Applying Domain Knowledge in Association Rules Mining Process – First Experience*

Jan Rauch and Milan Šimůnek

Faculty of Informatics and Statistics, University of Economics, Prague
nám W. Churchilla 4, 130 67 Prague 3, Czech Republic
{rauch,simunek}@vse.cz

Abstract. First experiences with utilization of formalized items of domain knowledge in a process of association rules mining are described. We use association rules - atomic consequences of items of domain knowledge and suitable deduction rules to filter out uninteresting association rules. The approach is experimentally implemented in the LISp–Miner system.

1 Introduction

One of great challenges in data mining research is application of domain knowledge in data mining process [3]. Our goal is to present first experiences with an approach to use domain knowledge in association rules mining outlined in [5]. We deal with association rules of the form $\varphi \approx \psi$ where φ and ψ are Boolean attributes derived from columns of an analyzed data matrix. An example of data matrix is in section 2. Not only conjunctions of *attribute-value* pairs but general Boolean expressions built from *attribute-set of values* pairs can be used. Symbol $\approx$ means a general relation of φ and ψ, see section 3.

We deal with formalized items of domain knowledge related to analyzed domain knowledge, see section 2. We apply the 4ft-Miner procedure for mining association rules. It deals with Boolean expressions built from *attribute-set of value*. An example of an analytical question solution of which benefits from properties of 4ft-Miner is in section 4.

The paper focuses on problem of filtering out of association rules which can be considered as consequences of given items of domain knowledge as suggested in [5]. Our approach is based on mapping of each item of domain knowledge to a suitable set of association rules and also on deduction rules concerning pairs of association rules. The approach is implemented in the LISp-Miner system which involves also the 4ft-Miner procedure. An example of its application is also in section 4. It can result in finding of interesting exceptions from items of domain knowledge in question, but the way of dealing with exceptions differs from that described in [8].

* The work described here has been supported by Grant No. 201/08/0802 of the Czech Science Foundation and by Grant No. ME913 of Ministry of Education, Youth and Sports, of the Czech Republic.

M. Kryszkiewicz et al. (Eds.): ISMIS 2011, LNAI 6804, pp. 113–122, 2011.

2 STULONG Data Set

2.1 Data Matrix Entry

We use data set STULONG concerning *Longitudinal Study of Atherosclerosis Risk Factors* [1]. Data set consists of four data matrices, we deal with data matrix *Entry* only. It concerns 1 417 patients – men that have been examined at the beginning of the study. Each row of data matrix describes one patient. Data matrix has 64 columns corresponding to particular attributes – characteristics of patients. The attributes can be divided into various groups, We use three groups defined for this paper - *Measurement*, *Difficulties*, and *Blood pressure*.

Group *Measurement* has three attributes - *BMI* i.e. Body Mass Index, *Subsc* i.e. skinfold above musculus subscapularis (in mm), and *Tric* i.e. skinfold above musculus triceps (in mm). The original values were transformed such that these attributes have the following possible values (i.e. categories):
BMI : $(16;21\rangle,\ (21;22\rangle,\ (22;23\rangle,\ \ldots,\ (31;32\rangle, > 32$ (13 categories)
Subsc : $\langle 4;10),\ \langle 10;12),\ \langle 12;14),\ \ldots,\ \langle 30;32),\ \langle 32;36), > 36$ (14 categories)
Tric : $1-4,\ 5,6,\ldots,12,13-14,15-17, \geq 18$ (12 categories).

Group *Difficulties* has three attributes with 2 - 5 categories, frequencies of particular categories are in brackets (there are some missing values, too):
Asthma with 2 categories: *yes* (frequency 1210) and *no* (frequency 192)
Chest i.e. *Chest pain* with 5 categories: *not present* (1019), *non ischaemic* (311), *angina pectoris* (52), *other* (19), *possible myocardial infarction* (3)
Lower limbs i.e. *Lower limbs pain* with 3 categories: *not present* (1282), *non ischaemic* (113), *claudication* (17).

Group *Blood pressure* has two attributes - *Diast* i.e. Diastolic blood pressure and *Syst* i.e. Systolic blood pressure The original values were transformed such that these attributes have the following categories:
Diast : $\langle 45;65),\ \langle 65;75),\ \langle 75;85),\ \ldots,\ \langle 105;115), > 115$ (7 categories)
Syst : $\langle 85;105),\ \langle 105;115),\ \langle 115;125),\ \ldots,\ \langle 165;175), > 175$ (9 categories).

2.2 Domain Knowledge

There are various types of domain knowledge related to STULONG data. Three of them in a formalized form are managed by the LISp-Miner system [7]: *groups of attributes*, *information on particular attributes* and *mutual influence of attributes*.

There are 11 basic groups (see `http://euromise.vse.cz/challenge2004/data/entry/`. These groups are mutually disjoint and their union is the set of

[1] The study (STULONG) was realized at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and University Hospital in Prague, under the supervision of Prof. F. Boudík, MD, DSc., with collaboration of M. Tomečková, MD, PhD and 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 CR(head. Prof. J. Zvárová, PhD, DSc.). The data resource is on the web pages `http://euromise.vse.cz/challenge2004/`.

all attributes. We call these groups *basic groups of attributes*, they are perceived by physicians as reasonable sets of attributes. It is also possible to define additional groups of atributes for some specific tasks, see e.g. groups *Measurement*, *Difficulties*, and *Blood pressure* introduced above.

Examples of information on particular attributes are boundaries for classification of overweight and obesity by BMI. Overweight is defined as BMI in $25.0 - 29.9$ and obesity as $BMI \geq 30$.

There are several types of influences among attributes. An example is expression $BMI \uparrow\uparrow Diast$ saying that if body mass index of patient increases then its diastolic blood pressure increases too.

3 Association Rules

The association rule is understood to be an expression $\varphi \approx \psi$ where φ and ψ are Boolean attributes. It means that the Boolean attributes φ and ψ are associated in the way given by the symbol $\approx$. This symbol is called the *4ft-quantifier*. It corresponds to a condition concerning a four-fold contingency table of φ and ψ. Various types of dependencies of φ and ψ can be expressed by 4ft-quantifiers.

The association rule $\varphi \approx \psi$ concerns analyzed data matrix $\mathcal{M}$. An example of a data matrix is data matrix *Entry* a fragment of which is in figure 1.

	attributes			examples of basic Boolean attributes	
patient	$Asthma$	BMI	...	$Asthma(yes)$	$BMI\ ((21;22\rangle, (22;23\rangle)$
o_1	*yes*	$(16;21\rangle$	...	1	0
⋮	⋮	⋮	⋱	⋮	⋮
o_{1417}	*no*	$(22;23\rangle$	...	0	1

Fig. 1. Data matrix $\mathcal{M}$ and examples of Boolean attributes

The Boolean attributes are derived from the columns of data matrix $\mathcal{M}$. We assume there is a finite number of possible values for each column of $\mathcal{M}$. Possible values are called *categories*. *Basic Boolean attributes* are created first. The basic Boolean attribute is an expression of the form $A(\alpha)$ where $\alpha \subset \{a_1, \ldots a_k\}$ and $\{a_1, \ldots a_k\}$ is the set of all possible values of the column A. The basic Boolean attribute $A(\alpha)$ is true in row o of $\mathcal{M}$ if it is $a \in \alpha$ where a is the value of the attribute A in row o. Set α is called a *coefficient* of $A(\alpha)$. Boolean attributes are derived from basic Boolean attributes using propositional connectives $\vee$, $\wedge$ and $\neg$ in a usual way.

There are two examples of basic Boolean attributes in figure 1 - $Asthma(yes)$ and $BMI((21;22\rangle, (22;23\rangle)$. Attribute $Asthma(yes)$ is true for patient o_1 and false for patient o_{1417}, we write "1" or "0" respectively. Attribute $BMI((21;22\rangle, (22;23\rangle)$ is false for o_1 because of $(16;21\rangle \notin \{(21;22\rangle, (22;23\rangle)\}$ and true for o_{1417} because of $(22;23\rangle \in \{(21;22\rangle, (22;23\rangle)\}$. Please note that we should write $Asthma(\{yes\})$ etc. but we will not do it. We will also usually write $BMI(21;23\rangle$ instead of $BMI\ ((21;22\rangle, (22;23\rangle)$ etc.

Table 1. 4ft table $4ft(\varphi, \psi, \mathcal{M})$ of φ and ψ in $\mathcal{M}$

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

The rule $\varphi \approx \psi$ is *true in data matrix* $\mathcal{M}$ if the condition corresponding to the 4ft-quantifier is satisfied in the four-fold contingency table of φ and ψ in $\mathcal{M}$, otherwise $\varphi \approx \psi$ is *false in data matrix* $\mathcal{M}$. The four-fold contingency table $4ft(\varphi, \psi, \mathcal{M})$ of φ and ψ in data matrix $\mathcal{M}$ is a quadruple $\langle a, b, c, d\rangle$ where a is the number of rows of $\mathcal{M}$ satisfying both φ and ψ, b is the number of rows of $\mathcal{M}$ satisfying φ and not satisfying ψ etc., see Table 1.

There are various 4ft-quantifiers, some of them are based on statistical hypothesis tests, see e.g. [1,6]. We use here a simple 4ft-quantifier i.e. quantifier $\Rightarrow_{p,Base}$ of *founded implication* [1]. It is defined for $0 < p \leq 1$ and $Base > 0$ by the condition $\frac{a}{a+b} \geq p \ \wedge \ a \geq Base$. The association rule $\varphi \Rightarrow_{p,Base} \psi$ means that at least $100p$ per cent of rows of $\mathcal{M}$ satisfying φ satisfy also ψ and that there are at least $Base$ rows of $\mathcal{M}$ satisfying both φ and ψ. We use this quantifier not only because of its simplicity but also because there are known deduction rules related to this quantifier [4].

4 Applying LISp-Miner System

The goal of this paper is to describe an application of an approach to filtering out association rules, which can be considered as consequences of given items of domain knowledge. This approach is based on mapping of items of domain knowledge in question to suitable sets of association rules and also on deduction rules concerning pairs of association rules. We deal with items of domain knowledge stored in the LISp-Miner system outlined in section 2.2. We use GUHA procedure 4ft-Miner [6] which mines for association rules described in section 3. In addition we outline how the groups of attributes can be used to formulate reasonable analytical questions.

An example of a reasonable analytical question is given in section 4.1. Input of the 4ft-Miner procedure consists of parameters defining a set of relevant association rules and of an analyzed data matrix. Output consists of all relevant association rules true in input data matrix. There are fine tools to define set of association rules which are relevant to the given analytical question. We use data *Entry*, see section 2.1. Input parameters of 4ft-Miner procedure suitable to solve our analytical question are described also in section 4.1. There are 158 true relevant association rules found for these parameters.

Our analytical question is formulated such that we are not interested in consequences of item of domain knowledge *BMI* $\uparrow\uparrow$ *Diast*. This item says that if body mass index of patient increases then his diastolic blood pressure increases too, see section 2.2. However, there are many rules among 158 resulting rules which can be considered as consequences of item *BMI* $\uparrow\uparrow$ *Diast*. We filter out these consequences in two steps.

In the first step we define a set $Cons(BMI \uparrow\uparrow Diast, Entry, \approx)$ of atomic consequences of $BMI \uparrow\uparrow Diast$. Each atomic consequence is an association rule of the form $BMI(\omega) \approx Diast(\delta)$ which can be considered as true in data matrix *Entry* if $BMI \uparrow\uparrow Diast$ is supposed to be true. In addition, $\approx$ is a 4ft-quantifier used in the 4ft-Miner application in question. For more details see section 4.2.

In the second step we filter out each association rule $\varphi \approx \psi$ from the output of 4ft-Miner which is equal to an atomic consequence or can be considered as a consequence of an atomic consequence. There are additional details in section 4.3.

4.1 Analytical Question and 4ft-Miner

Let us assume we are interested in an analytical question:

Are there any interesting relations between attributes from group Measurement *and attributes from group* Blood pressure *in the data matrix* Entry*? Attributes from group* Measurement *can be eventually combined with attributes from group* Difficulties*. Interesting relation is a relation which is strong enough and which is not a consequence of the fact* BMI $\uparrow\uparrow$ Diast.

This question can be symbolically written as

$$Measurement \wedge Difficulties \longrightarrow Blood\ pressure \quad [Entry\,;\, BMI \uparrow\uparrow Diast]\ .$$

We deal with association rules, thus we convert our question to a question concerning association rules. Symbolically we can express a converted question as

$$\mathcal{B}[Measurement] \wedge \mathcal{B}[Difficulties] \ \approx\ \mathcal{B}[Blood\ pressure] \quad [Entry\,;\, BMI \uparrow\uparrow Diast]\ .$$

Here B[*Measurement*] means a set of all Boolean attributes derived from attributes of the group *Measurement* we consider relevant to our analytical question, similarly for $\mathcal{B}[Difficulties]$ and $\mathcal{B}[Blood\ pressure]$.

We search for rules $\varphi_M \wedge \varphi_D \ \approx\ \psi_B$ which are true in data matrix *Entry*, cannot be understood as a consequence of $BMI \uparrow\uparrow Diast$ and $\varphi_M \in \mathcal{B}[Measurement]$, $\varphi_D \in \mathcal{B}[Difficulties]$, and $\psi_B \in \mathcal{B}[Blood\ pressure]$.

The procedure 4ft-Miner does not use the well known a-priori algorithm. It is based on representation of analyzed data by suitable strings of bits [6]. That's way 4ft-Miner has very fine tools to define such set of association rules. One of many possibilities how to do it is in figure 2. Remember that we deal with rules $\varphi \approx \psi$, φ is called *antecedent* and ψ is *succedent*. Set B[*Measurement*] is defined in column `ANTECEDENT` in row `Measurement Conj, 1-3` and in three consecutive rows.

Each φ_M is a conjunction of 1 - 3 Boolean attributes derived from particular attributes of the group *Measurement*. Set of all such Boolean attributes derived from attribute *BMI* is defined by the row `BMI(int), 1-3 B, pos`. It means that all Boolean attributes $BMI(\alpha)$ where α is a set of 1 - 3 consecutive categories (i.e. interval of categories) are generated. Examples of such Boolean attributes are $BMI(16; 21\rangle)$, $BMI((21; 22\rangle, (22; 23\rangle)$ i.e. $BMI(21; 23\rangle$, and $BMI((21; 22\rangle, (22; 23\rangle, (23; 24\rangle)$ i.e. $BMI(21; 24\rangle$. Sets of Boolean attributes derived from attributes *Subsc* and *Tric* are defined similarly. An example of $\varphi_M \in \mathcal{B}[Measurement]$ is conjunction $\varphi_M = BMI(21; 24\rangle \wedge Subsc\langle 4; 14)$.

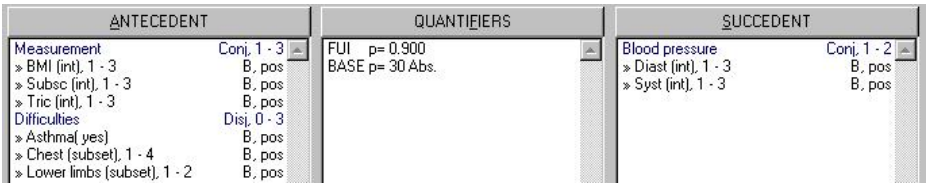


Fig. 2. Input parameters of the 4ft-Miner procedure

Each φ_D is a disjunction of 0 - 3 Boolean attributes derived from particular attributes of the group *Difficulties*. There is only one Boolean attribute derived from attribute *Asthma* i.e. *Asthma*(*yes*). Set of all such Boolean attributes derived from attribute *Chest* is defined by the row `Chest(subset), 1-4 B, pos`. It means that all Boolean attributes *Chest*(α) where α is a subset of 1 - 4 categories of attribute *Chest* are generated. In addition, category *not present* is not taken into account (not seen in figure 2). Similarly, all Boolean attributes *Lower limbs*(α) where α is a subset of 1 - 2 categories are generated and category *not present* is not considered. Please note, that a disjunction of zero Boolean attributes means that φ_D is not considred.

Set $\mathcal{B}$[*Blood pressure*] is defined in row `Blood pressure Conj, 1-2` of column `SUCCEDENT` and in two consecutive rows in a way similar to that in which set $\mathcal{B}$[*Measurement*] is defined. In column `QUANTIFIERS` the quantifier $\Rightarrow_{0.9,30}$ of founded implication is specified.

This task was solved in 171 minutes (PC with 2GB RAM and Intel T7200 processor at 2 GHz). $456 * 10^6$ association rules were generated and tested, 158 true rules were found. The rule $BMI(21;22\rangle \wedge Subsc(< 14) \Rightarrow_{0.97,33} Diast\langle 65;75)$ is the strongest one (i.e. with highest confidence). It means that 34 patients satisfy $BMI(21;22\rangle \wedge Subsc(< 14)$ and 33 from them satisfy also $Diast\langle 65;75)$.

Most of found rules have attribute *BMI* in antecedent and attribute *Diast* in succedent (as the above shown rule). We can expect that lot of such rules can be seen as a consequences of *BMI* $\uparrow\uparrow$ *Diast*.

4.2 Atomic Consequences of *BMI* $\uparrow\uparrow$ *Diast*

We define a set *Cons*(*BMI* $\uparrow\uparrow$ *Diast*,*Entry*, $\Rightarrow_{0.9,30}$) of simple rules in the form $BMI(\omega) \approx Diast(\delta)$ which can be considered as consequences of *BMI* $\uparrow\uparrow$ *Diast*. Such rules are called *atomic consequences of* BMI $\uparrow\uparrow$ Diast. We assume that this set is usually defined by a domain expert.

Examples of such atomic consequences are rules $BMI(low) \Rightarrow_{0.9,30} Diast(low)$ saying that at least 90 per cent of patients satisfying $BMI(low)$ satisfy also $Diast(low)$ and that there are at least 30 patients satisfying both $BMI(low)$ and $Diast(low)$. The only problem is to define suitable coefficients *low* for both attributes *BMI* and *Diast*.

Let us remember that there are 13 categories of *BMI* - $(16;21\rangle$, $(21;22\rangle$, $(21;22\rangle$, ..., $(31;32\rangle$, > 32 and 7 categories of *Diast* - $\langle 45;65)$, $\langle 65;75)$, ...,

$\langle 105; 115)$, > 115. We can decide that each Boolean attribute $BMI(\omega)$ where $\omega \subset \{(16; 21\rangle, (21; 22\rangle, (21; 22\rangle\}$ will be considered as $BMI(low)$ (we use low quarter of all categories) and similarly each Boolean attribute $Diast(\delta)$ where $\delta \subset \{\langle 45; 65), \langle 65; 75)\}$ will be considered as $Diast(low)$ (we use low third of all categories). We can say that rules $BMI(low) \Rightarrow_{0.9,30} Diast(low)$ are defined by a rectangle $\mathcal{A}_{low} \times \mathcal{S}_{low}$ = Antecedent $\times$ Succedent where

$$\text{Antecedent} \times \text{Succedent} = \{(16; 21\rangle, (21; 22\rangle, (21; 22\rangle\} \times \{\langle 45; 65), \langle 65; 75)\}$$

There is *LMDataSource* module in the LISp-Miner system which makes possible to do various input data transformations and in addition it also allows to define the set $Cons(BMI \uparrow\uparrow Diast, Entry, \Rightarrow_{0.9,30})$ as a union of several similar, possibly overlapping, rectangles $\mathcal{A}_1 \times \mathcal{S}_1, \ldots, \mathcal{A}_R \times \mathcal{S}_R$ such that $BMI(\omega) \Rightarrow_{0.9,30} Diast(\delta) \in Cons(BMI \uparrow\uparrow Diast, Entry, \Rightarrow_{0.9,30})$ if and only if there is an $i \in \{1, \ldots, K\}$ such that $\omega \subseteq \mathcal{A}_i$ and $\delta \subseteq \mathcal{S}_i$. An example of definition of a set $Cons(BMI \uparrow\uparrow Diast, Entry, \Rightarrow_{0.9,30})$ is in figure 3, six rectangles are used.

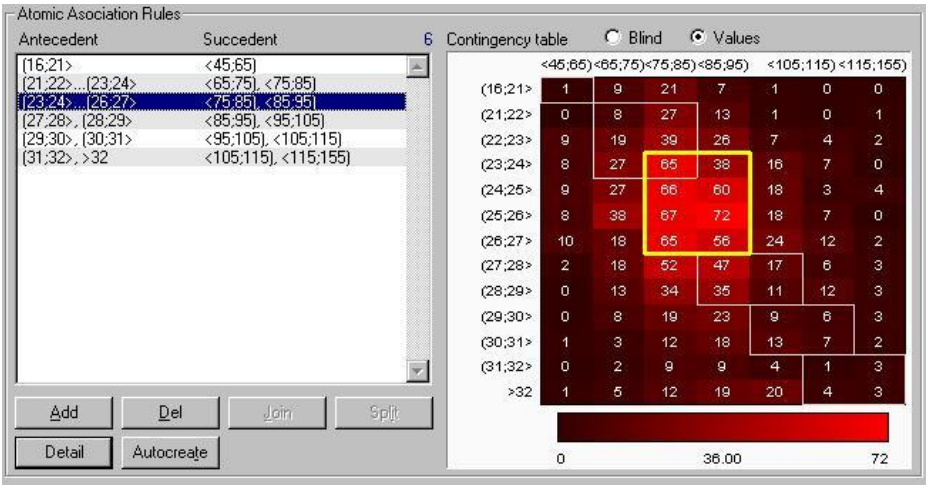


Fig. 3. Definition of atomic association rules

4.3 Filtering Out Consequences of *BMI* ↑↑ *Diast*

We will discuss possibilities of filtering out all rules from the output rules which can be considered as consequences of given item of domain knowledge. We take into account both strict logical deduction – see (ii), and specific conditions also supporting filtering out additional rules – see (iii). We use the set $Cons(BMI \uparrow\uparrow Diast, Entry, \Rightarrow_{0.9,30})$ of atomic consequences $BMI(\omega) \Rightarrow_{0.9,30} Diast(\delta)$ of $BMI \uparrow\uparrow Diast$ defined in figure 3. We filter out each of 158 output rules $\varphi \Rightarrow_{0.9,30} \psi$ satisfying one of conditions (i), (ii), (iii):

$\mathcal{M}$	$Diast\langle 65;85)$	$\neg Diast\langle 65;85)$
$BMI(21;22\rangle$	a	b
$\neg BMI(21;22\rangle$	c	d

$\mathcal{M}$	$Diast\langle 65;95)$	$\neg Diast\langle 65;95)$
$BMI(21;22\rangle$	a'	b'
$\neg BMI(21;22\rangle$	c'	d'

Fig. 4. $4ft(BMI(21;22\rangle, Diast\langle 65;85), \mathcal{M})$ and $4ft(BMI(21;22\rangle, Diast\langle 65;95), \mathcal{M})$

(i) $\varphi \Rightarrow_{0.9,30} \psi$ is equal to an atomic consequence $BMI(\omega) \Rightarrow_{0.9,30} Diast(\delta)$

(ii) $\varphi \Rightarrow_{0.9,30} \psi$ is a logical consequence of an atomic consequence $BMI(\omega) \Rightarrow_{0.9,30} Diast(\delta)$

(iii) $\varphi \Rightarrow_{0.9,30} \psi$ is in the form $\varphi_0 \wedge \varphi_1 \Rightarrow_{0.9,30} \psi_0$ where $\varphi_0 \Rightarrow_{0.9,30} \psi_0$ satisfies (i) or (ii). We filter out such rules because of patients satisfying $\varphi_0 \wedge \varphi_1$ satisfy also φ_0 and thus the rule $\varphi_0 \wedge \varphi_1 \Rightarrow_{0.9,30} \psi_0$ does not say something new in comparison with $\varphi_0 \Rightarrow_{0.9,30} \psi_0$ even if its confidence is higher than 0.9.

We give more details below.

(i): There is no atomic consequence $BMI(\omega) \Rightarrow_{0.9,30} Diast(\delta)$ belonging to $Cons(BMI \uparrow\uparrow Diast, Entry, \Rightarrow_{0.9,30})$ among the output rules.

(ii): There is output rule $BMI(21;22\rangle \Rightarrow_{0.9,30} Diast\langle 65;95)$ not belonging to $Cons(BMI \uparrow\uparrow Diast, Entry, \Rightarrow_{0.9,30})$. Rule $BMI(21;22\rangle \Rightarrow_{0.9,30} Diast\langle 65;85)$ belongs to $Cons(BMI \uparrow\uparrow Diast, Entry, \Rightarrow_{0.9,30})$, see second row in the left part of figure 3. In addition, rule $BMI(21;22\rangle \Rightarrow_{0.9,30} Diast\langle 65;95)$ logically follows from rule $BMI(21;22\rangle \Rightarrow_{0.9,30} Diast\langle 65;85)$.

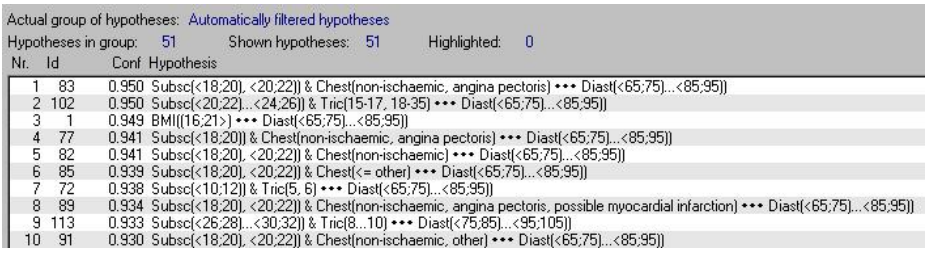
It means that if rule $BMI(21;22\rangle \Rightarrow_{0.9,30} Diast\langle 65;85)$ is true in a given data matrix $\mathcal{M}$ then rule $BMI(21;22\rangle \Rightarrow_{0.9,30} Diast\langle 65;95)$ is true in $\mathcal{M}$ too. Rule $BMI(21;22\rangle \Rightarrow_{0.9,30} Diast\langle 65;85)$ is for data matrix *Entry* considered as a consequence of $BMI \uparrow\uparrow Diast$ and thus rule $BMI(21;22\rangle \Rightarrow_{0.9,30} Diast\langle 65;95)$ can also be considered as a consequence of $BMI \uparrow\uparrow Diast$ for data matrix *Entry*. It means that rule $BMI(21;22\rangle \Rightarrow_{0.97,30} Diast\langle 65;95)$ is filtered out.

We demonstrate why rule $BMI(21;22\rangle \Rightarrow_{0.9,30} Diast\langle 65;95)$ logically follows from rule $BMI(21;22\rangle \Rightarrow_{0.9,30} Diast\langle 65;85)$. In figure 4 there are 4ft-tables $4ft(BMI(21;22\rangle, Diast\langle 65;85), \mathcal{M})$ of $BMI(21;22\rangle$ and $Diast\langle 65;85)$ in $\mathcal{M}$ and $4ft(BMI(21;22\rangle, Diast\langle 65;95), \mathcal{M})$ of $BMI(21;22\rangle$ and $Diast\langle 65;95)$ in $\mathcal{M}$. It is clear that $a + b = a' + b'$. In addition each patient satisfying $Diast\langle 65;85)$ satisfy also $Diast\langle 65;95)$ and thus $a' \geq a$ and $b' \leq b$ which means that if $\frac{a}{a+b} \geq 0.9 \wedge a \geq 30$ then also $\frac{a'}{a'+b'} \geq 0.9 \wedge a' \geq 30$.

Note that there is a theorem proved in [4] which makes possible to easy decide if association rule $\varphi' \Rightarrow_{p,Base} \psi'$ logically follows from $\varphi \Rightarrow_{p,Base} \psi$ or not.

(iii) It is also easy to show that rule $BMI(21;22\rangle \wedge Subsc(\leq 14) \Rightarrow_{0.9,30} Diast\langle 65;95)$ does not logically follow from rule $BMI(21;22\rangle \Rightarrow_{0.9,30} Diast\langle 65;95)$. However, patients satisfying $BMI(21;22\rangle \wedge Subsc(\leq 14)$ satisfy also $BMI(21;22\rangle$ and thus rule $BMI(21;22\rangle \wedge Subsc(\leq 14) \Rightarrow_{0.9,30} Diast\langle 65;95)$ does not say something new and can be also filtered out. (This could be of course a subject of additional discussion, however we will not discuss here due to limited space.)

From the same reason we filter out each rule $BMI(\rho) \wedge \varphi_1 \Rightarrow_{0.9,30} Diast(\tau)$ if the rule $BMI(\rho) \Rightarrow_{0.9,30} Diast(\tau)$ satisfies (i) or (ii). After filtering out all rules



Actual group of hypotheses: Automatically filtered hypotheses

Hypotheses in group: 51 Shown hypotheses: 51 Highlighted: 0

Nr.	Id	Conf	Hypothesis
1	83	0.950	Subsc(<18;20), <20;22)) & Chest(non-ischaemic, angina pectoris) ••• Diast(<65;75)...<85;95))
2	102	0.950	Subsc(<20;22)...<24;26)) & Tric(15-17, 18-35) ••• Diast(<65;75)...<85;95))
3	1	0.949	BMI([16;21>) ••• Diast(<65;75)...<85;95))
4	77	0.941	Subsc(<18;20)) & Chest(non-ischaemic, angina pectoris) ••• Diast(<65;75)...<85;95))
5	82	0.941	Subsc(<18;20), <20;22)) & Chest(non-ischaemic) ••• Diast(<65;75)...<85;95))
6	85	0.939	Subsc(<18;20), <20;22)) & Chest(<= other) ••• Diast(<65;75)...<85;95))
7	72	0.938	Subsc(<10;12)) & Tric(5, 6) ••• Diast(<65;75)...<85;95))
8	89	0.934	Subsc(<18;20), <20;22)) & Chest(non-ischaemic, angina pectoris, possible myocardial infarction) ••• Diast(<65;75)...<85;95))
9	113	0.933	Subsc(<26;28)...<30;32)) & Tric(8...10) ••• Diast(<75;85)...<95;105))
10	91	0.930	Subsc(<18;20), <20;22)) & Chest(non-ischaemic, other) ••• Diast(<65;75)...<85;95))

Fig. 5. Automatically filtered association rules

according to (i) – (iii), only 51 rules remain from the original 158 rules. Several examples are in figure 5.

We can see that there is true rule $BMI\langle 16;21\rangle \Rightarrow_{0.9,30} Diast\langle 65;95\rangle$ which satisfies neither (i) nor (ii) and thus it cannot be considered as a consequence of $BMI \uparrow\uparrow Diast$. This is a reason to study this rule in more detail, because it could be an interesting exception. It should be reported to the domain expert. However, let us emphasize that definition of $Cons(BMI \uparrow\uparrow Diast, Entry, \Rightarrow_{0.9,30})$ in figure 3 was done without a consultation with domain expert.

Additional remaining rules concern attributes *Subsc* and *Diast* in some cases combined with *Chest* and *Tric*. We assume that by a suitable analytical process we can offer a new item of domain knowledge $Subsc \uparrow\uparrow Diast$.

5 Conclusions and Further Work

Here presented approach allows filtering out all rules reasonable considered as consequences of domain knowledge, e.g. the above mentioned $BMI \uparrow\uparrow Diast$. This leads to a remarkable decrease of number of output association rules, so users could concentrate on interpretation of a smaller group of potentially more valuable association rules. An already available implementation has even more filtering features that could be moreover repeated in an iterative way.

Let us emphasize that there are several additional types of mutual influence of attributes [7]. An example is $Education \uparrow\downarrow BMI$ which says that if education increases then BMI decreases. All these types of knowledge can be treated in the above described way [5]. The described transformation of an item of domain knowledge into a set of association rues can be inverted and used to synthesize a new item of domain knowledge (e.g. $Subsc \uparrow\uparrow Diast$).

The whole approach seems to be understandable from the point view of a domain expert. However, a detailed explanation will be useful. This leads to necessity to prepare for each analytical question an analytical report explaining in details all of steps leading to its solution. There are first results in producing similar reports and presenting them at Internet [2].

Our goal is to elaborate the outlined approach into a way of automatized producing analytical reports answering given analytical question. Domain knowledge stored in the LISp-Miner system gives a possibility to automatically generate a

whole system of analytical questions. Successful experiments with running LISp-Miner system on a grid [9] makes possible to accept a challenge to create a system automatically formulating analytical question, getting new knowledge by answering these question and use new knowledge to formulate additional analytical question. Considerations on such a system are in [10].

References

1. Hájek, P., Havránek, T.: Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory. Springer, Heidelberg (1978)
2. Kliegr, T., et al.: Semantic Analytical Reports: A Framework for Post processing data Mining Results. In: Rauch, J., et al. (eds.) Foundations of Intelligent Systems, pp. 88–98. Springer, Heidelberg (2009)
3. Qiang, Y., Xindong, W.: 10 Challenging Problems in Data Mining Research. International Journal of Information Technology & Decision Making 5(4), 597–604 (2006)
4. Rauch, J.: Logic of Association Rules. Applied Intelligence 22, 9–28 (2005)
5. Rauch, J.: Considerations on Logical Calculi for Dealing with Knowledge in Data Mining. In: Ras, Z.W., Dardzinska, A. (eds.) Advances in Data Management. Studies in Computational Intelligence, vol. 223, pp. 177–199. Springer, Heidelberg (2009)
6. Rauch, J., Šimůnek, M.: An Alternative Approach to Mining Association Rules. In: Lin, T.Y., et al. (eds.) Data Mining: Foundations, Methods, and Applications, pp. 219–238. Springer, Heidelberg (2005)
7. Rauch, J., Šimůnek, M.: Dealing with Background Knowledge in the SEWEBAR Project. In: Berendt, B., Mladenič, D., de Gemmis, M., Semeraro, G., Spiliopoulou, M., Stumme, G., Svátek, V., Železný, F., et al. (eds.) Knowledge Discovery Enhanced with Semantic and Social Information. Studies in Computational Intelligence, vol. 220, pp. 89–106. Springer, Heidelberg (2009)
8. Suzuki, E.: Discovering interesting exception rules with rule pair. In: Fuernkranz, J. (ed.) Proceedings of the ECML/PKDD Workshop on Advances in Inductive Rule Learning, pp. 163–178 (2004)
9. Šimůnek, M., Tammisto, T.: Distributed Data-Mining in the LISp-Miner System Using Techila Grid. In: Zavoral, F., Yaghob, J., Pichappan, P., El-Qawasmeh, E. (eds.) NDT 2010. Communications in Computer and Information Science, vol. 87, pp. 15–20. Springer, Heidelberg (2010)
10. Šimůnek, M., Rauch, J.: EverMiner – Towards Fully Automated KDD Process, accepted for publication in Data Mining. In: TECH (2011) ISBN: 978-953-7619-X-X