Mining Quantitative Association Rules in a Atherosclerosis Dataset

Ansaf Salleb Teddy Turmeaux Christel Vrain Cyril Nortet

LIFO, Université d'Orléans
B.P. 6759, F-45067 Orléans France

{salleb, turmeaux, cv, nortet}@lifo.univ-orleans.fr

Abstract. Mining association rules in databases has become a popular task in data mining. However, most research has focused on categorical attributes in relational table whereas in practice, this table contains also numeric attributes. In this paper, we propose[1] to experiment *QuantMiner*, a genetic-based algorithm software for mining quantitative association rules on a atherosclerosis dataset. We give some experimental results obtained in both the description and the characterization of this disease.

Keywords: data mining, association rule, numeric attribute, discretization, categorical attribute, description, discrimination, evolutionary algorithm.

1 Introduction

In this paper, we focus on mining descriptive and characteristic rules in the context of the Stulong project.

"The study (STULONG) 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. M. Aschermann, MD, SDr, FESC), under the supervision of Prof. F. Boud ik, MD, ScD, with collaboration of M. Tomeckova, MD, PhD and Ass. Prof. J. Bultas, MD, PhD. 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. Zvarova, DrSc). The data resource is on the web pages `http://euromise.vse.cz/STULONG`. At present time the data analysis is supported by the grant of the Ministry of Education CR Nr LN 00B 107."

As stated in the description of this project, Stulong is a dataset concerning a twenty years lasting study of the risk factors of the atherosclerosis in a population of 1 419 middle aged men. The patients have been classified into three groups, namely the normal group, the risk group and the pathological group, and experts [1] have defined analytic questions related to the entry examination and to the long-term observation.

For such questions, descriptive tasks seem to be appropriated, since they aim at mining properties shared by "many" objects of a given group.

[1] This work is the fruit of our participation to the CNRS' "Action Spécifi que DiscoChallenge" animated by B. Crémilleux. We wish to thank all the participants for their insightful comments and discussions during our meetings. Thanks also for those who preprocessed the datasets.

An association rule [2] is an expression $\mathcal{C}_1 \Rightarrow \mathcal{C}_2$, where $\mathcal{C}_1$ and $\mathcal{C}_2$ express conditions on the attributes describing the objects.

The strength of a rule is usually evaluated by means of statistical measures, as for instance the *support* and the *confidence*, defined as follows:

- Support($\mathcal{C}$), where $\mathcal{C}$ expresses conditions on attributes, is the number of tuples in $\mathcal{T}$ that satisfy $\mathcal{C}$.
- $Support(\mathcal{C}_1 \Rightarrow \mathcal{C}_2) = Support(\mathcal{C}_1 \wedge \mathcal{C}_2)$
- $Confidence(\mathcal{C}_1 \Rightarrow \mathcal{C}_2) = \frac{Support(\mathcal{C}_1 \wedge \mathcal{C}_2)}{Support(\mathcal{C}_1)}$

Given two thresholds MINSUPP and MINCONF, a rule is *strong*, when its support is greater than MINSUP and its confidence greater than MINCONF.

Mining association rules involving categorical and numeric attributes [3–5], also called *quantitative association rules*, has been less studied. In this case, a condition is of the form $attribute = v_i$ or $attribute \in [l_i, u_i]$. For example, the rule:

$$\text{SUPER\_GROUP} = Normal \Longrightarrow \text{BMI} \in [22.5, 27.7]\ (17\% - 89\%)$$

means that 89% of people belonging to the normal group has a body mass index between the values 22.5 and 27.7. Moreover people having these 2 characteristics represent 17% of the database.

In that context, we aim at experimenting QuantMiner a tool developed by C. Nortet [6] at LIFO and BRGM for mining quantitative association rules.

2 QuantMiner briefly...

QuantMiner[2] is a genetic-based algorithm for mining quantitative association rules. It works directly on a set of rules patterns and looks for the best intervals for numeric attributes using an evolutionary-based algorithm.

In that framework, an individual is a set of couples $< attribute_i, [l_i, u_i] >$, where $attribute_i$ is the i^{th} numeric attribute in the rule pattern from the left to the right.

The algorithm starts with a set of randomly instantiated rules. During the process, genetic operators, mainly *mutation* and *crossover* are used in order to transform a given generation of rules into another one, improving its quality. The crossover operator consists in taking two individuals, called parents, at random and generating new individuals where for each attribute the interval is either inherited from one of the parents or formed by mixing intervals of the two parents. Mutation works on only one individual and modifies values in intervals (by increasing the lower bound or decreasing the upper one). Let us notice that the set of frequent categorical values is computed at the beginning of the process to define rule conditions on categorical attributes. This step is achieved using *apriori*[2].

[2] QuantMiner has been successfully used in another application related to geosciences at the BRGM - French Geological Survey, Mineral Resources Division.

3 Dataset Description

Table 1 shows the set of attributes of the table Entry_Dead we have used, their types (numeric or Categorical) and their descriptions. This table has been built using the table Entry preprocessed by Lucas et al. [7], the table Entry given in the PKDD challenge and the table Death. The size of the search space for the quantitative association rules depends on the number of attributes and their domain sizes. Without restrictions on the possible forms of the rule, the search space may become so large that the mining process is intractable and would lead to a huge number of rules.

For this reason, in our experiments, we have focused [3] on rule patterns such that the antecedent of the rule describes a subset of the population of patients, and the consequent gives a description and better a characterization of the population specified in the antecedent. Descriptions are of the form $attribute = v_i$ or $attribute \in [l_i, u_i]$. In other words, rules are of the form: $target\ set\ of\ patients \Longrightarrow descriptions$

Attribute	Type	Description	Attribute	Type	Description
ICO	C	Identifi cation of a patient	MOC_ALB	C	Urine albumen
ACTIV_JOB	C	Physical activity in a job	BOLHR	C	Chest pain
ACTIV_AFT	C	Physical activity after a job	CHLST	N	Cholesterol in mg%
TRANSP_JOB	C	Transport to go to work	TRIGL	N	Triglycerides in mg%
TIME_JOB	C	Time to get to work	SYST	N	Blood pressure systolic
BIRTH_YEAR	N	Year of birth	DIAST	N	Blood pressure diastolic
AGE	N	Age of the patient	HEIGHT	N	Height (cm)
ENTRY_YEAR	N	Year of entry into the study	WEIGHT	N	Weight (kg)
ALCO_CONS	N	Alcohol consumption	BMI	N	Body Mass Index
TOBA_CONS	N	Tobacco consumption	TRIC	N	Skin fold triceps (mm)
TOBA_DURA	N	Smoking duration	SUBSC	N	Skin fold subscapularis (mm)
MARIT_STAT	C	Marital status	RSK_FAMI	C	Family risk
EDUCATION	C	Reached education	RSK_OBES	C	Obesity risk
IM	C	Myocardial infarction	RSK_TOBA	C	Smoking risk
ICT	C	Ictus	RSK_HYPE	C	Hypertension risk
HT	C	Hypertension	RSK_CHOL	C	Cholesterol risk
HTL	C	Medicines in HT	GROUP	C	Group: 1..6
DIAB	C	Diabetes	SUPER_GROUP	C	Super Group: N, R, P
DIABD	C	Diet in DIAB	DEATH?	C	Patient dead?
HYPL	C	Hyperlipidemia	DEATH_YEAR	N	Year of death
HYPLL	C	Medicines in hyperlipidemia	DEATH_CAUS	C	Cause of death
MOC_SUC	C	Urine sugar	DEATH_DELTA	N	#year between entry and death

Table 1. Attributes of the table Entry_Dead

4 Experiments on the Atherosclerosis Dataset

We have experimented QuantMiner on the table Entry_Dead described above. Particularly, we are interested in the rules of the form:

$$DEATH? = value \Longrightarrow descriptions$$
$$SUPER\_GROUP = value \Longrightarrow descriptions$$

where descriptions are conjonctions of conditions $attribute = v_i$ or $attribute \in [l_i, u_i]$.

[3] Let us notice again that this is not a limitation in QuantMiner, but we do this for the reasons mentioned above.

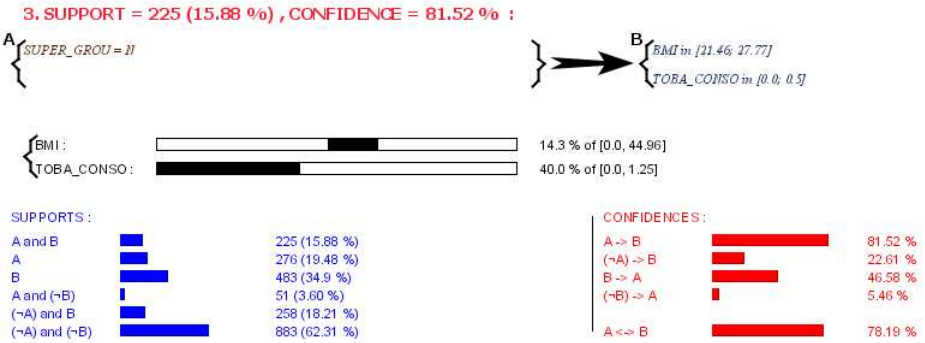

Fig. 1. Example of rules mined by QuantMiner

In other words, we have searched for rules containing only the attribute DEATH? or SUPER_GROUP in the left handside and some attributes among: ALCO_CONS, TOBA_CONSO, TOBA_DURA, BMI, EDUCATION, MARITAL_STATUS, AGE, etc. in the right hand side. Tables 2, 3, 4, 5 give a set of selected rules[4] obtained in Entry_Dead dataset. Figure 1 shows an example of rules as actually displayed in *QuantMiner*.

Characterization of super-groups by alcohol, tobacco consumption and BMI: Table2.

N	Rule ($A \Rightarrow B$)	Supp($A \Rightarrow B$)	Conf($A \Rightarrow B$)	Conf($\neg A \Rightarrow B$)
1	$SUPER\_GROUP = N \Longrightarrow ALCO\_CONS \in [1.0; 1.2] \wedge BMI \in [19.73; 27.77] \wedge TOBA\_CONSO \in [0.0; 0.5]$	13%	69%	19%
2	$SUPER\_GROUP = R \Longrightarrow ALCO\_CONS \in [1.0; 1.29] \wedge BMI \in [22.28; 30.72] \wedge TOBA\_CONSO \in [0.5; 1.25] \wedge TOBA\_DURA \in [15.0; 20.0]$	39%	64%	38%
3	$SUPER\_GROUP = P \Longrightarrow ALCO\_CONS \in [1.0; 1.53] \wedge BMI \in [21.98; 33.14] \wedge TOBA\_CONSO \in [0.5; 1.25]$	5%	64%	60%
4	$DEATH? = yes \Longrightarrow ALCO\_CONS \in [1.0; 1.28] \wedge TOBA\_CONSO \in [0.5; 1.25] \wedge TOBA\_DURA \in [15.0; 20.0]$	18%	68%	53%
5	$DEATH? = no \Longrightarrow ALCO\_CONS \in [1.0; 1.23] \wedge TOBA\_CONSO \in [0.0; 0.85]$	49%	67%	55%

Table 2. Characterization of super groups/Death? by alcohol, tobacco consumption and BMI

[4] Further rules are available on `http://www.univ-orleans.fr/lifo/Members/salleb/Challenge2004/`.

Characterization of super-groups by social characteristics: No interesting rules found for the time being.

Characterization of super-groups by biochemical examination: Table 3.

N	Rule ($A \Rightarrow B$)	Supp($A \Rightarrow B$)	Conf($A \Rightarrow B$)	Conf($\neg A \Rightarrow B$)
9	$SUPER\_GROUP = N \Longrightarrow MOC\_ALBUMI = 0 \wedge MOC\_SUCRE\_ = 0 \wedge CHLST \in [149.0; 259.0] \wedge TRIGL \in [56.0; 195.0]$	15%	78%	53%
10	$SUPER\_GROUP = R \Longrightarrow MOC\_ALBUMI = 0 \wedge MOC\_SUCRE\_ = 0 \wedge CHLST \in [179.0; 305.0] \wedge TRIGL \in [89.0; 299.0]$	43%	71%	66%
11	$SUPER\_GROUP = P \Longrightarrow MOC\_ALBUMI = 0 \wedge MOC\_SUCRE\_ = 0 \wedge CHLST \in [200.0; 353.0] \wedge TRIGL \in [97.0; 416.0]$	5%	70%	63%

Table 3. Characterization of super groups by biochemical examination

Characterization of super-groups by physical examination: Table 4

N	Rule ($A \Rightarrow B$)	Supp($A \Rightarrow B$)	Conf($A \Rightarrow B$)	Conf($\neg A \Rightarrow B$)
12	$SUPER\_GROUP = R \Longrightarrow SUBSC \in [8.0; 29.0]$	48%	80%	73%
13	$SUPER\_GROUP = R \Longrightarrow TRIC \in [4.0; 14.0]$	48%	79%	73%
14	$SUPER\_GROUP = N \Longrightarrow TRIC \in [3.0; 12.0]$	15%	78%	71%
15	$SUPER\_GROUP = N \Longrightarrow SUBSC \in [7.0; 24.0]$	14%	76%	68%
16	$SUPER\_GROUP = P \Longrightarrow TRIC \in [5.0; 20.0]$	6%	75%	80%
17	$SUPER\_GROUP = P \Longrightarrow SUBSC \in [10.0; 29.0]$	5%	70%	74%
18	$SUPER\_GROUP = N \Longrightarrow DIAST \in [60.0; 92.0]$	13%	69%	55%
19	$SUPER\_GROUP = P \Longrightarrow SYST \in [105.0; 195.0]$	5%	67%	65%
20	$SUPER\_GROUP = R \Longrightarrow SYST \in [95.0; 175.0]$	40%	66%	65%
21	$SUPER\_GROUP = P \Longrightarrow DIAST \in [60.0; 110.0]$	5%	65%	66%
22	$SUPER\_GROUP = R \Longrightarrow DIAST \in [60.0; 110.0]$	39%	64%	69%
23	$SUPER\_GROUP = N \Longrightarrow SYST \in [95.0; 150.0]$	12%	64%	58%

Table 4. Characterization of super groups by physical examination

Characterization of super-groups by physical ativities: Table 5.

N	Rule ($A \Rightarrow B$)	Supp($A \Rightarrow B$)	Conf($A \Rightarrow B$)	Conf($\neg A \Rightarrow B$)
24	$SUPER\_GROUP = N \Longrightarrow ACTIV\_AFT = 2 \wedge BMI \in [19.94; 27.98]$	13%	69%	51%
25	$SUPER\_GROUP = N \Longrightarrow TIME\_JOB = 5 \wedge BMI \in [20.37; 27.77]$	13%	67%	47%
26	$SUPER\_GROUP = R \Longrightarrow ACTIV\_AFT = 2 \wedge BMI \in [21.56; 31.41]$	39%	65%	65%
27	$SUPER\_GROUP = P \Longrightarrow ACTIV\_AFT = 2 \wedge BMI \in [23.78; 33.61]$	5%	64%	54%
28	$SUPER\_GROUP = R \Longrightarrow TIME\_JOB = 5 \wedge BMI \in [20.52; 31.49]$	38%	64%	62%

Table 5. Characterization of super groups by physical activities

5 Conclusion

In this paper, we have presented an application of mining quantitative association rules in the atherosclerosis dataset proposed in the PKDD Challenge 2004. Quantitative association rules have the nice feature to handle both categorical and numeric attributes.

QuantMiner is an interesting tool for mining descriptive rules. Nevertheless, it should be extended to take into account discrimination measures during the mining process. We have also developped a relational characterization tool named *CaracteriX* [8] which can handle multi-instance datasets. In order to experiment it to the *Control* table, we are currently extending it to deal with temporal data.

References

1. Tomeckov, M., Rauch, J., Berka, P.: Data from a Longitudinal Study of Atherosclerosis Risk Factors. In: ECML/PKDD 2002 Discovery Challenge Workshop program. (2002)
2. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In Buneman, P., Jajodia, S., eds.: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C. (1993) 207–216
3. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In Jagadish, H.V., Mumick, I.S., eds.: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada (1996) 1–12
4. Rastogi, R., Shim, K.: Mining optimized support rules for numeric attributes. In: Proceedings of the 15th International Conference on Data Engineering, 23-26 March 1999, Sydney, Austrialia, IEEE Computer Society (1999) 206–215
5. Aumann, Y., Lindell, Y.: A statistical theory for quantitative association rules. In: Knowledge Discovery and Data Mining. (1999) 261–270
6. Nortet, C.: Extraction de Règles d'Association Quantitatives. Master's thesis, BRGM and LIFO University of Orléans (2003)
7. Lucas, N., Az, J., Sebag, M.: Atherosclerosis Risk Identifi cation and Visual Analysis. In: ECML/PKDD 2002 Discovery Challenge Workshop program. (2002)
8. Turmeaux, T., Salleb, A., Vrain, C., Cassard, D.: Learning Characteristic Rules Relying on Quantifi ed Paths. In et al., N., ed.: 7th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD), Springer–Verlag, Lecture Notes in Computer Science (2003) 471–482