Republic. Her specialisation is non-invasive cardiology, namely epidemiology of atherosclerosis.