Attribute-Value and First Order Data Mining within the STULONG project

Anneleen Van Assche, Sofie Verbaeten, Darek Krzywania, Jan Struyf, Hendrik Blockeel

Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200 A, B-3001 Heverlee, Belgium
{anneleen,sofie,darek,jan,hendrik}@cs.kuleuven.ac.be

Abstract. The ECML/PKDD-2003 Discovery Challenge concerns knowledge discovery in the field of atherosclerosis risk factors. This paper reports on our data mining effort within this challenge. Since the data is presented in attribute-value format, a first obvious approach is to mine the data using standard attribute-value techniques. We used a variety of methods implemented in the Weka data mining tool. However, since the data can be seen as a propositionalisation of a multi-relational database, it might be advantageous to first transform the data into a relational form and then use an inductive logic programming technique. We also followed this approach and used the ACE data mining tool built in our research lab. We discuss and compare the results of these different approaches.

1 Introduction

In the early seventies of the twentieth century, a longitudinal study (STULONG) of atherosclerosis risk factors was developed in Czechoslovakia. The ECML/PKDD-2003 Discovery Challenge concerns knowledge discovery in data collected in this STULONG project. The data was gathered during twenty years in the population of 1417 middle aged men: each man is described by a variety of attributes collected during an entry examination, and the various control examinations of a patient are described by the changes of the patient w.r.t. the previous examination.

Since the data is presented in attribute-value (AV) format, a first obvious approach for mining this data is to use standard AV data mining techniques. We used different techniques implemented in the data mining tool Weka [5]. However, since the data can be seen as a multi-relational database, a second approach for mining the data is to use an inductive logic programming (ILP) technique. The relational form of the data avoids many empty as well as redundant entries. Also, background knowledge

can be easily added. We also followed this approach and used ACE [2], an inductive logic programming tool developed in our research lab.

In the following section, we first describe the data and data preprocessing step. We next present the data mining techniques we used, as well as the evaluation criteria. In section 3, we present the discovered knowledge. We conclude in section 4.

2 Data Mining Effort

2.1 The Data

The STULONG data set consists of four data matrices, two of them were studied in this work:

- The Entry data matrix:
  This data matrix contains 1417 instances describing middle aged men using 64 attributes. The attributes were collected during an entry examination. They are divided into the following groups: identification data, social characteristics, physical activity, smoking, drinking of alcohol, sugar/coffee/tea, personal anamnesis, questionnaire A_2, physical examination, biochemical examination, and risk factors.
- The Control data matrix:
  This data matrix contains 10572 instances each describing a control examination of a particular man (one of the 1417 men described in the Entry data matrix) using 66 attributes. The attributes have been recorded for the duration of 20 years. They are divided into the following groups: identification data, patient's changes since the last examination, sickness leave, A_2 questionnaire, physical examination, and biochemical examination.

Furthermore, the men in the Entry data are divided into 3 subgroups: the normal group (NG; 276 instances; 19.48 %), the risk group (RG; 859 instances; 60.62%), and the pathological group (PG; 114 instances; 8.05%). The division into these groups is based on the occurrence of risk factors (e.g. heavy smoker, overweight, high cholesterol level, ...), observed at the entry examination. The group NG includes men that did not show any of these risk factors, men in RG had at least one of the risk factors, and men in PG had a manifested cardio-vascular disease or other serious disease (making their long-term observation impossible). Full details of the STULONG project and challenge can be found at http://euromise.vse.cz/challenge2003.

2.2 Data Pre-Processing

As already observed in the ECML/PKDD-2002 Discovery Challenge on Atherosclerosis Data (see e.g. [4]) the encoding of missing values and values stating that a certain property does not hold is handled inconsistently in the data. The value for "not stated" (mostly code 6) is in some cases used to denote a missing value, in other cases to denote a "no" (the property does not hold). Also, the missing values (empty entries) in some cases denote a "no". We took a closer look at all attributes and mapped their values to a unambiguous encoding.[1] We converted the Entry and Control data sets in Weka .arff format [5].

We also decided to represent the data in a relational form, since both the Entry and Control data matrices can be seen as a propositionalisation of two relational databases. As a result of this propositionalisation they have many empty entries that aren't relevant for a certain patient or control. For example, most of the attributes relating to the personal anamnesis are redundant for the patients, since they don't have those particular diseases. This can be avoided in a relational representation of the data. Also the Entry and Control data matrices together are actually a multi-relational database (there is a 1-n relation from the Entry to the Control data matrix).

Another advantage of the relational form is that background knowledge can be used, so it is very easy to introduce new characteristics that can be computed from the original data set. In the AV representation, new attributes (based on attributes already in the data sets) were added by using the Weka filters or by running some scripts. Some of these new attributes will be discussed in section 3.

2.3 Data Mining

The data sets in .arff format were mined using the Weka [5] data mining tool. In particular we used the following Weka methods for *classification*: ZeroR, OneR, IBk (with k = 1,5,10), Naive Bayes, Decision Stump, Decision Table, j48.PART, j48.J48, j48.J48 with binary splits, Logistic regression, and also Linear Regression and M5' in classification via regression mode (all with default settings). For *regression* we used: Linear Regression and M5'. Apriori was used to find *association rules*.

The relational representation of the data sets was mined using the ILP system ACE [2]. In particular we used TILDE (Top-down Induction of

[1] This was not always possible; for some attributes no distinction could be made between "no" and truly missing values.

Logical Decision Trees) [1]. TILDE is an ILP extension of the C4.5 decision tree algorithm [3]. Instead of using attribute-value tests in the nodes of the tree, logical queries are used. TILDE can be used in classification or regression mode. Since for many tasks the data distributions are skewed, it makes sense just to predict the chance of being positive/negative, instead of trying to build a classification tree. We therefore used TILDE mostly in regression mode.

2.4 Evaluation Criteria

For classification and regression tasks we performed a 10-fold cross-validation. The classifiers were evaluated based on their ROC performance and accuracy. The regression models were evaluated based on their relative absolute error (RAE) and correlation.

3 Discovered Knowledge

We first present an initial exploration of the Entry data set. Then, we show some interesting association rules, also found in the Entry data set. In section 3.3 we present and discuss the AV and ILP predictive models that were built for the different discovery challenge tasks in the Entry data set. Most of the tasks are answered at a high level, one is discussed in more detail. Finally, in section 3.4, we report on our results of mining the Control data set. For the Control data set, we only used ILP techniques.

3.1 Initial Exploration of the Entry Data Set

As a first step in the exploration of the Entry data set, we compare the averages of some attributes in the different subgroups NG, RG, and PG. In particular, we compare the averages of attributes that are non-risk factors[2], and for which the values can be ordered. In Fig. 1, two attributes related to social characteristics, two attributes related to physical activities, and two attributes concerning skin fold are compared for the three subgroups: `VZDELANI` (reached education), `ZODPOV` (responsibility in a job), `TELAKTZA` (physical activity in a job), `AKTPOZAM` (physical activity after a job), `TRIC` (skin fold above musculus triceps in mm), and `SUBSC` (skin fold above musculus subscapularis in mm).

From this first, simple step we can conclude that people who are not at risk and aren't ill already (NG) have reached a higher education level,

[2] It is trivial that the risk factors, like e.g. smoking, will have higher values in the groups RG and PG.

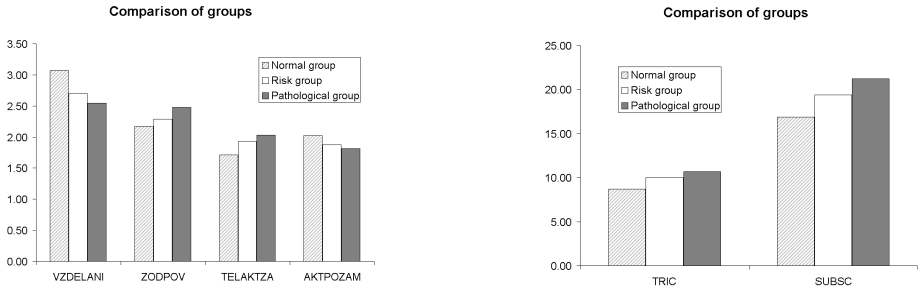

Fig. 1. Comparison of mean values of attributes for the three subgroups

have a higher responsibility in their job, do less physical activities during their work and more in their leisure time, than the other two groups. We can draw the same conclusion for the risk group in comparison to the pathological group. Comparing the mean values of `TRIC` and `SUBSC`, we observe that the risk group shows an increase of 15 percent for both attributes `TRIC` and `SUBSC` in comparison to the normal group, and the pathological group, 7 and 9 percent respectively in comparison to the risk group.

The following graph (Fig. 2) gives us some idea of the correlation between skin folds and the `BMI` (see also section 3.3). The attribute skin

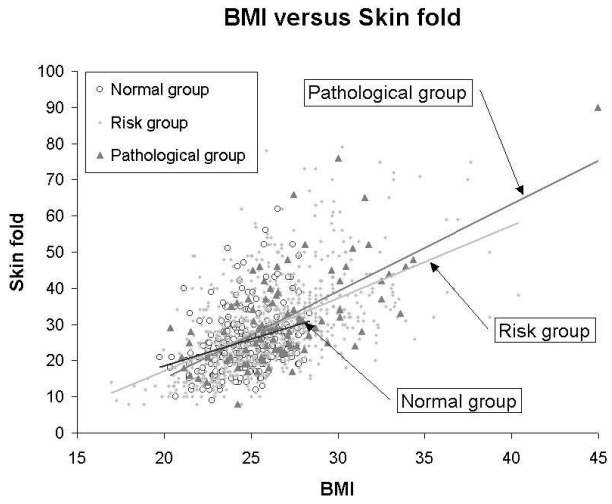

Fig. 2. Correlation between BMI and skin fold for the three subgroups

fold is the average of the attributes TRIC and SUBSC. For each of the three subgroups a trend line is drawn. The slopes of these trend lines are 1,47 (NG), 2,01 (RG), and 2,41 (PG). Because the slopes increase with the risk level and the trend lines approximately cross at BMI value 24, we can conclude that from two people with the same BMI (>24), the one with the highest skin fold is more at risk.

3.2 Association Rules in the Entry Data Set

For some of the discovery challenge tasks association rules were computed in Weka using Apriori. The tasks concern finding relations between two groups of attributes (see section 2.1 for the different groups of attributes in the Entry data).

Below some association rules are shown. The number in the premise denotes the number of examples that are covered by it and the number in the consequence refers to the number of examples that are covered by the complete association rule. Confidence (conf) is the proportion of examples covered by the premise that are also covered by the consequence. Lift is confidence divided by the proportion of all examples that are covered by the consequence. This is a measure of the importance of the association that is independent of support.

```
Relation between social factors and physical activity

university 397 ==> married & mainly sits during job 286
conf:(0.72) < lift:(1.57)
university 397 ==> mainly sits during job 316
conf:(0.8) < lift:(1.53)

Relation between physical activity and smoking

didn't stop smoking & smoker 803 ==> mainly stands during job & smokes +21 years 438
conf:(0.55) < lift:(1.63)
way to work takes 1/2 hour & smoker 697 ==> smokes +21 years 449
conf:(0.64) < lift:(1.41)

Relation between physical activity and cholesterol level

way to work takes 1/2 hour 962 ==> way to work on foot & no cholesterol risk 228
conf:(0.24) < lift:(1.34)

Relation between physical activities

way to work on foot 353 ==> mainly stands during job & way to work takes 1/2 hour 231
conf:(0.65) < lift:(1.34)
```

3.3 Classifiers for the Entry Data Set

We also answered the discovery challenge task w.r.t. the Entry data set by building predictive models. For each question, the relevant attributes were selected and models were built on the data set including all 1417 instances, and on NG, RG, and PG. Different methods for building these models were used (see subsection 2.3), and we compared them using different evaluation criteria (see subsection 2.4). We also took a closer look at the trees, model trees, and linear equations to study the influences of the different attributes. We compared such influences in the models built on the whole data set versus in those built on the different subgroups.

Because of lack of space, we decided to report only one of the results in detail and briefly provide an overview of the results on the other questions. We only mention the different risk groups if there are differences w.r.t. the whole group of patients.

Relations between social factors and other characteristics
We discovered the following relations concerning physical activity of the patients. The higher educated (secondary school and university vs. basic and apprentice school), the less physical activity in job. Other workers than managerial and partly independent workers seem to do more physical activity in job. In the risk and pathological group singles also appear to do more physical activity in their job. Married people on the other hand do less physical activity in their job. After their job, managerial and partly independent workers tend to do more physical activity than others.

With regard to smoking, higher educated people seem to smoke less. Furthermore, singles and widowers smoke less than married and divorced people.

With respect to alcohol consumption we found that pensioners tend to drink less than other people. Moreover people with a higher education level appear to drink less.

People with apprentice school as their highest education level have higher BMI and from that group those that are married have even higher BMI. Other workers than managerial and partly independent workers also have higher BMI.

Blood pressure seems to increase with respect to age. In the normal group of patients, people with secondary school as their highest education level turn out to have more chance to have a high pressure.

No significant relations were found in the whole data set between social factors and cholesterol. In the risk group on the other hand the

oldest people have a smaller cholesterol level. In the pathological group we also found that pensioners because of ICHS have smaller chance to have increased cholesterol.

Relations between physical activities and other characteristics
Concerning smoking we found that people that do great physical activity after their job smoke less than others. People that go to work by car seem to smoke slightly more. The first relation is also represented in the normal and pathological group and the last in the risk group. In the risk group we also found that people that go to work on foot appear to smoke less.

Alcohol consumption turns out to be smaller for people whose transport to work takes more than one hour.

With respect to BMI we found that people that do great activity after their job have less chance to be overweight. For the pathological group we experienced that people that carry have loads in their job have more chance to be overweight.

No significant relations were found in the whole data set between physical activities and blood pressure. In the risk group however we found that people that go to work on foot have higher blood pressure.

No significant relations were found in the whole data set between physical activities and cholesterol. For the normal group it seems that people who use public means of transport have less cholesterol.

Relations between alcohol consumption and other characteristics
With respect to smoking we experienced that the more people drink, the more they usually smoke. For wine this is not always clear.

Regarding alcohol consumption, people that drink daily up to one liter of beer have more chance to be overweight. In the pathological group it seems that non-drinkers have less chance to be overweight.

People that drink daily more than one liter of beer have also more chance to have a high blood pressure.

Skin folds and BMI
The (regression) task is to predict `BMI` using the attributes `SUBSC` (skin fold above musculus subscapularis) and `TRIC` (skin fold above musculus triceps). We also consider a classification task, namely predicting `OVERWEIGHT` (with value 1 if `BMI` > 25, and 0 otherwise).

From Table 1 with results from TILDE can be seen that the ROC performance for `OVERWEIGHT` for the whole data set is quite good. The results

reveal that the correlation between skin folds and BMI is the strongest in the risk group.

Looking at the trees built by TILDE, we also noticed that people from NG and RG with SUBSC smaller than 15 both have an average BMI of 23.9, while the same people from PG have a BMI of 24,5. The trees indicate that, for all the different groups, SUBSC with a threshold around 15 is the best way to distinguish between overweight en non-overweight. From the different classifiers we could also conclude that the higher the values of SUBSC the higher the BMI. The influence of TRIC on BMI and OVERWEIGHT is less than that of SUBSC. Also, the influence of TRIC isn't always monotonous. For higher values of SUBSC, TRIC has sometimes a negative influence on the BMI value.

Table 1. Statistics on the experiments of BMI and OVERWEIGHT (OW) for the different groups (T is whole group, NG normal group, RG riks group and PG pathological group). Size is the number of nodes in the tree, ACC the accuracy, RAE the relative absolute error, r the correlation coefficient and AUC the area under the ROC curve, all measured using 10-fold cross-validation. AUC(33%) is the ROC area measured using a separate 33% test set. Since BMI is a regression task, only Size, RAE and r could be provided.

Experiment	Size (nodes)	ACC	RAE	r	AUC	AUC(33%)
OW_T	6.0	71%	0.83	0.417	0.71	0.74
OW_NG	0.6	53%	1.01	-0.009	0.50	0.53
OW_RG	3.9	74%	0.78	0.467	0.73	0.72
OW_PG	1.0	75%	0.97	0.205	0.55	0.56
BMI_T	6.9		0.77	0.486		
BMI_NG	1.0		0.90	0.309		
BMI_RG	3.9		0.72	0.529		
BMI_PG	1.7		0.86	0.304		

Staying healthy in the risk group

The task is to predict if a person of RG (859 instances) came down with the observed cardiovascular diseases in the course of 20 years, using the attributes in the groups: social characteristics, physical activities, smoking, alcohol consumption, BMI, blood pressure, and cholesterol level.

An extra attribute ILL was added to the AV representation of the persons in RG with values 0 (68.22 %; meaning that the person did not come down with the observed cardiovascular diseases in the course of 20 years) and 1 (31.78 %; otherwise). The value of ILL is based on the HODNO attribute in the Control data set. With Weka we built models that predict ILL based on the proposed characteristics. The classifiers did not perform much better than random prediction.

We also used our relational representation of the data, and built regression trees with TILDE for the different aspects separately, predicting the chance of staying healthy. Also here, the performance was not very good. We note that the highest correlation coefficient we found was 0.15 for the cholesterol level. In particular, we found that people with cholesterol level smaller than 250 have more chance to stay healthy. The correlation coefficients for the other characteristics were all smaller than 0.05.

3.4 Classifiers for the Control Data Set

In this section we present our results for the Control data set. We use the relational Control data set that stores for each patient information about a number of control examinations. The goal is to predict whether a person in the risk group comes down with a cardiovascular disease (class 1) or not (class 0). We use TILDE for this classification task with different sets of input attributes corresponding to the interesting questions identified by the domain expert. We only use control examinations *ce* that come before the start of the patient's first cardiovascular disease ($ce.year \leq ROK_i$), in order to make sure that the results can be used for prior prediction.

For numeric attributes, we include extra features (defined in TILDE's background knowledge) that compute the trend of the numeric attribute over the different examinations. This trend is computed as the slope of a least squares model of the attribute over the time interval from $T - N$ tot T, with T the start of the patient's first disease and N a parameter that TILDE can fill in. For patients in class 0 we set T to the year of their last examination. We do not consider persons with less than two control examinations before their first disease because we need at least two examinations to compute a trend. After this filter step, we obtain 719 examples, 230 in class 1 and 489 in class 0. For each experiment we use TILDE in the "classification via regression" mode and set the F-test pruning parameter to 0.005 (all other parameters are left at their default values).

Table 2 presents some statistics about the different experiments, measured with 10-fold cross-validation. For most sets of input attributes the correlation r is small. If there is no correlation then a possible conclusion is that there is no relation between the particular set of input attributes and the fact that the patient gets a cardiovascular disease. Here this is the case for "physical activity", "change of diet" and "BMI". The combined experiments do not perform much better than the best subset (e.g. "BMI & Cholesterol" and "Cholesterol"). The model built with all input

Table 2. Statistics on the Control data experiments. Size is the number of nodes in the tree, ACC the accuracy, RAE the relative absolute error, r the correlation coefficient and AUC the area under the ROC curve, all measured using 10-fold cross-validation. AUC(33%) is the ROC area measured using a separate 33% test set.

Input attributes	Size (nodes)	ACC	RAE	r	AUC	AUC(33%)
Job	1.0	68%	0.79	0.162	0.54	0.55
Physical activity	0.1	68%	1.01	-0.09	0.43	0.51
Smoking	3.7	67%	0.98	0.153	0.57	0.61
Diet	0.0	68%	1.00	-0.11	0.43	0.50
BMI	1.4	67%	1.03	-0.04	0.46	0.50
Blood pressure	3.3	63%	1.03	0.106	0.55	0.58
Cholesterol	9.1	64%	1.03	0.147	0.59	0.57
Glycaemia & Uric acid	3.3	66%	0.81	0.435	0.71	0.71
BMI & Cholesterol	10.6	63%	1.05	0.147	0.59	0.57
Smoking & Cholesterol	12.5	63%	1.04	0.164	0.59	0.60
All	8.5	66%	0.91	0.351	0.71	0.69

features does reasonably well, but "Glycaemia & Uric acid" alone does better. Note that the statistics in Table 2 are negatively biased (e.g., $r < 0$, $AUC < 0.5$) because they were estimated with cross-validation. In order to determine the effect of this bias, we also measured the AUC on a separate 33% test set.

We now list some interesting subgroups that can be derived from the decision trees. A subgroup of patients is interesting if it is large and if the proportion of class 1 persons in the group deviates significantly from the proportion in the population (32%). The predicted proportion and subgroup size is shown in the THEN part of each rule.

IF glycaemia > 7.2 and BMI > 23.4 in each examination and diastolic blood pressure slope during last 10 years < -0.77 THEN 64% (103)

IF systolic blood pressure slope during last 20 years < -0.97mm THEN 53% (122)

IF systolic blood pressure slope during last 20 years > 2.39mm THEN 52% (87)

IF HDL slope in last 20 years > -0.03mmol and > 1.55mg% THEN 52% (61)

IF slope in number of cigarettes during last 20 years > 0.48, and during last 10 years > -0.11 THEN 51% (35)

IF glycaemia > 7.2 in each examination THEN 48% (434)

IF the patient leaves to full retirement in some examination THEN 20% (233)

IF reduced smoking in some examination and slope in number of cigarettes during last 20 years < 0.48 but during last 10 years > -0.11 THEN 16% (116)

IF HDL slope in last 20 years < 1.55mg%, triglycerides slope during last 5 years

$<$ 14.1mg% and HDL slope in last 3 years $>$ 0.053mmol THEN 10% (165)

IF glycaemia $<$ 7.2 in some examination THEN 7% (285)

To summarize, "Glycaemia" seems to be the most important attribute. This can be seen from the correlation coefficient $r = 0.435$ in Table 2 and from the 3 interesting rules that include this attribute. Also blood pressure, cholesterol and smoking are important factors. The features that compute the slope of numeric attributes proved to be useful (they occur in almost every rule).

4 Conclusions

We have experimented with a variety of data mining techniques, including well-known propositional techniques (using the Weka data mining tool) as well as multi-relational techniques (using the ACE data mining tool). We have obtained a number of results that we think may be of interest. Our interpretation of these results is necessarily limited to the data miner's point of view; we lack the necessary background to interpret them in terms of the application domain. Further interpretation of these results by domain experts is therefore necessary.

Acknowledgements

AVA and DK are supported by the GOA/2003/08(B0516) on Inductive Knowledge Bases. SV and HB are Postdoctoral Fellows and JS a research assistent of the Fund for Scientific Research - Flanders (Belgium)(F.W.O. - Vlaanderen).

References

1. H. Blockeel and L. De Raedt. Top-down induction of first order logical decision trees. *Artificial Intelligence*, 101(1-2):285–297, June 1998.
2. H. Blockeel, L. Dehaspe, B. Demoen, G. Janssens, J. Ramon, and H. Vandecasteele. Improving the efficiency of inductive logic programming through the use of query packs. *Journal of Artificial Intelligence Research*, 16:135–166, 2002.
3. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann series in machine learning. Morgan Kaufmann, 1993.
4. D. Wettschereck. Educational Data Preprocessing. In P. Berka, J. Rauch, and S. Tsumoto, editors, *The ECML/PKDD 2002 Discovery Challenge on Atherosclerosis Data*, 2002.
5. I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.