Using Domain Knowledge in Association Rules Mining
Case Study

Jan Rauch and Milan Šimůnek
Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic

Keywords: Data Mining, Domain Knowledge, Association Rules, Patterns of Domain Knowledge, Logic of Association Rules.

Abstract: A case study concerning an approach to application of domain knowledge in association rule mining is presented. Association rules are understood as general relations of two general Boolean attributes derived from columns of an analysed data matrix. Interesting items of domain knowledge are expressed in an intuitive form distinct from association rules. Each particular pattern of domain knowledge is mapped onto a set of all association rules which can be considered as its consequences. These sets are used when interpreting results of data mining procedure. Deduction rules concerning association rules are applied.

1 INTRODUCTION

An approach to dealing with domain knowledge in association rules data mining is sketched in (Rauch and Šimůnek, 2011). Association rules of the form $\varphi \approx \psi$ are used. Here φ and ψ are general Boolean attributes derived from columns of an analysed data matrix. Symbol $\approx$ is a 4ft-quantifier, it corresponds to a condition concerning a contingency table of φ and ψ (Rauch, 2013).

Boolean attributes φ and ψ are built from basic Boolean attributes of the form $A(\alpha)$ where A is an attribute i.e. a column of a data matrix with possible values $a_1, \ldots, a_t$ and $\alpha \subset \{a_1, \ldots, a_t\}$. Basic Boolean attribute $A(\alpha)$ is true in row o of an analysed data matrix if it holds $A(o) \in \alpha$ i.e. if the value $A(o)$ of the attribute A in row o belongs to the set α.

We use the 4ft-Miner procedure (Rauch, 2013; Rauch and Šimůnek, 2005) which mines for such association rules. The 4ft-Miner procedure is an enhanced implementation of the ASSOC procedure introduced in (Hájek and Havránek, 1978). Its implementation is based on dealing with suitable strings of bits making possible to easy deal with basic Boolean attributes $A(\alpha)$. We do not use the a-priori algorithm (Agrawal et al., 1996).

We deal with SI-formulas expressing mutual influence of attributes. The expression *BMI* $\uparrow\uparrow$ *Diastolic* is an example of SI-formula. It concerns attributes *BMI* i.e. body mass index and *Diastolic* i.e. diastolic blood pressure. Its meaning is: *if BMI increases then diastolic blood pressure increases as well.*

For each SI-formula Ω and 4ft-quantifier $\approx$ a set $Cons(\Omega, \approx)$ of association rules $\varphi \approx' \psi$ which can be considered as consequences of Ω is defined. This set is then used when interpreting results of the 4ft-Miner procedure. Deduction rules rules of the form $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ where both $\varphi \approx \psi$ and $\varphi' \approx \psi'$ are association rules play an important role. If the deduction rule $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ is correct and the association rule $\varphi \approx \psi$ is true in a given data matrix $\mathcal{M}$, then the association rule $\varphi' \approx \psi'$ is also true in $\mathcal{M}$. These deduction rules are studied in (Rauch, 2013) in details together with additional features of special logical calculi of association rules.

An application of this approach for SI-formula *BMI* $\uparrow\uparrow$ *Diastolic* is outlined in (Rauch and Šimůnek, 2011). The goal of the paper is to present this approach for additional SI-formulas in details. The goal of this paper is not to get new medical knowledge, the goal is to present new possibilities of dealing with domain knowledge in association rules data mining. Well known items of domain knowledge together with freely downloadable medical data set are used to achieve this goal.

No similar approach based on domain knowledge and logical calculi of association rules is known to the authors. However, various alternative approaches are published e.g. in (Delgado et al., 2001; Delgado et al., 2011; Brossette et al., 1998; Ordonez et al., 2006; Roddick et al., 2003). Their detailed comparison with the approach presented here is beyond of the scope of this paper and it is left as a further work.

The STULONG medical data set is introduced in section 2 together with related items of domain knowledge. An analytical question concerning this

Rauch J. and Šimůnek M..
Using Domain Knowledge in Association Rules Mining - Case Study.
DOI: 10.5220/0004539101040111
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval and the International Conference on Knowledge Management and Information Sharing* (KDIR-2013), pages 104-111
ISBN: 978-989-8565-75-4

data set and related items of domain knowledge is presented in section 3. Applications of the 4ft-Miner procedure relevant to this analytical question are described in section 4. Sets of association rules which can be considered as consequences of items of domain knowledge in question are introduced in section 5. These sets are used to interpret results of the 4ft-Miner procedure, see section 6. Results related to an additional analytical question are shortly presented in section 7. Concluding remarks are in section 8.

2 STULONG DATA SET

2.1 Data Matrix Entry

We use data set STULONG concerning *Longitudinal Study of Atherosclerosis Risk Factors*, see http://euromise.vse.cz/challenge2004/. 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, see e.g. http://euromise.vse.cz/challenge2004/data/entry/.

We use four groups defined for this paper – *Social + BMI*, *Vices*, *Problems*, and *Examinations*. The groups are introduced in Tab. 1 together with attributes belonging to particular groups. Let us note: names of categories are followed by the frequencies of particular categories, *married*/1207 means that there are 1207 married patients in the data matrix *Entry*. Frequencies of categories of the attribute *BMI* are in Fig. 1. Frequencies of categories of the attribute *Cholesterol* are distributed similarly. There are missing values in the data matrix *Entry*, thus the sum of frequencies of all particular categories of some attributes can be less 1417. Names of some categories are abbreviated, we use *manager* instead of *managerial worker*, etc., see http:// euromise.vse.cz/ challenge2004/ data/ entry/ social.html#zodpov. *Tric* means *Skinfold above musculus triceps (mm)* and *Subsc* means *skinfold above musculus subscapularis (mm)*.

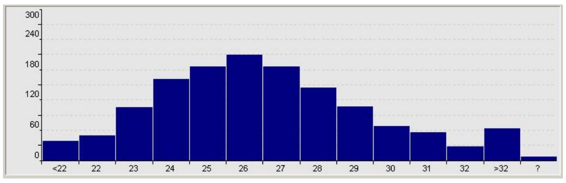


Figure 1: Frequencies of categories of *BMI*.

Table 1: Group of attributes and attributes.

Group of attributes	
Attribute	Categories
Social + BMI	
M_Status	*married*/1207, *divorced*/104 *single*/95, *widover*/10
Education	*basic*/151, *apprentice*/405 *secondary*/444, *university*/397
Responsibility	*manager*/286, *independent*/435 *others*/636, *pensioner*/25
BMI 13 categories	$< 22, 22, \ldots, 32, > 32$ frequencies see Fig. 1
Vices (given amount/day)	
Beer (litres)	*not*/465, $\leq$ 1/777, $>$ 1/157
Vine (litres)	*not*/675, $\leq$ 0.5/689, $>$ 0.5/36
Liquers (cc)	*not*/759, $\leq$ 100/574, $>$ 100/76
Smoking (cigarettes)	*not*/383, *1-4*/45, *5-14*/206 *15-20*/391, *21+*/346, *pipe*/29
Coffee (cups)	*not*/488, *1-2*/45, *3+*/643
Problems	
Hypertension	*yes*/220, *no*/1192
Infarction	*yes*/34, *no*/1378
Diabetes	*yes*/30, *no*/1378
Hyperlipidemia	*yes*/54, *no*/815
Examinations	
Tric (in cm)	$\langle 0;5\rangle$/176, $(5;10\rangle$/667, $(10;15\rangle$/303, $(15;20\rangle$/92, $(20;40\rangle$/43
Subsc (in cm)	$\langle 0;10\rangle$/130, $(10;15\rangle$/323, $(15;20\rangle$/399, $(20;25\rangle$/189, $(25;30\rangle$/118, $(30;55\rangle$/121
Cholesterol (in mg%) (10 categories)	$\langle 100;160)$, $\langle 160;180)$, ..., $\langle 300;320)$, $\langle 320;540\rangle$

2.2 Mutual Influence of Attributes

There are various items of domain knowledge concerning mutual influence of attributes and related to the STULONG data set. We use several of them, they correspond to the following SI-formulas.

SI-formula BMI $\uparrow^{+}$ Hypertension(yes) means: *if BMI increases, then the relative frequency of Hypertension(yes) increases as well.* Here BMI is a general ordinal attribute, the attribute *BMI* i.e. a column of data matrix *Entry* is an example of an instance of the general ordinal attribute BMI. Similarly, Hypertension(yes) is a general Boolean attribute, the attribute *Hypertension*(*yes*) is its instance. This approach can be described formally, SI-formulas can be understood as an enhancement of a logical calculus of association rules (Rauch, 2011). We use a less formal approach here.

We assume that BMI $\uparrow^{+}$ Hypertension(yes) is an approved and generally accepted medical knowledge. There are additional and similar SI-formulas BMI $\uparrow^{+}$ Infarction(yes), BMI $\uparrow^{+}$ Diabetes(yes), BMI $\uparrow^{+}$ Hyperlidemia(yes), and a similar set of SI-formulas can be created for each of ordinal attributes *Education*, *Beer*, *Vine*, *Liquers*, *Smoking*, *Tric*, *Subsc*, and *Cholesterol*. However, to show possibilities of formulation and answering analytical questions based on SI-formulas, we assume that BMI $\uparrow^+$ Hypertension(yes) is the only approved relevant medical knowledge here.

3 ANALYTICAL QUESTIONS

We have groups of attributes *Social + BMI*, *Vices*, *Problems*, and *Examinations*, see Tab. 1. In addition, we have SI-formulas concerning ordinal attributes from the groups *Social + BMI* and *Vices* and Boolean attributes from the group *Problems*. In addition, we assume that SI-formula BMI $\uparrow^+$ Hypertension(yes) is the only approved and generally accepted medical knowledge concerning data matrix *Entry*. Thus, it is natural to ask the following question Q_1:

Q_1: In the data matrix *Entry*, are there any interesting relations between attributes of the groups *Social + BMI* and *Vices* on the one side and the attributes of the group *Problems* on the other side which cannot be considered as consequences of $BMI \uparrow^+ Hypertension(yes)$?

We deal with association rules and thus the question Q_1 can be formulated as the question QAR_1:

QAR_1: In the *Entry* data matrix, are there any interesting true association rules $\varphi \approx \psi$ such that φ is a Boolean characteristics of the groups *Social + BMI* and *Vices*, ψ is a Boolean characteristics of the group *Problems*, $\approx$ is a suitable 4ft-quantifier, and these rules $\varphi \approx \psi$ cannot be considered as consequences of $BMI \uparrow^+ Hypertension(yes)$?

We use the procedure 4ft-Miner to solve QAR_1 in the following four steps. Below, we write Hpt(yes) instead of Hypertension(yes) (also BMI $\uparrow^+$ Hpt(yes) etc.):

1. We define a set Φ of interesting Boolean characteristics of the groups *Social + BMI* and *Vices* and a set Ψ of interesting Boolean characteristics of the group *Problems*. An example is in section 4.2.
2. We find a set $True(Entry, \Phi, \approx, \Psi)$ of all rules $\varphi \approx \psi$ which are true in *Entry*, $\varphi \in \Phi$, $\psi \in \Psi$, and $\approx$ is a suitable 4ft-quantifier. Several variants of definitions of Φ, Ψ, and $\approx$ can be used. Examples are in section 4.3.
3. We define a set $Cons(BMI \uparrow^+ Hpt(yes), \approx)$ of all rules $\varphi \approx \psi$ which can be considered as consequences of $BMI \uparrow^+ Hpt(yes)$, see section 5.
4. We investigate the set $\mathcal{TC}$ defined as $True(Entry, \Phi, \approx, \Psi) \cap Cons(BMI \uparrow^+ Hpt(yes), \approx)$. Depending on results of investigation we can get the following conclusions.

If $\mathcal{TC}$ contains only rules from the set $Cons(BMI \uparrow^+ Hpt(yes), \approx)$ then we conclude: All rules from $True(Entry, \Phi, \approx, \Psi)$ can be considered as consequences of $BMI \uparrow^+ Hpt(yes)$; there is no interesting rule $\varphi \approx \psi$ indicating an additional item of knowledge.

If $\mathcal{TC}$ contains no (or only several) rules from the set $Cons(BMI \uparrow^+ Hpt(yes), \approx)$ then we conclude: There are no (or only too few) rules in $True(Entry, \Phi, \approx, \Psi)$ which can be considered as consequences of $BMI \uparrow^+ Hpt(yes)$. Assuming that the definitions of sets Φ and Ψ are reasonable we can further conclude that this is suspicious and start investigation of circumstances of acquisition of the *Entry* data matrix.

If $\mathcal{TC}$ contains rules which are not from the set $Cons(BMI \uparrow^+ Hpt(yes), \approx)$ then we start interpretation of these rules. One of ways how to do this is to look if there are rules which can be considered as consequences of additional SI-formulas which correspond to items of knowledge which are not approved and generally accepted. This way is outlined below.

We use additional SI-formulas BMI $\uparrow^+$ ATR(yes) where ATR is one of the attributes Infarction(yes), Diabetes(yes), and Hyperlidemia(yes). For each such SI-formula Ω we continue this way: We define a set $Cons(\Omega, \approx)$ of rules $\varphi \approx \psi$ which can be considered as consequences of Ω. Then we investigate the set $\mathcal{TC}_\Omega$ defined as $True(Entry, \Phi, \approx, \Psi) \cap Cons(\Omega, \approx)$. Remember that we assume that Ω is not and approved and generally accepted item of medical knowledge.

Depending on results of investigation we can get the following conclusions.

If $\mathcal{TC}_\Omega$ does not contain rules from the set $Cons(\Omega, \approx)$ then we conclude: There is no indication of Ω in the *Entry* data matrix.

If $\mathcal{TC}_\Omega$ contains some rules from the set $Cons(\Omega, \approx)$ then we conclude: There are indications of Ω in the *Entry* data matrix. Then we can start suitable activity (e.g. confirmation analysis starting with getting additional observations) to decide if Ω is a generally acceptable item of knowledge.

Examples of such conclusions related to the question QAR_1 are given in section 6. Examples of conclusions related to an additional task are in section 7.

4 APPLYING 4ft-Miner

4.1 Association Rules

Association rule is an expression $\varphi \approx \psi$ where φ and ψ are Boolean attributes. It means that φ and ψ are associated in a way given by the symbol $\approx$. φ is called *antecedent*, ψ is called *succedent*. Symbol $\approx$ is a *4ft-quantifier*. It corresponds to a condition concerning a four-fold contingency table of φ and ψ. The association rule $\varphi \approx \psi$ concerns analysed data matrix. Data matrix *Entry* is an example of such data matrix.

The rule $\varphi \approx \psi$ is *true in data matrix* $\mathcal{M}$ if the condition corresponding to 4ft-quantifier $\approx$ is satisfied in a four-fold contingency table of φ and ψ in $\mathcal{M}$, otherwise $\varphi \approx \psi$ is *false in* $\mathcal{M}$. 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 2. There are various 4ft-quantifiers, see e.g. (Rauch, 2013).

Table 2: 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

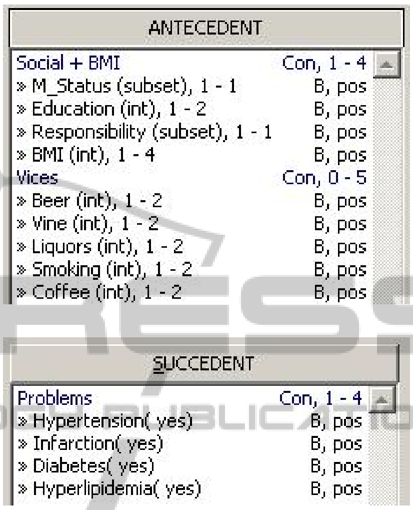
We use here the 4ft-quantifier $\Rightarrow_{p,B}$ of *founded implication*. It is defined for $0 < p \leq 1$ and $B > 0$ by the condition $\frac{a}{a+b} \geq p \ \wedge \ a \geq B$. The association rule $\varphi \Rightarrow_{p,B} \psi$ means that at least $100p$ per cent of rows of $\mathcal{M}$ satisfying φ satisfy also ψ and that there are at least B rows of $\mathcal{M}$ satisfying both φ and ψ.

We also use 4ft-quantifier $\sim^{+}_{q,B}$ of *above average dependence* defined for $0 < q$, $B > 0$ by the condition $\frac{a}{a+b} \geq (1+q)\frac{a+c}{a+b+c+d} \wedge a \geq B$. This says that among rows satisfying φ is at least $100p$ per cent more rows satisfying ψ than among all rows and that there are at least B rows satisfying both φ and ψ.

4.2 Set of Relevant Rules

We solve analytical question QAR_1: In the *Entry* data matrix, are there any interesting true association rules $\varphi \approx \psi$ such that φ is a Boolean characteristics of the groups *Social + BMI* and *Vices*, ψ is a Boolean characteristics of the group *Problems*, $\approx$ is a suitable 4ft-quantifier, and $\varphi \approx \psi$ cannot be considered as a consequence of $BMI \uparrow^{+} Hypertension(yes)$?

We use the 4ft-Miner procedure in four steps introduced in section 3. In the first step we define a set Φ of relevant antecedents i.e. Boolean characteristics of the groups *Social + BMI* and *Vices* and a set Ψ of relevant succedents i.e. Boolean characteristics of the group *Problems*. The set Φ is defined as a set of all conjunctions $\varphi_S \wedge \varphi_V$ where $\varphi_S \in \mathcal{B}(Social + BMI)$ and $\varphi_V \in \mathcal{B}(Vices)$. Here $\mathcal{B}(Social + BMI)$ means a set of all Boolean attributes derived from the attributes of the group *Social + BMI* we consider relevant to our analytical question, similarly for $\mathcal{B}(Vices)$. The set Ψ can be similarly denoted as $\mathcal{B}(Problems)$.



ANTECEDENT

Social + BMI	Con, 1 - 4
» M_Status (subset), 1 - 1	B, pos
» Education (int), 1 - 2	B, pos
» Responsibility (subset), 1 - 1	B, pos
» BMI (int), 1 - 4	B, pos
Vices	Con, 0 - 5
» Beer (int), 1 - 2	B, pos
» Vine (int), 1 - 2	B, pos
» Liquors (int), 1 - 2	B, pos
» Smoking (int), 1 - 2	B, pos
» Coffee (int), 1 - 2	B, pos

SUCCEDENT

Problems	Con, 1 - 4
» Hypertension(yes)	B, pos
» Infarction(yes)	B, pos
» Diabetes(yes)	B, pos
» Hyperlipidemia(yes)	B, pos

Figure 2: Definitions of relevant antecedents and succedents.

The set $\mathcal{B}(Social + BMI)$ is defined in the frame `ANTECEDENT` in Fig 2 in row `Social + BMI Conj, 1-4` and in four consecutive rows. Each φ_S is a conjunction of 1 - 4 basic Boolean attributes derived from particular attributes of the group *Social + BMI*.

Set of basic Boolean attributes derived from attribute *M_Status* is defined by the row `M_Status(subset), 1-1 B, pos`. It means that all basic Boolean attributes *M_Status*(α) where α is a subset of all categories of attribute *M_Status* containing just one category are generated: *M_Status*(*married*), *M_Status*(*divorced*), *M_Status*(*single*), and *M_Status*(*widower*). Set of basic Boolean attributes derived from attribute *Responsibility* is defined similarly.

Set of basic Boolean attributes derived from attribute *Education* is defined by the row `Education(int), 1-2 B, pos`. Thus, all 7 basic Boolean attributes *Education*(α) where α is a set of 1 or 2 consecutive categories (i.e. interval of categories) are generated. *Education*(*basic school*), *Education*(*basic school*, *apprentice school*) are examples.

Set of all Boolean attributes derived from the attribute *BMI* is defined by the row `BMI(int), 1-4 B, pos`. It means that all Boolean attributes *BMI*(α) where α is a set of 1 - 4 consecutive categories (i.e.

interval of categories) are generated. The Boolean attributes BMI(< 22) and $BMI(22,23,24,25)$ are examples. 46 basic Boolean attributes are defined this way and more than 6 500 conjunctions φ_S are defined altogether.

The set $\mathcal{B}(Vices)$ is defined similarly, see row `Vices Conj, 0-5` and five consecutive rows in the frame `ANTECEDENT` in Fig 2. The number 0 in this row means that the attribute φ_V can be skipped. Altogether, there are more than 11 000 conjunctions φ_V and more than 73×10^6 conjunctions $\varphi_S \wedge \varphi_V$.

The set Ψ i.e. $\mathcal{B}(Problems)$ of relevant succedents is defined in row `Problems Conj, 1-4` and four consecutive rows in the frame `SUCCEDENT` in Fig 2. Each $\varphi_P \in \mathcal{B}(Problems)$ is a conjunction of 1 - 4 basic Boolean attributes derived from attributes of the group *Problems*. There is only one basic Boolean attribute derived from attribute *Hypertensions* i.e. *Hypertensions*(yes), see `Hypertension(yes) B, pos.` The same is true for remaining attributes of the group *Problems*. Thus, there are 15 relevant succedents.

In addition, there are more than 10^9 association rules $\varphi_S \wedge \varphi_V \approx \varphi_P$ where $\varphi_S \in \mathcal{B}(Social + BMI)$, $\varphi_V \in \mathcal{B}(Vices)$, $\varphi_P \in \mathcal{B}(Problems)$ and $\approx$ is a 4ft-quantifier.

4.3 True Relevant Association Rules

We used three runs of the 4ft-Miner procedure to get sets $True(Entry, \Phi, \approx, \Psi)$ of all rules $\varphi \approx \psi$ which are true in *Entry*, see the second step in section 3. We used Φ and Ψ defined in the previous section.

The 4ft-quantifier $\Rightarrow_{0.9,30}$ defined by the condition $\frac{a}{a+b} \geq 0.9 \wedge a \geq 30$ (see section 4.1) was used first. The task was solved in 4 minutes (PC with 2GB RAM and Intel T7200 processor at 2 GHz) and 3.35×10^6 association rules were generated and tested. Various optimization techniques are implemented in the 4ft-Miner procedure, see (Rauch and Šimůnek, 2005). Thus, not all $> 10^9$ rules are truly generated and tested. However, no true rules was found, i.e. $True(Entry, \Phi, \Rightarrow_{0.9,30}, \Psi) = \emptyset$.

Thus we used 4ft-quantifier $\Rightarrow_{0.3,30}$ instead of $\Rightarrow_{0.9,30}$. This setting led to 24 true rules. In other words, the set $True(Entry, \Phi, \Rightarrow_{0.3,30}, \Psi)$ contains 24 rules. The strongest rule is the rule (we denote this rule as $\mathcal{R}$) $BMI(\geq 30) \wedge \Gamma \Rightarrow_{0.314,32} Hpt(yes)$ with 4ft-table $4ft(BMI(\geq 30) \wedge \Gamma, Hpt(yes), Entry)$ shown in Figure 3. We write $Hpt(yes)$ instead of $Hypertensions(yes)$. Γ abbreviates $Beer(not, \leq 1) \wedge Vine(not, \leq 0.5) \wedge Liquors(not, \leq 100)$.

Entry	$Hpt(yes)$	$\neg Hpt(yes)$
$BMI(\geq 30) \wedge \Gamma$	32	70
$\neg(BMI(\geq 30) \wedge \Gamma)$	187	1118

Figure 3: $4ft(BMI(\geq 30) \wedge \Gamma, Hpt(yes), Entry)$.

Rule $\mathcal{R}$ says that relative frequency of patients satisfying $Hpt(yes)$ among patients satisfying $BMI(\geq 30) \wedge \Gamma$ (i.e. confidency) is 0.314 and that there are 32 patients satisfying both $BMI(\geq 30) \wedge \Gamma$ and $Hpt(yes)$.

We also used 4ft-quantifier $\sim^{+}_{0.1,30}$ instead of $\Rightarrow_{0.9,30}$. This settings led to 3 754 true rules, which means that the set $True(Entry, \Phi, \sim^{+}_{0.1,30}, \Psi)$ contains 3 754 rules. Succedents of 3 749 of them are equal to $Hypertensions(yes)$. The strongest rule (what concerns lift related to 0.1 in the 4ft-quantifier $\sim^{+}_{0.1,30}$, see below) is the rule (we denote this rule as $\mathcal{R}_1$) $BMI(\geq 31) \wedge \Gamma_1 \sim^{+}_{1.02,31} Hpt(yes)$ with 4ft-table $4ft(BMI(\geq 31) \wedge \Gamma_1, Hpt(yes), Entry)$ shown in Fig. 4. $Hpt(yes)$ means the same as above, Γ_1 abbreviates $M\_Status(married) \wedge Vine(not, \leq 0.5) \wedge Liquors(not, \leq 100) \wedge Coffee(not, 1\text{-}2)$.

Entry	$Hpt(yes)$	$\neg Hpt(yes)$
$BMI(\geq 31) \wedge \Gamma_1$	31	68
$\neg(BMI(\geq 31) \wedge \Gamma_1)$	187	1119

Figure 4: $4ft(BMI(\geq 31) \wedge \Gamma_1, Hpt(yes), Entry)$.

Rule $\mathcal{R}_1$ says that relative frequency of patients satisfying $Hpt(yes)$ among patients satisfying $BMI(\geq 31) \wedge \Gamma_1$ (i.e. $\frac{31}{31+68}$) is 102 per cent higher than relative frequency of patients satisfying $Hpt(yes)$ among all patients (i.e. $\frac{31+187}{31+68+187+1119}$) and that there are 31 patients satisfying both $BMI(\geq 31) \wedge \Gamma_1$ and $Hpt(yes)$.

Let us note that the value 0.1 in the 4ft-quantifier $\sim^{+}_{0.1,30}$ means that $\frac{a}{a+b} \geq (1+0.1)\frac{a+c}{a+b+c+d}$, see section 4.1. This correspond to the fact that lift of the association rule in question is ≥ 1.1.

5 CONSEQUENCES OF ITEMS OF DOMAIN KNOWLEDGE

Runs of the 4ft-Miner procedure resulted into two non-empty sets – $True(Entry, \Phi, \Rightarrow_{0.3,30}, \Psi)$ containing 24 rules and $True(Entry, \Phi, \sim^{+}_{0.1,30}, \Psi)$ with 3 754 rules. We define a set $Cons(BMI \uparrow^{+} Hpt(yes), \approx)$ of all rules $\varphi \approx \psi$ which can be considered as consequences of $BMI \uparrow^{+} Hpt(yes)$ for both $\Rightarrow_{0.3,30}$ and $\sim^{+}_{0.1,30}$, see section 3.

The set $Cons(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$ of all rules $\varphi \Rightarrow_{p,B} \psi$ which can be considered as consequences of $BMI \uparrow^{+} Hpt(yes)$ is defined in four steps:

1) A set $AC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$ of *atomic consequences* of $BMI \uparrow^{+} Hpt(yes)$ for $\Rightarrow_{0.3,30}$ is defined as a set of simple rules $BMI(\delta) \Rightarrow_{p',B'} Hpt(yes)$

such that $p' \geq 0.3$, $B' \geq 30$ and $BMI(\delta)$ is a basic Boolean attribute expressing (informally speaking) that *BMI* is high enough.

2) A set $AgC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$ of *agreed consequences* of $BMI \uparrow^{+} Hpt(yes)$ for $\Rightarrow_{0.3,30}$ is defined. A rule $\rho \Rightarrow_{p,B} \sigma$ belongs to the set $AgC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$ if the following conditions are satisfied:

- $\rho \Rightarrow_{p,B} \sigma \notin AC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$
- there is no $\kappa \Rightarrow_{p',B'} \lambda$ belonging to $AC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$ such that $\rho \Rightarrow_{p,B} \sigma$ logically follows from $\kappa \Rightarrow_{p',B'} \lambda$.
- there is $\kappa \Rightarrow_{p',B'} \lambda$ belonging to $AC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$ such that, according to a domain expert, it is possible to agree that $\rho \Rightarrow_{p,B} \sigma$ says nothing new in addition to $\kappa \Rightarrow_{p',B'} \lambda$.

3) A set $LgC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$ of *logical consequences* of $BMI \uparrow^{+} Hpt(yes)$ for $\Rightarrow_{0.3,30}$ is defined. A rule $\varphi \Rightarrow_{p,B} \psi$ belongs to the set $LgC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$ if the following conditions are satisfied:

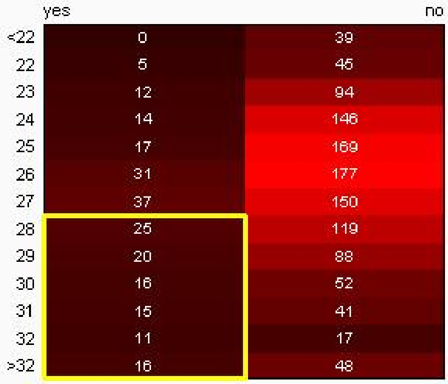
- $\varphi \Rightarrow_{p,B} \psi \notin (AC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30}) \cup AgC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30}))$
- there is $\tau \Rightarrow_{p',B'} \omega$ belonging to the set $AC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30}) \cup AgC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$ such that $\varphi \Rightarrow_{p,B} \psi$ logically follows from $\tau \Rightarrow_{p',B'} \omega$.

4) We define $Cons(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30}) =$
$AC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30}) \cup$
$AgC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30}) \cup$
$LgC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$.

We give more details to particular steps 1) – 3).

1) The set $AC(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$ is defined as a set of all rules $BMI(\delta) \Rightarrow_{p,B} Hpt(yes)$ where $p' \geq 0.3$, $B' \geq 30$ and $BMI(\delta)$ is a basic Boolean attribute expressing that *BMI* is high enough. Procedure 4ft-Miner has a simple tool facilitating a definition of a set of basic Boolean attributes $BMI(\delta)$ expressing that *BMI* is high enough, see Fig. 5. This is a contingency table of the attributes *BMI* and *Hypertension*. Based on this table we can decide in co-operation with a domain expert that we will consider all basic Boolean attributes $BMI(\delta)$ such that $\delta \subset \{28, 29, 30, 31, 32, > 32\}$. It is crucial that this definition is stored and further used.

2) An association rule $\varphi' \Rightarrow_{p,B} \psi'$ logically follows from a rule $\varphi \Rightarrow_{p,B} \psi$ if the following is true: If $\varphi \Rightarrow_{p,B} \psi$ is true in a given data matrix $\mathcal{M}$ then $\varphi' \Rightarrow_{p,B} \psi'$ is also true in $\mathcal{M}$. It is easy to prove that the association rule $BMI(\delta) \wedge \chi \Rightarrow_{0.3,30} Hpt(yes)$ does not logically follow from $BMI(\delta) \Rightarrow_{0.3,30} Hpt(yes)$.



	yes	no
<22	0	39
22	5	45
23	12	94
24	14	146
25	17	169
26	31	177
27	37	150
28	25	119
29	20	88
30	16	52
31	15	41
32	11	17
>32	16	48

Figure 5: Frequencies of categories of *BMI*.

The core of the proof is the fact that if there are at least *B* rows of a data matrix $\mathcal{M}$ satisfying $BMI(\delta) \wedge Hpt(yes)$ then there still can be no row of $\mathcal{M}$ satisfying $BMI(\delta) \wedge \chi \wedge Hpt(yes)$.

However, in *some* cases it can be reasonable from the point of view of a domain expert to agree that $BMI(\delta) \wedge \chi \Rightarrow_{0.3,30} Hpt(yes)$ is a consequence of $BMI(\delta) \Rightarrow_{0.3,30} Hpt(yes)$. Then we call the rule $BMI(\delta) \wedge \chi \Rightarrow_{0.3,30} Hpt(yes)$ an *agreed consequence* of the rule $BMI(\delta) \Rightarrow_{0.3,30} Hpt(yes)$.

The rule $BMI(\delta) \wedge Beer(not) \Rightarrow_{0.3,30} Hpt(yes)$ is an example of an agreed consequence of $BMI \uparrow^{+} Hpt(yes)$ because the truthfulness of Boolean attribute $Beer(not)$ has no influence on the relation of *BMI* and $Hpt(yes)$. The same is true for all basic Boolean attributes we can derive from the attributes of the groups $Social + BMI$ and *Vices* (except *BMI*).

3) A criterion making possible to decide if an association rule $\varphi' \Rightarrow_{p,B} \psi'$ logically follows from a rule $\varphi \Rightarrow_{p,B} \psi$ is proved in (Rauch, 2013). It is, e.g., easy to prove that $BMI(\delta) \Rightarrow_{0.3,30} Hpt(yes) \vee Infarction(yes)$ logically follows from $BMI(\delta) \Rightarrow_{0.3,30} Hpt(yes)$.

Let us emphasize that the same approach can be used to get the set $Cons(BMI \uparrow^{+} Hpt(yes), \sim^{+}_{0.1,30})$. In addition, it is important that the 4ft-Miner procedure has tools making possible to apply this approach to get a set $Cons(\Omega, \approx)$ of all rules $\varphi \approx \psi$ which can be considered as consequences of an item Ω of domain knowledge for a 4ft-quantifier $\approx$. This is possible for various types of items Ω of domain knowledge and many important 4ft-quantifiers $\approx$. Results on deduction rules of the form $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ (Rauch, 2013) are used.

6 INTERPRETING RESULTS

We have two non-empty sets of rules – $True(Entry, \Phi, \Rightarrow_{0.3,30}, \Psi)$ containing 24 rules and $True(Entry, \Phi, \sim^{+}_{0.1,30}, \Psi)$ containing 3754 rules, see section 4.3. We interpret these sets according the point 4 introduced in section 3. The 4ft-Miner procedure makes possible to easy compare these sets with sets $Cons(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$ and $Cons(BMI \uparrow^{+} Hpt(yes), \sim^{+}_{0.1,30})$ respectively. Results of comparison are summarized in Table 3. Here $\mathcal{C}(\Omega, \approx)$ abbreviates $Cons(BMI \uparrow^{+} Hpt(yes), \approx)$.

Table 3: 4ft table $4ft(\varphi, \psi, \mathcal{M})$ of φ and ψ in $\mathcal{M}$.

$\approx$	in $\mathcal{C}(\Omega, \approx)$	not in $\mathcal{C}(\Omega, \approx)$	Σ
$\Rightarrow_{0.3,30}$	24	0	24
$\sim^{+}_{0.1,30}$	1 407	1 347	3 754

The first row in the body of Table 3 means that all rules in $True(Entry, \Phi, \Rightarrow_{0.3,30}, \Psi)$ can be considered as consequences of $BMI \uparrow^{+} Hpt$. Thus we can conclude that there are no interesting rules $\varphi \Rightarrow_{0.3,30} \psi$ true in data matrix *Entry* such that φ is a Boolean characteristics of the groups *Social + BMI* and *Vices*, ψ is a Boolean characteristics of the group *Problems*, and $\varphi \Rightarrow_{0.3,30} \psi$ cannot be considered as a consequence of $BMI \uparrow^{+} Hpt(yes)$.

The second row in the body of Table 3 means that there are 1 347 interesting rules $\varphi \sim^{+}_{0.1,30} \psi$ true in data matrix *Entry* such that φ is a Boolean characteristics of the groups *Social + BMI* and *Vices*, ψ is a Boolean characteristics of the group *Problems*, and $\varphi \sim^{+}_{0.1,30} \psi$ cannot be considered as a consequence of $BMI \uparrow^{+} Hpt(yes)$.

The 4ft-Miner procedure has additional tools enabling the following conclusions:

(i) There are only 4 rule concerning *Infarction*(yes) and 1 rule concerning *Hyperlipidemia*(yes) among the mentioned 1 347 rules, all remaining 1 342 rules concern $Hpt(yes)$. This is because frequencies of attributes *Infarction*(yes), *Hyperlipidemia*(yes), and *Diabetes*(yes) are very low, see Table 1.

(ii) Among remaining 1 342 rules concerning $Hpt(yes)$, there are 411 rules which can be written as $BMI(\delta) \wedge \tau \sim^{+}_{0.1,30} Hpt(yes)$ where $\delta \subset \{24, 25, 26\}$. This means that these rules are not consequences of $BMI \uparrow^{+} Hpt(yes)$ in the sense of the definition in section 5. They can be seen as candidates of exceptions from the item $BMI \uparrow^{+} Hpt(yes)$. However, a deeper discussion on this topic is out of the scope of this paper.

These conclusions are made under the assumptions that the sets of interesting rules are defined in the way described in section 4.2 and that the sets $Cons(BMI \uparrow^{+} Hpt(yes), \Rightarrow_{0.3,30})$ and $Cons(BMI \uparrow^{+} Hpt(yes), \sim^{+}_{0.1,30})$ defined in section 5 are used. Let us emphasize that all these definitions can be modified in various ways and thus various variants of these conclusions can be formulated.

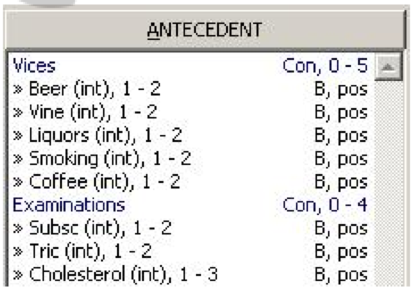
Let us also emphasize that rules with minimal confidence 0.3 used above are not too much suitable to express interesting relations. We use them here only to show principles of the presented approach. In addition, there are lot of rules with stronger quantifiers than $\sim^{+}_{0.1,30}$ in the set $True(Entry, \Phi, \sim^{+}_{0.1,30}, \Psi)$ containing 3 754 rules. The strongest one is a rule with quantifier $\sim^{+}_{1.02,31}$ (i.e. lift 2.02).

7 ADDITIONAL RESULTS

The analytical questions QAR_1 solved above can be modified to the question QAR_2 (note that attribute BMI is not involved):

QAR_2: In the *Entry* data matrix, are there any interesting true association rules $\varphi \approx \psi$ such that φ is a Boolean characteristics of the groups *Vices* and *Examinations*, ψ is a Boolean characteristics of the group *Problems*, $\approx$ is a suitable 4ft-quantifier?

To solve this question, we used a run of the 4ft-Miner procedure with the definition of relevant antecedents according to Fig. 6, the definition of relevant succedents according to Fig. 2, and the quantifier $\sim^{+}_{0.5,30}$ (i.e. lift = 1.5). This resulted to 71 true rules, all of them concern $Hpt(yes)$.



ANTECEDENT

Vices Con, 0 - 5
» Beer (int), 1 - 2 B, pos
» Vine (int), 1 - 2 B, pos
» Liquors (int), 1 - 2 B, pos
» Smoking (int), 1 - 2 B, pos
» Coffee (int), 1 - 2 B, pos
Examinations Con, 0 - 4
» Subsc (int), 1 - 2 B, pos
» Tric (int), 1 - 2 B, pos
» Cholesterol (int), 1 - 3 B, pos

Figure 6: Definition of relevant antecedents for QAR_2.

If we define a set of atomic consequences of $Subsc \uparrow^{+} Hpt$ as a set of all rules $Subsc(\delta) \sim^{+}_{0.5,30} Hpt$ where $\delta \subset \{\langle 20; 25), \langle 25; 30), \langle 30; 55\rangle\}$ (see "> 20" in the row *Subsc* of Table 4) and if we use an analogous approach as described in Section 5, we can conclude that there are 28 rules which can be considered as consequences of $Subsc \uparrow^{+} Hpt$ and 1 rule which cannot be considered as a consequence of $Subsc \uparrow^{+} Hpt$. In addition there are 29 rules concerning the attribute *Subsc*, see row *Subsc* of Table 4. There are analogous information for attributes *Tric* and *Cholesterol* which is abbreviated as *Chol*, see Table 4.

We can conclude that there are lot of rules which can be considered as consequences of the SI-formula $Subsc \uparrow^{+} Hpt$ and only one rule which cannot be con-

Table 4: Summary of results for QAR_2.

Attribute	Atomic	$\in$ *Cons*	$\notin$ *Cons*	Total
Subsc	> 20	28	1	29
Tric	> 10	7	4	11
Chol	≥ 240	18	12	30

sidered as a consequence of this SI-formula. Thus it seems reasonable to try to confirm the hypothesis *Subsc* $\uparrow^+$ *Hpt*. In addition, we can conclude that there are no strong indications of *Tric* $\uparrow^+$ *Hpt* and *Chol* $\uparrow^+$ *Hpt*.

However, there are various possibilities of modifications of parameters of the set of relevant antecedents, modifications of the quantifier $\sim^+_{0.5,30}$ and modifications of the definitions of sets of consequences of *Tric* $\uparrow^+$ *Hpt* and *Chol* $\uparrow^+$ *Hpt*. This can lead to revision of the introduced conclusions.

8 CONCLUSIONS

We have presented a new way of dealing with domain knowledge in association rules data mining. This is based on mapping items of domain knowledge to sets of association rules which can be considered as their consequences. It was shown that there is both necessary theory based on logic of association rules and the 4ft-Miner procedure realizing relevant operations with data, items of knowledge and rules. This makes possible to formulate interesting analytical questions and answer them in an efficient way. There is a very fine way to define sets of relevant association rules. These association rules, when true in data, can be considered as the smallest possible indications of more complex dependences among related attributes.

However, there is still a challenge concerning sensitivity of the presented approach to various parameters. There is also a challenge of combining of the 4ft-Miner procedure for mining the presented syntactically rich association rules with additional data mining procedures, namely with procedures of the LISp-Miner system dealing with various contingency tables (Hájek et al., 2010). These topics are subjects of further work.

ACKNOWLEDGEMENTS

The work described here has been supported by the grant IGA 20/2013 of the University of Economics, Prague.

REFERENCES

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I. (1996). Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press.

Brossette, S. E., Sprague, A. P., Hardin, J. M., Waites, K. B., Jones, W. T., and Moser, S. A. (1998). Research paper: Association rules and data mining in hospital infection control and public health surveillance. *JAMIA*, 5(4):373–381.

Delgado, M., Ruiz, M., and Sanchez, D. (2011). New approaches for discovering exception and anomalous rules. *International Journal of Uncertainty and Knowledge-based Systems*, 19(2):361–399.

Delgado, M., Sanchez, D., Martin-Bautista, M., and Vila, M. (2001). Mining association rules with improved semantics in medical databases. *Artificial Intelligence in Medicine*, 21(1–3):241–245.

Hájek, P. and Havránek, T. (1978). *Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory*. Springer–Verlag, Berlin Heidellberg New York, 1st edition.

Hájek, P., Holeňa, M., and Rauch, J. (2010). The GUHA method and its meaning for data mining. *Journal of Computer and System Science*, 76(1):34–48.

Ordonez, C., Ezquerra, N., and Santana, C. A. (2006). Constraining and summarizing association rules in medical data. *Knowledge and Information Systems (KAIS)*, 9(3):259–283.

Rauch, J. (2011). Consideration on a formal frame for data mining. In Hong, T.-P., Kudo, Y., Kudo, M., Lin, T. Y., Chien, B.-C., Wang, S.-L., Inuiguchi, M., and Liu, G., editors, *GrC*, pages 562–569. IEEE.

Rauch, J. (2013). *Observational Calculi and Association Rules*, volume 469 of *Studies in Computational Intelligence*. Springer.

Rauch, J. and Šimůnek, M. (2005). 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*, volume 6 of *Studies in Computational Intelligence*, pages 211–231. Springer.

Rauch, J. and Šimůnek, M. (2011). Applying domain knowledge in association rules mining process - first experience. In Kryszkiewicz, M., Rybinski, H., Skowron, A., and Ras, Z. W., editors, *ISMIS*, volume 6804 of *Lecture Notes in Computer Science*, pages 113–122. Springer.

Roddick, J. F., Fule, P., and Graco, W. J. (2003). Exploratory medical knowledge discovery: experiences and issues. *SIGKDD Explorations*, 5(1):94–99.