Mining for association rules in medical data

Jean-Marc Trémeaux, Yan Liu

Department of Informatics and Statistics, Université Lumière Lyon 2
Faculté de Sciences Économiques et de Gestion
5 avenue Pierre Mendès-France, 69676 BRON Cedex
jm.tremeaux@gmail.com
louisliu_fr@yahoo.fr

Abstract

Association rules are logical propositions of the form $\varphi \approx \psi$ that are frequently observed in a given set of data. An analyst can express hypotheses in the form of these rules and verify their validity in the data, as well as find hints to relevant, previously unknown relationships. In this paper, we propose to answer some analytical questions about STULONG, a study of risk factors of atherosclerosis among a population of Czech patients. We use the 4ft-Miner application, a procedure part of the GUHA method, and present some results.

1 Introduction

STULONG[1] is an epidemiologic study carried out in order to elaborate the risk factors of atherosclerosis in a population of middle aged men. Atherosclerosis is a disease more frequent among men, developping slowly and involving multiple causes such as high blood pressure, dyslipidemia (high cholesterol or triglyceride levels), alcohol consumption and tobacco smoking. Thus the need for a long term, transversal study in order to characterize the most important risk factors and their interrelationship. This kind of study lends pretty well to a data mining methodology, and thus has been used as the basis of the discovery challenge in ECML/PKDD 2004. Some tasks of this challenge were to answer STULONG analytical questions such as *"what is the relation of physical activity to blood pressure?"*, and to discover unsuspected relationships.

Models of association rules have been proposed to describe relationships that are frequently observed in a data set. These rules are in the form $\varphi \approx \psi$, where φ and ψ are combinations of categorial attribute values and $\approx$ a quantifier. As an example, the rule $BeerConsumption(High) \Rightarrow BMI(Normal, High)$ express the facts that heavy beer consumers have either normal or high body mass index. The theoretical aspects of these models have been studied in GUHA[2], and practical procedures such as 4ft-Miner[5] are used to automatically extract a set of relevant rules.

[1] The study (STULONG) was realized at the 2^{nd} Department of Medicine, 1^{st} 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. Boudík, MD, ScD, with collaboration of M. Tomečková, 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. Zvárová, DrSc). The data resource is on the web pages http://euromise.vse.cz/challenge2004. At present time the data analysis is supported by the grant of the Ministry of Education CR Nr LN 00B 107.

The purpose of our work is to conduct an analysis of the STULONG data, using association rules models in order to answer simple analytical questions. As a methodologic guide, we use the widespread CRISM-DM[1] process model. It is an iterative process composed of six important phases of data-mining as described in figure 1. The plan of this paper roughly follows one iteration of the whole process. In section 2, we present the background knowledge related to the STULONG study and the objectives we want to achieve. In section 3, we describe the data and its preparation (acquisition and transformation). In section 4, we introduce the model of association rules in details. In section 5, we present some results of our experiments in modeling. Finally, in section 6 we make some concluding remarks.

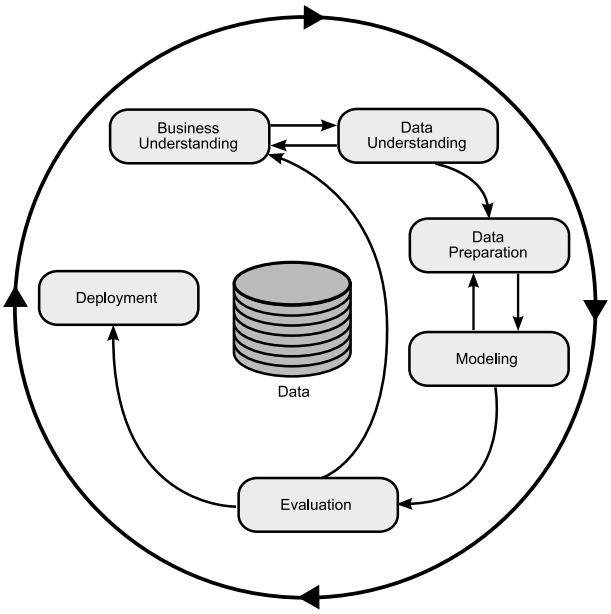

Figure 1: *Overview of the CRISP-DM process*

2 Background knowledge

In the years 1970's, a broad study of atherosclerosis cardiovascular disease was conducted in the former Czechoslovakia. The aims of this study were multiple:

- Identify the risk factors involved in the disease.
- Follow the development of these risk factors and their impact on health.
- Investigate the relationship of these risk factors with cardiovascular morbidity and mortality.

The target of the study was a population of 2370 middle-aged men, living in Prague. This part of the population was considered to be at risk concerning atherosclerosis disease. The protocol consisted of a first examination, followed up by a 10-12 years long observation and pharmacological treatment for some of them. Upon invitation for the first examination, 1419 men showed up. They were segmented into 3 groups – normal, at risk, pathologic – according to the results of the examination.

The primary goal of our work is to answer four analytical questions stated in the ECML/PKDD 2004 discovery challenge. All of these questions are related to the entry examination. We want to identify the relations of the following characteristics of men in the respective groups:

- alcohol consumption and smoking.
- alcohol consumption and BMI (Body mass index).
- alcohol consumption and blood pressure.
- alcohol consumption and level of total cholesterol HDL cholesterol, triglycerides.

Aside from this systematic approach of identifying the relationship between characteristic attributes, we want to test some hypotheses about atherosclerosis that can be found in the literature. Our work consists of finding such hypotheses in medical the literature that are relevant to STULONG, reformulating them in terms of association rules, and then mining for such associations in the STULONG data.

Such knowledge is expressed in textual form, like *heavy drinking induces overweight*, *90% of alcoholics smoke too* or *smoking increases the craving for alcohol*, that can be translated by the analyst into formal propositions like $Smoking(True) \Rightarrow AlcoholIntake(Normal, High)$.

One such hypothesis found in the literature [3] is about the relationship between alcohol intake and blood pressure among young adults. It has been observed that the lowest systolic pressure levels is in subjects consuming 1 to 2 drinks per day. Among abstainers, systolic pressure is higher by 4.0 mm Hg in those who drank < 1 drink per day or 2 to 3 drinks per day, and 8.1 mm Hg in those who drank ≥ 3 drinks per day. In summary, these results suggested a J-shaped relationship of alcohol intake with blood pressure, with the lowest levels in consumers of 1-3 drinks per day. We want to find out how this former knowledge relates to the STULONG data.

Our final objective is to investigate the dissimilarity in different groups of the population – normal, at risk, pathologic –. These differences can be expressed in terms of relationships that are more prevalent in some groups, and then evaluated in order to better understand the quality, structure and properties of the segmentation. Medical decisions have been taken on the basis of group membership, *e.g.* treatment and thorough examination in the risk group. That's why a quantitative, well-justified approach to group formation is of prime matter.

3 Data understanding and preparation

The analytical tasks we want to address are data-driven, which means that the studied data have been collected in the 1970's, and our objectives were defined *a posteriori*, based on the data at hand. Therefore, an important work has to be done on understanding the data, selecting the information relevant to our study and preparing it for the modeling methods we use.

The STULONG data is in form of two relations, *entry* and *control*. The relation *entry* describes the results of the first examination of 1419 pations, while *control* describes the 20 years long follow-up of some patients.

In this study, we only consider the *entry* relation. Each patient examination is described by 244 attributes, grouped into 11 families: identification, social characteristics, physical activity, smoking, drinking of alcohol, sugar / coffee / tea, family anamnesis, personal anamnesis, questionnaire A_2, physical examination and biochemical examination. Each attribute can have either categorial or continuous values.

Of all 244 attributes, we have manually selected only those attributes that were the most relevant to our analytical goals, namely 5 groups:

- Identification: identification number, group.
- Physical examination: height, weight, blood pressure (systolic I, II, diastolic I, II), skinfold above musculus triceps, skinfold above musculus subscapularis.
- Alcohol intake: frequency of drinking, beer 7°, beer 10°, beer 12°, wine, liquors, amount of beer / wine / liquors per day.
- Smoking: frequency of smoking, smoking for how many years, stopped smoking for how many years.
- Biochemical examination: cholesterol, triglycerides.

It should be remarked that manually selecting attributes introduces a bias in our study, as some more attributes (like age) not considered here could have proved relevant to our analytical questions. But as we shall see, the modeling procedure we use allows us to express associations in terms of many attributes, and only retain the most discriminative.

Most of those attributes are self-explanatory, but some of them need particular attention and will be discussed here. *Group* is a special attribute that map each patient to one of three pre-defined groups. Normal group is a group of men without any risk factor or apparent cardiovascular disease. Risk group is a group of men that exhibit at least one risk factor (such as obesity or familial antecedants). Pathological group is a group of men with manifestation of cardiovascular or another serious disease. Grouping similar people has some benefits for the analysis:

- Alleviate data exploration by exhibiting common properties among similar people.
- Exhibit dissimilarities between distinct groups (*e.g.* relationships that are verified in one group and not one another.
- Take action (*e.g.* further examination or medical treatment) based on these properties.

One new attribute BMI had to be constructed to investigate an analytical question:

$$BMI = \frac{weight\ in\ kg}{(height\ in\ m)^2}$$

The attributes relative to physical and biochemical examination are measures taking real values, so they had to be discretized as association rules only deals with categorial attributes. For this task, three main approaches were made:

- Based on field knowledge: if some agreement exists on possible attribute values, use it. Thus, BMI was discretized into pre-defined categories (underweight, ideal, overweight, obese) with boundaries agreed upon.
- Equidistant intervals: used for attributes that don't have normalized values.
- Intervals of equal frequency: used for very dissymetric distributions.

Please note that some attributes may have normalized values (*e.g.* an ideal systolic pressures rests in the range 90-135 mm Hg), but for the sake of knowledge extraction, it was more useful to use a finer grained discretization. Dealing with the number and size of intervals was a complex task, as the quality of discretization influences the quality of the produced models, which in turns hints at a better discretization. Lots of small intervals leads to a better adequation of the model to the data, but with the drawback of a larger search space to explore, risks of overfitting and more difficult model understanding. Tables 1 and 2 show two discretizations that were used for systolic pressure.

Category	Frequency	Description
<0;110)	69	low
<110;135)	741	normal
<135;+inf)	607	high

Table 1: Coarse discretization of Syst1 based on field knowledge

Category	Frequency
<90;100)	7
<100;110)	62
<110;120)	207
<120;130)	305
<130;140)	311
<140;150)	249
<150;160)	142
<160;170)	58
<170;180)	43
<180;190)	13
<190;200)	7
<200;210)	8
<210;220)	4
<220;230)	1

Table 2: Fine discretization of Syst1 into intervals of 10mm Hg

The opposite issue arises when attribute values are too coarse. This is the case of alcohol intake frequency, which takes only 3 possible values (never, occasionaly, frequently). For the sake of examining the relationship of alcohol intake to blood pressure on the same ground as [3], we would need to construct a new attribute *alcohol intake in g/day*, which requires some additional effort on data understanding.

4 Association rules and the 4ft-Miner procedure

Association rules are propositions of the form $\varphi \approx \psi$ that are true for a set of observed individuals. Let $\mathcal{M}$ the studied data matrix shown in Table 3. $\mathcal{M}$ is composed of a set of observations $o_1, ..., o_n$ described by the attributes $A_1, ..., A_K$.

A *basic boolean attribute* is a predicate $A(\omega)$, where A is an attribute and ω a subset of the values taken by the attribute A. $A(\omega)$ is evaluated to true for the observation o if and only if o has the value a for the attribute A, and $a \in \omega$. As an example, the basic boolean attribute $A_1(v_1, v_3)$ is true for observation o_1 if and only if $a_{1,1} = v_1$ or $a_{1,1} = v_3$.

The *derived boolean attribute* φ and ψ are conjunctions of basic boolean attributes.

The association rule $\varphi \approx \psi$ is evaluated on the basis of a contigency table called *4ft table*, represented in Table 4. In this matrix, a is the number of observations satisfying both φ and ψ in that data matrix $\mathcal{M}$, b is the number of observations satisfying φ and not satisfying ψ, etc.

observation	A_1	A_2	...	A_K	φ	ψ
o_1	$a_{1,1}$	$a_{1,2}$	...	$a_{1,K}$	1	0
.	.	.	...	.	.	.
.	.	.	...	.	.	.
.	.	.	...	.	.	.
o_n	$a_{n,1}$	$a_{n,2}$	...	$a_{n,K}$	0	1

Table 3: Data matrix $\mathcal{M}$

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

Table 4: 4ft table of $\mathcal{M}$, φ and ψ

Various types of dependencies of φ and ψ can be expressed by using the suitable conditions on the 4ft table. Each condition is associated to a *4ft quantified* $\approx$. We say that $\varphi \approx \psi$ is true in the data matrix $\mathcal{M}$ if the corresponding conditions in the 4ft table of the data matrix $\mathcal{M}$ for derived attributes φ and ψ are satisfied.

Here is a list of 4ft quantifiers:

- **founded implication** $\Rightarrow_{p,base}$ for $0 < p \leq 1$ and $base > 0$ is defined by the condition $\frac{a}{a+b} \geq p \wedge a \geq base$. It means that at least $100p$ percent of observations satisfying φ also satisfy ψ, and that there are at least $base$ observations in $\mathcal{M}$ satisfying both φ and ψ.

- **founded double implication** $\Leftrightarrow_{p,base}$ for $0 < p \leq 1$ and $base > 0$ is defined by the condition $\frac{a}{a+b+c} \geq p \wedge a \geq base$. It means that at least $100p$ percent of observations satisfying φ or ψ also satisfy both φ and ψ, and that there are at least $base$ observations in $\mathcal{M}$ satisfying both φ and ψ.

- **p-equivalence** $\equiv_{p,base}$ for $0 < p \leq 1$ and $base > 0$ is defined by the condition $\frac{a+d}{a+b+c+d} \geq p \wedge a \geq base$. It means that φ and ψ have the same value (either true or false) for at least $100p$ percent of observations in $\mathcal{M}$, and that there are at least $base$ observations in $\mathcal{M}$ satisfying both φ and ψ.

- **above average implication** $\Rightarrow^{+}_{p,base}$ for $0 < p$ and $base > 0$ is defined by the condition $\frac{a}{a+b} \geq (1+p)\frac{a+c}{a+b+c+d} \wedge a \geq base$. It means that among the obervations satisfying φ, there are at least $100p$ percent more observations satisfying ψ than there are observations satisfying ψ in the whole data matrix, and that there are at least $base$ observations in $\mathcal{M}$ satisfying both φ and ψ.

In addition, the relationship between attributes can be expressed as a **conditional association rule**, that is a proposition of the for $\varphi \approx \psi/\chi$, where φ, ψ and χ are derived boolean attributes. This means that the relation $\varphi \approx \psi$ is investigated only among the individuals matching the condition χ.

The procedure 4ft-Miner takes in input the data matrix $\mathcal{M}$, and a pattern composed of the attributes and their possible values to be investigated. It uses an efficient bitstring-based algorithm to return the set of all potentially interesting association rules.

An additional procedure, SD4ft-Miner [6] (for Set Differs) compares the 4ft-table of two groups defined by the conditions χ_1 and χ_2, in order to tell how much they differ. Given the 4ft-tables (a_1, b_1, c_1, d_1) and (a_2, b_2, c_2, d_2), we can for example define the condition $|\frac{a_1}{a_1+b_1} - \frac{a_2}{a_2+b_2}| \geq \delta \wedge a_1 \geq base_1 \wedge a_2 \geq base_2$, which express the comparison of a founded implication in two groups. This condition will only be evaluated to true for rules that express a sufficiently dissimilar distribution in the two groups.

5 Results

In this section, we present and discuss some of the models mined by the 4ft-Miner and SD4ft-Miner procedures. The models address 9 analytical tasks shown in Table 5.

Id	Task Type	Task name
T_1	Basic	Alcohol(?) $\approx$ Smoking(?)
T_2	Basic	Alcohol(?) $\approx$ Blood pressure(?)
T_3	Basic	Alcohol(?) $\approx$ BMI(?)
T_4	Basic	Alcohol(?) $\approx$ Cholesterol(?), triglycerides(?)
T_5	Specific	Alcohol(?) $\approx$ Blood pressure(?)
T_6	Set Difference	Alcohol(?) $\approx$ Smoking(?)
T_7	Set Difference	Alcohol(?) $\approx$ Blood pressure(?)
T_8	Set Difference	Alcohol(?) $\approx$ BMI(?)
T_9	Set Difference	Alcohol(?) $\approx$ Cholesterol(?), triglycerides(?)

Table 5: Task list

Basic tasks T_1, T_2, T_3 and T_4 are systematic investigations of relationships between 2 groups of attributes. The specific task T_5 address an hypothesis about the relationship of alcohol intake with blood pressure described in section 2. The SD tasks T_4–T_9 underline some dissimilarities between groups *Normal* and *Risk* related to the same attributes as in T_6, T_7, T_8 and T_9.

The tested rules are evaluated in terms of their internal validity in the data (sufficient p and *base*, and not subsumed by a simple rule), and their relevance to the studied analytical task. Only the most relevant rules were selected and discussed here.

Association rule	p	*base*	*avgd*
Alcohol(Regularly) $\Rightarrow$ $\neg$Daily smoking(0)	0.766	354	
Alcohol(Occasionaly) $\wedge$ Daily liquor(no) $\wedge$ Beer 10°(No) $\wedge$ Beer 12°(No) $\Rightarrow$ Daily smoking(0) / Group(Normal)	0.72	18	
Alcohol(Regularly) $\wedge$ Beer consumtion(1+L) $\wedge$ Daily liquor(0-100cc) $\wedge$ Wine consumption(0-0.5L) $\wedge$ Beer 12°(No) $\Rightarrow$ Daily smoking(15-20) / Group(Risk)	0.85	11	
Alcohol(Regularly) $\wedge$ Beer consumtion(1+L) $\wedge$ Daily liquor(0-100cc) $\wedge$ Wine consumption(0-0.5L) $\wedge$ Beer 10°(Yes) $\Rightarrow$ Daily smoking(15-20) / Group(Pathologic)	0.69	11	
Beer consumtion(1+L) $\wedge$ Daily liquor(0-100cc) $\wedge$ Wine consumption(no) $\wedge$ Beer 12°(No) $\Rightarrow^{+}$ Daily smoking(21+)	0.48	34	0.98

Table 6: Task T_1

Rules in Table 6 investigate the relationship between alcohol intake and smoking. The first rule express the knowledge that people drinking regularly smoke too. Among the population, 77% of the people satisfy this rule, with a base of 354 people drinking alcohol regularly. The second rule states that people in the normal group drinking alcohol occasionnaly, but no liquor or 10°–12° beer (they could be drinking lighter beer or wine) don't smoke. The third and fourth rules state that heavy drinkers in the normal and pathologic group also smoke regularly. Finally, the fifth rule states

that among heavy drinkers, there are 91% more people that smoke 21 cigarettes or more per day than in the whole population.

Association rule	p	$base$	$avgd$
Alcohol(Occasionaly) ∧ Wine consumption(no) ∧ Daily liquor(0-100cc) ∧ Beer 7°(No) $\Rightarrow^{+}$ Diast1(low)	0.3	23	0.75
Alcohol(Regularly) ∧ Daily liquor(0-100cc, 100+cc) ∧ Beer consumption(1+L) ∧ Beer 10°(Yes) $\Rightarrow^{+}$ Syst1(high)	0.69	20	0.61

Table 7: Task T_2

Rules in Table 7 investigate the relationship between alcohol intake and blood pressure. The first rule states that occasional drinkers have lower diastolic pressure than on average. The second rules states that heavy drinkers have higher systolic pressure than on average. Please note that only 3 values for blood pressure were taken into accound (low, normal, high) based on medical knowledge. The relationship between alcohol intake and blood pressure is addressed with a finer grain in task T_5.

Association rule	p	$base$	$avgd$
Alcohol(Occasionaly) ∧ Wine consumption(0-0.5L) ∧ Daily liquor(no) ∧ Beer 7°(No) ∧ Beer 10°(Yes) ∧ Beer 12°(No) ⇒ BMI(Ideal) / Group(Normal)	0.83	10	
Beer consumption(no, 0-1L) ∧ Daily liquor(0-100cc) ∧ Beer 12°(No) ⇒ BMI(Overweight) / Group(Risk)	0.64	152	
Alcohol(Occasionaly) ∧ Wine consumption(0-0.5L) ∧ Daily liquor(no) ∧ Beer 7°(No) ∧ Beer 10°(Yes) ∧ Beer 12°(No) $\Rightarrow^{+}$ BMI(Ideal) / Group(Normal)	0.83	10	0.56

Table 8: Task T_3

Rules in Table 8 investigate the relationship between alcohol intake and BMI. The first rule states that occasional drinkers of wine or 10° beer in the normal group have an ideal BMI. The second rules states that people consuming up to 1L of 7° or 10° beer and up to 100cc of liquor in the risk group tend to be overweight. The last rule states that occasional drinkers of wine and 10° beer in the normal groups are fitter than on average.

Association rule	p	$base$	$avgd$
Beer consumption(no) ∧ Daily liquor(0-100cc) ∧ Wine consumption(0-0.5L) ⇒ Trigryceride(<28;178))	0.82	47	
Alcohol(Occasionaly) ∧ Beer 12°(No) ∧ Beer consumption(0-1L) ∧ Daily liquor(0-100cc) $\Rightarrow^{+}$ Cholesterol(<112;232)) ∧ Trigryceride(<178;328))	0.18	25	1.12

Table 9: Task T_4

Rules in Table 9 investigate the relationship between alcohol intake and biochemicals. The desirable range for triglycerides stand below 150 mg/dL, and for cholesterol below 200 mg/dL (ranges set by the American Heart Association [4]). The first rule states that low drinkers of liquor and wine have triglyceride levels in the normal range. The second rule states that occasional drinkers of beer and liquor have borderline levels of cholesterol and triglicerides more than on average.

Rules in table 10 address the specific analytical task of observing the systolic pressure levels in relation to alcohol intake. The first rule states that among people

Association rule	p	$base$	$avgd$
Beer consumption(no) ∧ Daily liquor(no) ∧ Wine consumption(no) $\Rightarrow^+$ Syst_10(<140;150)...<170;180))	0.41	31	0.18
Alcohol(Occasionaly) ∧ Daily liquor(0-100cc) $\Rightarrow^+$ Syst_10(<90;100), <100;110))	0.06	23	0.30
Alcohol(Regularly) $\Rightarrow^+$ Syst_10(<170;180), <180;190))	0.06	29	0.59

Table 10: Task T_5

that don't drink alcohol, average to high ranges are observed more frequently than in the whole population. The second rule states that among people drinking liquor occasionaly, there are 30% more people having low systolic pressure than in the whole population. Finally, the third rule states that among people drinking alcohol regularly, higher systolic pressure levels are observed than on average. These rules seem to go in the direction of the medical hypothesis, with higher blood pressure ranges observed for people consuming alcohol never or regularly.

Association rule	p (group 1)	p (group 2)
Alcohol(Occasionaly) ∧ Beer 10°(No) ∧ Beer 12°(No) ∧ Beer consumption(no, 0-1L) ∧ Daily liquor(no) ⇒ Daily smoking(0)	0.72	0.18

Table 11: Task $T_6 : Group(Normal) \times Group(Risk)$

The rule in table 11 express a difference between the people in normal and risk group. People in the normal group (with a base similar to task T_1) drinking alcohol occasionaly tend to not smoke, a relation that is less frequently observed in this risk group.

Association rule	p (group 1)	p (group 2)
Alcohol(Regularly) ∧ Beer 12°(No) ∧ Beer consumption(no, 0-1L) ∧ Daily liquor(0-100cc) ⇒ Syst2(normal)	0.90	0.49

Table 12: Task $T_7 : Group(Normal) \times Group(Risk)$

The rule in table 12 states that people drinking alcohol regularly have a normal systolic pressure more frequently in the normal group than in the risk group.

Association rule	p (group 1)	p (group 2)
Alcohol(Occasionaly) ∧ Beer 10°(Yes) ∧ Beer 12°(No) ∧ Beer 7°(No) ∧ Daily liquor(no) ∧ Wine consumption(0-0.5L) ⇒ BMI(Ideal)	0.83	0.32

Table 13: Task $T_8 : Group(Normal) \times Group(Risk)$

The rule in table 13 states that people drinking beer or wine occasionaly have an ideal BMI more frequently in the normal group than in the risk group.

The rule in table 14 states that people drinking alcohol regularly have a cholesterol level in the normal range more frequently in the normal group than in the risk group.

Association rule	p (group 1)	p (group 2)
Alcohol(Regularly) $\wedge$ Beer 7°(No) $\wedge$ Daily liquor(0-100cc, 100+cc) $\Rightarrow$ Cholesterol(<112;232))	0.93	0.38

Table 14: Task $T_9 : Group(Normal) \times Group(Risk)$

6 Conclusion

In this work, we have seen how association rules can be used to express medical knowledge and study its validity in data. We addressed 9 tasks, ranging from systematic discovery of potentially userful rules, to specific analytical tasks and finding the dissimilarities between subsets of the population. Some of the results might be useful to the practicians. Difficulties lied in finding good parameters of the models (especially for discretization of attributes). We used trial and error, which may be automated in part to find models of better quality and with less human interference.

References

[1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. CRISP-DM 1.0. Technical report, The CRISP-DM Consortium, 2000.

[2] P. Hájek and T. Havránek. *Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory*. Springer-Verlag, 1978.

[3] Gillman MW, Cook NR, Evans DA, Rosner B, and Hennekens CH. Relationship of alcohol intake with blood pressure in young adults. *Hypertension*, 25:1106–1110, 1995.

[4] Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, and Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*, pages 1837–1847, 1998;97.

[5] J. Rauch and M. Šimůnek. An Alternative Approach to Mining Association Rules. In Lin T.Y., Ohsuga S., Liau C.J., Hu X., and Tsumoto S., editors, *Foundations of Data Mining and Knowledge Discovery*, pages 211–232. Berlin, Springer, 2005.

[6] J. Rauch and M. Šimůnek. GUHA Method and Granular Computing. In HU, Xiaohua, LIU, Qing, SKOWRON, Andrzej, LIN, Tsau Young, YAGER, Ronald R., Zang, and Bo, editors, *Proceedings of Granular computing*, pages 630–635. Piscataway, IEEE, 2005.