Language of Conclusions and Formal Framework for Data Mining with Association Rules

Jan Rauch

Faculty of Informatics and Statistics, University of Economics, Prague *

Abstract. FOFRADAR is a formal framework describing a process of data mining with association rules. Its purpose is to serve as a theoretical basis for automation of the data mining process. Association rule is understood as a couple of general Boolean attributes derived from columns of a data matrix and mutually related in an interesting way. FOFRADAR is based on a logical calculus of association rules, which is enhanced by languages and procedures making possible to deal with items of domain knowledge. Items of domain knowledge correspond to general expressions good understandable to domain experts. One of the languages of FOFRADAR is a language formulas which correspond to conclusions we can draw from results of data mining process. New features of this language are presented.

1 Introduction

A formal framework FOFRADAR (FOrmal FRAmework for Data mining with Association Rules) describing a process of data mining with association rules is introduced in [7]. Its goal is to describe the process such that formalized items of domain knowledge can be used both in formulation of reasonable analytical questions and in interpretation of results of a mining procedure. FOFRADAR is assumed to serve as a theoretical basis for the EverMiner project [9, 14].

The goal of the EverMiner project is to study data mining as a permanent knowledge driven process. It is assumed that there is a knowledge repository containing both relevant domain knowledge and hypotheses on new items of knowledge based on results of the analysis. We also assume there are tools that formulate reasonable data mining tasks, search in the analysed data for true patterns relevant to the formulated tasks, filter out found patterns which can be understood as the consequences of items of knowledge stored in the repository, synthesize hypotheses on new items of knowledge from the remaining patterns and store these hypotheses in the knowledge repository.

Let us emphasize that EverMiner is rather a long-term project which can bring interesting partial results. A natural part of the EverMiner project is a study of possibilities of automation of data mining. A formal description of the

* The work described here has been supported by Grant No. IGA 20/2013 of the University of Economics, Prague.

data mining process is a necessary prerequisite of its automation. We start with mining of association rules which are known patterns used in data mining.

However, we deal with more general association rules than introduced in [1]. The association rule is understood as an expression $\varphi \approx \psi$ where φ and ψ are Boolean attributes derived from columns of analysed data matrices and $\approx$ stands for a condition concerning a contingency table of φ and ψ [10]. Symbol $\approx$ is called *4ft-quantifier*. Boolean attributes are derived from basic Boolean attributes i.e expressions $A(\alpha)$. Here A is an attribute corresponding to a column of an analysed data matrix with possible values (i.e. categories) $a_1, \ldots, a_k$ and α is a subset of the set of categories, $\alpha \subset \{a_1, \ldots, a_k\}$. Basic Boolean attribute $A(\alpha)$ is true in a row o of a given data matrix $\mathcal{M}$ if $A(o) \in \alpha$ where $A(o)$ is a value of attribute A in row o. This means that we do not deal only with Boolean attributes - conjunctions of attribute-value pairs $A(a)$ where $a \in \{a_1, \ldots, a_k\}$ but with general Boolean attributes derived from columns of an analyzed data matrix. The 4ft-Miner procedure mines for such association rules [11].

FOFRADAR is a result of enhancing of a logical calculus of association rules [10] by additional languages and procedures. It is strongly related to the rules of the above introduced form $\varphi \approx \psi$ and to the possibilities of the 4ft-Miner procedure to mine for such rules. The goal of this paper is to present new considerations on language of formulas which correspond to conclusions of a data mining process. Before that main features of FOFRADAR are introduced. We proceed very informally, formal approach is introduced in [7, 8], see also [10].

The structure of the paper is as follows. An overview of FOFRADAR is in section 2. Particular languages and procedures of FOFRADAR are described in sections 3 – 6 in a way introduced in section 2. Language of conclusions is discussed in section 7. Remarks to related works are in section 8.

2 FOFRADAR Overview

FOFRADAR is sketched in Fig. 1 together with relations of its elements to the CRISP-DM. It is a result of an enhancement of a calculus of association rules by

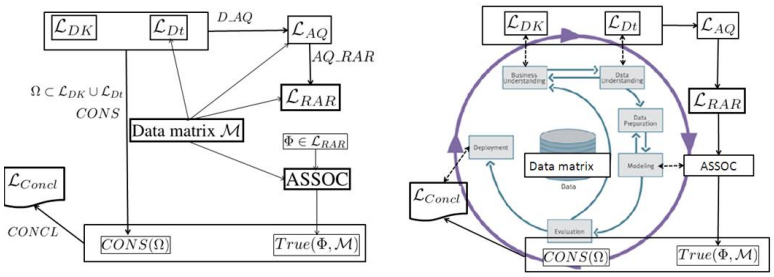

Fig. 1. *FOFRADAR* and CRISP-DM

several languages and procedures. Overview of elements of FOFRADAR follows.

Language $\mathcal{L}_{DK}$ – formulas of this language correspond to items of domain knowledge, they can be considered as results of business understanding. Language $\mathcal{L}_{DK}$ includes SI-formulas expressing mutual influence of attributes. Formula *BMI* $\uparrow\uparrow$ *Diastolic* meaning that if Body Mass Index of patients increases then diastolic blood pressure increases too is an example. Additional details are in section 3.

Language $\mathcal{L}_{Dt}$ – formulas of this language correspond to relevant information on analyzed data, see section 3. *Language* $\mathcal{L}_{AQ}$ – formulas of this language correspond to analytical questions. Analytical questions – formulas of $\mathcal{L}_{AQ}$ are formulated using formulas of languages $\mathcal{L}_{DK}$ and $\mathcal{L}_{Dt}$. This can be seen as an application of a *procedure* D_AQ, see section 4.

The core of the data mining process is the procedure ASSOC [3]. Its input consists of an analysed data matrix $\mathcal{M}$ and of a definition of a set of association rules to be verified to solve an analytical question given by a formula Θ of the language $\mathcal{L}_{AQ}$. The set of association rules to be verified is given by a formula Φ of a language $\mathcal{L}_{RAR}$ (i.e. a language of definitions of sets of Relevant Association Rules). This set is denoted as $\mathcal{S}(\Phi)$. We assume that the formula Φ is a result of an application of a procedure AQ_RAR to a given formula Θ, we write $\Phi = AQ\_RAR(\Theta)$. Output of the ASSOC procedure is a set $True(\mathcal{S}(\Phi), \mathcal{M})$ of all rules $\varphi \approx \psi$ which belong to $\mathcal{S}(\Phi)$ and which are true in $\mathcal{M}$. We use the procedure 4ft-Miner as an implementation of the ASSOC, see section 5.

An answer to a given analytical question is based on a comparison of the set $True(\mathcal{S}(\Phi), \mathcal{M})$ and a set $CONS(\Omega)$ of association rules which can be understood as consequences of a set Ω of items of domain or data knowledge used in the formulation of the solved analytical question Θ. An example of a set of consequences of SI-formula *BMI* $\uparrow\uparrow$ *Diastolic* is in section 6. Possible conclusion of analysis i.e. formulas of language $\mathcal{L}_{Concl}$ are introduced in section 7.

3 Asociation rules, Domain and Data Knowledge

We use a concrete data matrix to introduce a logical calculus of association rules in a very informal way. This is done in section 3.1, a formal definition is given in [10]. Language $\mathcal{L}_{DK}$ of domain knowledge is introduced in section 3.2. We will not describe the language $\mathcal{L}_{Dt}$ of data knowledge here in more details. Information that given data matrix *Entry* concerns only male patients is an example of an item of data knowledge.

3.1 Calculus $\mathcal{LC}_{\mathcal{E}}$ of Association Rules

We deal with association rules $\varphi \approx \psi$ where φ and ψ are Boolean attributes derived from basic Boolean attributes of the form $A(\alpha)$ and $\approx$ is a 4ft-quantifier. Here A is an attribute corresponding to a column of an analysed data matrix with possible values (i.e. categories) $a_1, \ldots, a_k$ and $\alpha \subset \{a_1, \ldots, a_k\}$. This means that a set of all basic Boolean attributes we can derive from a given column is

given by a set of categories for this column. Consequently, a set of all Boolean attributes concerning a given data matrix is determined by sets of possible values for particular columns of this data matrix.

We usually consider data matrices containing only natural numbers. There is only a finite number of possible values for each column. Let us assume that the number of possible values of a column is t and the possible values in this column are integers $1, \ldots, t$. Then all the possible values in a data matrix are described by the number of its columns and by the numbers of possible values for each column. These numbers determine a type of a data matrix and also a type of a logical calculus of association rules.

A *type of a logical calculus of association rules* is a K-tuple $\mathcal{T} = \langle t_1, \ldots, t_K \rangle$ where $K \geq 2$ is an integer and $t_i \geq 2$ are integers for $i = 1, \ldots, K$. A *data matrix of type* $\mathcal{T}$ has columns – attributes $A_1, \ldots, A_K$. Possible values (categories) for attribute A_i are $1, \ldots, t_i$ and *a type of attribute* A_i *is* t_i where $i = 1, \ldots, K$. A language $\mathcal{L}_\mathcal{T}$ of association rules of the type $\mathcal{T}$ is given by the attributes $A_1, \ldots, A_K$ and 4ft-quantifiers $\approx_1, \ldots, \approx_Q$. Association rule of $\mathcal{L}_\mathcal{T}$ is each rule $\varphi \approx \psi$ built from $A_1, \ldots, A_K$ and $\approx_1, \ldots, \approx_Q$. A logical calculus $\mathcal{LC}_\mathcal{T}$ of association rules of a type $\mathcal{T}$ consists of a language $\mathcal{L}_\mathcal{T}$, a set of all data matrices of type $\mathcal{T}$ and of instructions on how to decide if a given association rule is true in a given data matrix.

We use a data matrix *Entry* to introduce an example of a calculus of association rules. *Entry* is a part of the data set STULONG [1]. Data matrix *Entry* concerns 1 417 patients – men that have been examined at the beginning of the study. Each row of *Entry* describes one patient. *Entry* has 64 columns corresponding to particular attributes $A_1, \ldots, A_{64}$ – characteristics of patients. We use only first 12 attributes introduced in Tab. 1. These attributes and their categories have alternative names also introduced in Tab. 1 together with their frequencies in data matrix *Entry*. Types of these 12 attributes are also in Tab. 1. In Table 1, there is also a C-type for these attributes. C-type is introduced in the next section.

We can say that there is a type $\mathcal{T}_\mathcal{E} = \langle 4, 4, 4, 3, 3, 13, 7, 9, 10, 2, 2, 2, t_{13}, \ldots, t_{64} \rangle$ of a logical calculus $\mathcal{LC}_\mathcal{E}$ of association rules. Data matrix *Entry* is a data matrix of the type $\mathcal{T}_\mathcal{E}$ and each its update is also a data matrix of the type $\mathcal{T}_\mathcal{E}$. $A_1, \ldots, A_{64}$ are attributes of language $\mathcal{L}_\mathcal{E}$ of calculus $\mathcal{LC}_\mathcal{E}$. The alternative name of the attribute A_1 is *M_Status*, the alternative names of its categories 1,2,3,4 are *married*, *divorced*,*single*, *widover* respectively; similarly for additional attributes and categories. Let us note that there are missing values and thus the sum of frequencies of categories of particular attributes can be less than 1417.

[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/`

Table 1. Attributes and categories of $\mathcal{LC}_{\mathcal{E}}$ calculus of association rules

Attribute				Names of categories	
Name		Type	C-type		
Def.	Alternative			Definition	Alternative / frequency
A_1	*M_Status*	4	N	1,2,3,4	*married*/1207, *divorced*/104, *single*/95, *widover*/10
A_2	*Education*	4	O	1,2,3,4	*basic*/151, *apprentice*/405, *secondary*/444, *university*/397
A_3	*Responsibility*	4	N	1,2,3,4	*manager*/286, *independent*/435, *others*/636, *pensioner*/25
A_4	*Alcohol*	3	O	1,2,3	*no*/131, *occasionally*/748, *regularly*/462
A_5	*Coffee*	3	O	1,2,3	*no*/488, *1-2 cups*/643, *3+ cups*/258
A_6	*BMI*	13	O	1,..., 13	$(16;21\rangle/39$, $(21;22\rangle/50$, ..., ..., $(31;32\rangle/28$, $>32/64$
A_7	*Diastolic*	7	O	1, ..., 7	$\langle 50;70)/74$, $\langle 70;80)/281$, ..., ..., $\langle 110;120)/43$, $\langle 120;150)/16$
A_8	*Systolic*	9	O	1, ..., 9	$\langle 90;110)/69$, $\langle 110;120)/207$, ..., ..., $\langle 170;180)/43$, $\langle 180;220)/33$
A_9	*Cholesterol*	10	O	1, ..., 10	$\langle 100;160)/45$, $\langle 160;180)/97$, ..., ..., $\langle 300;320)/57$, $\langle 320;540)/57$
A_{10}	*Hypertension*	2	N	1,2	*yes*/220, *no*/1192
A_{11}	*Ictus*	2	N	1,2	*yes*/2, *no*/1408
A_{12}	*Infarction*	2	N	1,2	*yes*/34, *no*/1378
$A_{13}, \ldots, A_{64}$		see `http://euromise.vse.cz/challenge2004/data/entry/`			

Examples of basic Boolean attributes of the calculus $\mathcal{LC}_{\mathcal{E}}$ are: $A_1(1)$ – alternatively *M_Status*(*married*), $A_2(1,2)$ – alternatively *Education*(*basic*, *apprentice*), $A_6(1,2,3)$ – alternatively $BMI((16;21\rangle, (21;22\rangle, (21;22\rangle)$ i.e. $BMI(16;22\rangle$.

Examples of Boolean attributes are: *M_Status*(*married*) $\wedge$ $BMI(16;22\rangle$ and *Hypertension*(*yes*) $\vee$ *Ictus*(*yes*) $\vee$ *Infarction*(*yes*).

Association rule $\varphi \approx \psi$ is true in a data matrix $\mathcal{M}$ if a condition given by the 4ft-quantifier $\approx$ is satisfied for a contingency table of ψ and φ in $\mathcal{M}$. This contingency table is also called *4ft-table* $4ft(\varphi, \psi, \mathcal{M})$ *of* φ *and* ψ *in* $\mathcal{M}$ and denoted as $4ft(\varphi, \psi, \mathcal{M})$. It is is a quadruple $\langle a, b, c, d\rangle$ of non-negative integers where a is the number of rows of $\mathcal{M}$ satisfying both φ and ψ, b is the number of rows satisfying φ and not satisfying ψ etc., see Fig. 2. Two 4ft-quantifiers are

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

Fig. 2. 4ft table $4ft(\varphi, \psi, \mathcal{M})$ of φ and ψ in $\mathcal{M}$

introduced below, about 40 additional 4ft-quantifiers are defined in [3, 10].

4ft-quantifier $\Rightarrow_{p,B}$ of *founded implication* is defined for $0 < p \leq 1$ and $B > 0$ in [3] by the condition $\frac{a}{a+b} \geq p \wedge a \geq B$. Here $F_{\Rightarrow_{p,B}}$ is the associated function of $\Rightarrow_{p,B}$. Rule $\varphi \Rightarrow_{p,B} \psi$ means that at least $100p$ per cent of objects satisfying φ satisfy also ψ and that there are at least B rows of $\mathcal{M}$ satisfying both φ and ψ.

4ft-quantifier $\sim^+_{q,B}$ of *above average dependence* is for $0 < q$ and $B > 0$ defined in [10] by the condition $\frac{a}{a+b} \geq (1+q)\frac{a+c}{a+b+c+d} \wedge a \geq B$. Rule $\varphi \sim^+_{q,B} \psi$ means that among the rows satisfying φ, there are at least $100p$ per cent more rows atisfying ψ than among all rows of $\mathcal{M}$ and that there are at least B rows of $\mathcal{M}$ satisfying both φ and ψ.

This means that we can say that the calculus logical calculus $\mathcal{LC}_{\mathcal{E}}$ of association rules has two 4ft-quantifiers: $\Rightarrow_{p,B}$ and $\sim^+_{q,B}$. It is easy to add additional 4ft-quantifiers defined in [3, 10].

Correct deduction rules $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ where both $\varphi \approx \psi$ and $\varphi' \approx \psi'$ are association rules play a very important role in the FOFRADAR. Deduction rule $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ is correct if it holds for each data matrix $\mathcal{M}$: *if $\varphi \approx \psi$ is true in $\mathcal{M}$ then also $\varphi' \approx \psi'$ is true in $\mathcal{M}$*. The rules $\frac{A(1) \Rightarrow_{p,B} B(1)}{A(1) \Rightarrow_{p,B} B(1,2)}$ and $\frac{A(1) \Rightarrow_{p,B} B(1)}{A(1) \Rightarrow_{p,B} B(1) \vee C(1)}$ are very simple examples of correct deduction rules. There are relatively simple criteria making possible to decide if a given deduction rule $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ is correct. These criteria are known for most of important 4ft-quantifiers [10].

3.2 Language of Domain Knowledge

Language $\mathcal{L}_{DK}$ of domain knowledge is an enhancement of a language $\mathcal{L}_{\mathcal{T}}$ of a calculus $\mathcal{LC}_{\mathcal{T}}$ of association rules of type $\mathcal{T} = \langle t_1, \ldots, t_K \rangle$ [8]. The following items of domain knowledge can be expressed by formulas of $\mathcal{L}_{DK}$: (i) *C-types of basic attributes*, (ii) *groups of basic attributes*, (iii) *simple mutual influence of attributes*.

C-types of basic attributes are defined by a K-tuple $\mathcal{L}_{CT} = \langle \sigma_1, \ldots, \sigma_K \rangle$ where $\sigma_i \in \{N, O, C\}$ for $i = 1, \ldots, K$. This K-tuple is called *C-types of attributes*. If $\sigma_i = N$ then *the attribute A_i is nominal*, i.e., we do not assume to deal with ordering of its categories $1, \ldots, t_i$. If $\sigma_i = O$ then *the attribute A_i is ordinal* and we assume to use ordering of its categories $1, \ldots, t_i$. If $\sigma_i = C$ then *the attribute A_i is cyclical*, i.e., we assume ordering of its categories and in addition we assume that the category 1 follows the category t_i. An attribute *WeekDays* with categories *Su*, *Mo*, *Tu*, *We*, *Th*, *Fr*, *Sa* is an example of a cyclical attribute. C-types of attributes of calculus $\mathcal{LC}_{\mathcal{E}}$ are given in Tab. 1.

We use two types of groups of basic attributes – basic and additional. There are L *basic groups* $G_1, \ldots G_L$ *of basic attributes* – subsets of $\{A_1, \ldots, A_K\}$ satisfying $L < K$, $\cup_{i=1}^{L} G_i = \{A_1, \ldots, A_K\}$ and $G_i \cap G_j = \emptyset$ for $i \neq j$, $i, j = 1, \ldots L$. Calculus $\mathcal{LC}_{\mathcal{E}}$ has 11 basic groups, see `http://euromise.vse.cz/challenge2004/data/entry/`. The additional groups of basic attributes are usually defined for ad hoc analyses. We use here two such groups for calculus $\mathcal{LC}_{\mathcal{E}}$: group *Personal* consisting of 6 attributes *M_Status*, *Education*, *Responsibility*, *Alcohol Coffee*,

and *BMI* and group *Measurement* consisting of 3 attributes *Diastolic*, *Systolic*, and *Cholesterol*.

Mutual simple influence among attributes is expressed by *SI-formulas*. There are several types of SI-formulas. Below, we assume that A_i, A_j, $i \neq j$ are ordinal attributes and φ is a Boolean attribute. Examples of types of SI-formulas follow:

- *ii*-formula (i.e. increases - increases) $A_i \uparrow\uparrow A_j$ meaning *if* A_i *increases then* A_j *increases too*, *BMI* $\uparrow\uparrow$ *Diastolic* being an example
- *id*-formula (i.e. increases - decreases) $A_i \uparrow\downarrow A_j$ meaning *if* A_i *increases then* A_j *decreases*, *Education* $\uparrow\downarrow$ *Diastolic* being an example
- i^+b^+-formula has a form $A_i \uparrow^+ \varphi$ and its meaning is: *if A increases, then relative frequency of* φ *increases too*, *BMI* $\uparrow^+$ *Hypertension*(*yes*) being an example
- i^+b^--formula, i^-b^+-formula, i^-b^--formula have form $A_i \uparrow^- \varphi$, $A_i \downarrow^+ \varphi$, $A_i \downarrow^- \varphi$ respectively, their meaning is analogous to that of $A \uparrow^+ \varphi$.

4 From Domain Knowledge to Analytical Questions

Items of domain knowledge are used to formulate analytical questions. We introduce two types of such questions – GG-questions and negative GG_SI-question.

GG-question has form $[\mathcal{M} : G'_1, \ldots, G'_U \approx^? G''_1, \ldots G''_V]$ where $\mathcal{M}$ is a data matrix and $G'_1, \ldots, G'_U$ and $G''_1, \ldots G''_V$ are groups of attributes. A simple example is a formula Θ_1 defined as $\Theta_1 = [Entry : Personal \approx^? Measurement]$. Its meaning is: *In the data matrix* Entry*, are there any interesting relations between combinations of values of attributes of group Personal on one side and combinations of values of attributes of group Measurement on the other side?*

Negative GG_SI-question – its general form is
$[\mathcal{M} : (\Omega_1, \ldots, \Omega_P) \not\rightarrow G'_1, \ldots, G'_U \approx^? G''_1, \ldots G''_V]$ where $\mathcal{M}$, $G'_1, \ldots, G'_U$ and $G''_1, \ldots G''_V$ are as above and $\Omega_1, \ldots, \Omega_P$ are SI-formulas. A formula Θ_2 defined as $\Theta_2 = [Entry : BMI \uparrow\uparrow Diastolic \not\rightarrow Personal \approx^? Measurement]$ is a simple example of negative GG_SI-question. Its meaning is: *In the data matrix* Entry*, are there any interesting relations between combinations of values of attributes of group Personal on one side and combinations of values of attributes of group Measurement on the other side which are not consequences of BMI* $\uparrow\uparrow$ *Diastolic?*

The questions Θ_1 and Θ_2 are examples of formulas of a language $\mathcal{LAQ}_{\mathcal{E}}$ of analytical questions which is an enhancement of the language $\mathcal{L}_{\mathcal{E}}$ of calculus $\mathcal{LC}_{\mathcal{E}}$. Let us remember that a procedure DK_AQ is a part of the FOFRADAR. Its goal is to generate reasonable analytical questions. Input of DK_AQ consists of a list of groups and a list of SI-formulas of the language $\mathcal{L}_{DK}$. Let us only note that DK_AQ can be realized by suitable nested cycle statements.

We deal with the calculus $\mathcal{LC}_{\mathcal{E}}$. We assume that the only item of known domain knowledge is *BMI* $\uparrow\uparrow$ *Diastolic* and we are going to solve the question $\mathcal{Q}_2$. Our goal is to introduce in more details a language $\mathcal{L}_{Concl}$ of conclusions we can accept on results of mining association rules $\varphi \approx \psi$. Our goal is not to get new medical knowledge. Let as note that similar question is solved in [12], however, without details on a corresponding language $\mathcal{L}_{Concl}$.

5 Applying 4ft-Miner

5.1 Principles

We use the procedure 4ft-Miner [11] to solve the analytical question $\Theta_2 = [Entry : B\text{MI} \uparrow\uparrow D\text{iastolic} \not\rightarrow P\text{ersonal} \approx^? M\text{easurement}]$ introduced above. We deal with association rules, thus we formulate this question such that it deals with association rules:

$$\Theta_2 = [Entry : B\text{MI} \uparrow\uparrow D\text{iastolic} \not\rightarrow \mathcal{B}(P\text{ersonal}) \approx^? \mathcal{B}(M\text{easurement})] .$$

Here $\mathcal{B}(P\text{ersonal})$ denotes a set of all relevant Boolean attributes derived from attributes of the group *P*ersonal, similarly for $\mathcal{B}(M\text{easurement})$. We search association rules $\varphi_P \approx^? \psi_M$ which are true in data matrix *Entry*, cannot be understood as consequences $B\text{MI} \uparrow\uparrow D\text{iastolic}$, $\approx^?$ is a suitable 4ft-quantifier, $\varphi_P \in \mathcal{B}(P\text{ersonal})$ and $\psi_M \in \mathcal{B}(M\text{easurement})$.

There are very fine possibilities to define a set of relevant association rules, they are given by input parameters of 4ft-Miner. A definition of values of these parameters can be understood as an expression of a language $\mathcal{L}_{RAR}$. There are enough experience [10] making possible to construct a procedure *AQ_RAR* assigning to each analytical question Θ (i.e. a formula of language $\mathcal{L}_{AQ}$) a formula $AQ\_RAR(\Theta)$ of language $\mathcal{L}_{RAR}$ such that $AQ\_RAR(\Theta)$ defines parameters for a run of the 4ft-Miner procedure suitable to solve Θ.

In section 5.2 a formula $\Phi = AQ\_RAR(\Theta_2)$ is introduced. It defines a set $\mathcal{S}(\Phi)$ of association rules relevant to the question Θ_2. Input of the 4ft-Miner procedure consists of the formula Φ and data matrix *Entry*. The output is a set $True(\mathcal{S}(\Phi), Entry)$ of all rules $\varphi \approx \psi$ which belong to $\mathcal{S}(\Phi)$ and which are true in *Entry*, see section 5.3.

5.2 From Analytical Questions to Parameters of 4ft-Miner

In Fig. 3, there are input parameters of the 4ft-Miner. They can be seen as the formula $\Phi = AQ\_RAR(\Theta_2)$. We search for association rules $\varphi_P \approx^? \psi_M$ where $\approx^?$ is a suitable 4ft-quantifier, $\varphi_P \in \mathcal{B}(P\text{ersonal})$ and $\psi_M \in \mathcal{B}(M\text{easurement})$.

Let us also note that $\Phi = AQ\_RAR(\Theta_2)$ is constructed to get a reasonable output without necessity to modify parameters. Actually, the application of 4ft-Miner requires modifications of an initial setting of parameters. This can also be included into the *AQ_RAR* procedure. Description of this possibility is not the goal of this paper.

The set $\mathcal{B}(Personal)$ is defined in Fig. 3 in the column `ANTECEDENT` in the row `Personal Conj, 1-3` and in the six consecutive rows. This means that φ_P is a conjunction of 1 - 3 basic Boolean attributes derived from attributes of the group *P*ersonal introduced in section 3.2, see also Tab. 1.

A set of all basic Boolean attributes derived from attribute *M*_Status is defined by the row `M_Status(subset), 1-1 B, pos`. This means that basic Boolean attributes *M*_Status(*m*arried), *M*_Status(*d*ivorced), *M*_Status(*s*ingle),

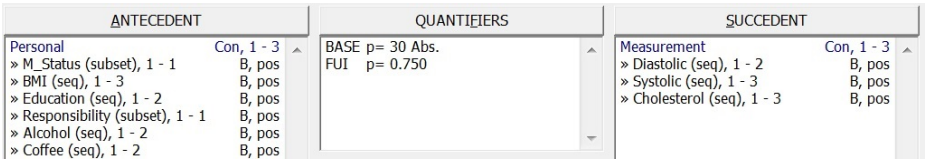

Fig. 3. Input parameters of the 4ft-Miner procedure

and $M\_\mathrm{Status}(\mathit{widover})$ are generated. A set of all basic Boolean attributes derived from attribute *R*esponsibility is defined similarly.

A set of all basic Boolean attributes $B\mathrm{MI}(\alpha)$ derived from attribute *B*MI is defined by the row `BMI(int), 1-3 B, pos`. This means that all $B\mathrm{MI}(\alpha)$ are generated such that α is a set of 1 - 3 consecutive categories (i.e. sequence of categories). Expression $B\mathrm{MI}(\langle 21;22\rangle, \langle 22;23\rangle)$ i.e. $B\mathrm{MI}(\langle 21;23\rangle$ is an example of such basic Boolean attribute. Sets of Boolean attributes derived from attributes *E*ducation, *A*lcohol and *C*offee are defined similarly.

The set $\mathcal{B}(\mathit{Measurement})$ is defined analogously in Fig. 3 in the column `SUCCEDENT`. In the column `QUANTIFIERS`, the quantifier $\approx^?$ is specified as a 4ft-quantifier $\Rightarrow_{0.75,30}$ of founded implication.

The formula $\Phi = \mathit{AQ\_RAR}(\Theta_2)$ can be seen as a triple $\langle ANT_{\Theta_2}, \Rightarrow_{0.75,30}, SUC_{\Theta_2}\rangle$ where ANT_{Θ_2} and SUC_{Θ_2} are definitions of sets $\mathcal{B}(ANT_{\Theta_2})$ and $\mathcal{B}(SUC_{\Theta_2})$ of Boolean attributes respectively. The triple $\langle ANT_{\Theta_2}, \Rightarrow_{0.75,30}, SUC_{\Theta_2}\rangle$ defines a set $\mathcal{S}(ANT_{\Theta_2}, \Rightarrow_{0.75,30}, SUC_{\Theta_2})$ of association rules $\varphi \Rightarrow_{0.75,30} \psi$ such that $\varphi \in \mathcal{B}(ANT_{\Theta_2})$, $\psi \in \mathcal{B}(SUC_{\Theta_2})$ and φ and ψ have no common basic attributes. The definitions ANT_{Θ_2} and SUC_{Θ_2} can be seen as sets of parameters in columns `ANTECEDENT` and `SUCCEDENTS` in Fig. 3 respectively. Then we have $\mathcal{B}(ANT_{\Theta_2}) = \mathcal{B}(\mathit{Personal})$ and $\mathcal{B}(SUC_{\Theta_2}) = \mathcal{B}(\mathit{Measurement})$.

5.3 4ft-Miner Output

The task specified in Fig. 3 was solved in 2 minutes (PC with 4GB RAM and Intel(R) Core(TM) i5-3320 processor at 2.6 GHz). 10^7 association rules were generated and tested, there are 341 output true rules. List of 10 rules with the highest confidence is in Fig. 4.

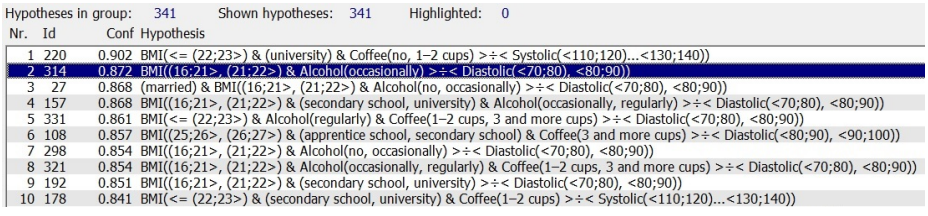

Fig. 4. Example of 4ft-Miner output

The rule $B\text{MI}(16;22) \wedge A\text{lcohol}(occasionally) \Rightarrow_{0.872,34} D\text{iastolic}\langle 70;90)$ is the second strongest one. This rule means that it holds in data matrix *Entry*: at least 87.2 per cent of patients satisfying $B\text{MI}(16;22) \wedge A\text{lcohol}(occsionally)$ satisfy also $D\text{iastolic}\langle 70;90)$ and there are at least 34 patients satisfying both $B\text{MI}(16;22) \wedge A\text{lcohol}(occsionally)$ and $D\text{iastolic}\langle 70;90)$, this information about 34 patients can be seen only in detailed output, not in Fig. 4.

Most of found rules have both the attribute *B*MI in antecedent (i.e. left part of a rule) and the attribute *D*iastolic in succedent (i.e. right part of a rule). We can expect that lot of such rules can be seen as consequences of SI-formula *B*MI $\uparrow\uparrow$ *D*iastolic introduced in section 3.2.

6 Consequences of SI-Formulas

Let Γ be an SI-formula and $\approx$ be a 4ft-quantifier, then $Cons(\Gamma, \approx)$ denotes a set of association rules which can be considered as consequences of Γ. The set $Cons(\Gamma, \approx)$ is defined in four steps [8].

1. A set $AC(\Gamma, \approx)$ of *atomic consequences of* Γ for $\approx$ is defined as a set of very simple rules $\kappa \approx' \lambda$ which can be, according to the domain expert, considered as direct consequences of Γ.
2. A set $AgC(\Gamma, \approx)$ of *agreed consequences of* Γ for $\approx$ is defined. A rule $\rho \approx' \sigma$ belongs to $AgC(\Gamma, \approx)$ if the following conditions are satisfied:
   - $\rho \approx' \sigma \notin AC(\Gamma, \approx)$
   - there is no $\kappa \approx' \lambda \in AC(\Gamma, \approx)$ such that $\rho \approx' \sigma$ logically follows from $\kappa \approx' \lambda$
   - there is $\kappa \approx' \lambda \in AC(\Gamma, \approx)$ such that, according to the domain expert, it is possible to agree that $\rho \approx' \sigma$ says nothing new in addition to $\kappa \approx' \lambda$.
3. A set $LgC(\Gamma, \approx)$ of *logical consequences of* Γ for $\approx$ is defined. A rule $\varphi \approx' \psi$ belongs to $LgC(\Gamma, \approx)$ if the following conditions are satisfied:
   - $\varphi \approx' \psi \notin (AC(\Gamma, \approx) \cup AgC(\Gamma, \approx))$
   - there is $\tau \approx' \omega \in AC(\Gamma, \approx) \cup AgC(\Gamma, \approx)$ such that $\varphi \approx' \psi$ logically follows from $\tau \approx' \omega$.
4. We define $Cons(\Gamma, \approx) = AC(\Gamma, \approx) \cup AgC(\Gamma, \approx) \cup LgC(\Gamma, \approx)$.

We outline a way in which a set $Cons(BMI \uparrow\uparrow Diastolic, \Rightarrow_{0.75,30})$ can be defined. Note that the 4ft-quantifier $\Rightarrow_{0.75,30}$ is used in the example in section 5. A rule $B\text{MI}(low) \Rightarrow_{0.75,30} D\text{iastolic}(low)$ saying that at least 75 per cent of patients satisfying $B\text{MI}(low)$ satisfy also $D\text{iastolic}(low)$ and that there are at least 30 patients satisfying both $B\text{MI}(low)$ and $D\text{iastolic}(low)$ can be considered as a simple consequence of *B*MI $\uparrow\uparrow$ *D*iastolic. In addition, if we consider $B\text{MI}(low) \Rightarrow_{0.75,80} D\text{iastolic}(low)$ as a consequence of *B*MI $\uparrow\uparrow$ *D*iastolic, then also each rule $B\text{MI}(low) \Rightarrow_{p,B} D\text{iastolic}(low)$ where $p \geq 0.75 \wedge B \geq 30$ is a consequence of *B*MI $\uparrow\uparrow$ *D*iastolic. The only problem is to define suitable coefficients *low* for both attributes *B*MI and *D*iastolic.

Attribute *B*MI has 13 categories: $(16;21\rangle$, $(21;22\rangle$, $(22;23\rangle$, $(23;24\rangle$, ..., $(31;32\rangle$, > 32. Attribute *D*iastolic has 7 categories: $\langle 50;70)$, $\langle 70;80)$, $\langle 80;90)$, ...,

$\langle 110;120), \langle 120;150)$. We can decide that each basic Boolean attribute $B\text{MI}(\alpha)$ satisfying condition $\alpha \subseteq \{(16;21\rangle, (21;22\rangle, (22;23\rangle, (23;24\rangle\}$ will be considered as $B\text{MI}(low)$, and similarly, each basic Boolean attribute $D\text{iastolic}(\beta)$ where $\beta \subseteq \{\langle 50;70), \langle 70;80), \langle 80;90)\}$ will be considered as $S\text{yst}(low)$. We can say that rules $B\text{MI}(low) \Rightarrow_{0.75,30} D\text{iastolic}(low)$ are defined by a rectangle $\mathcal{A}_{low} \times \mathcal{S}_{low}$:

$$\mathcal{A}_{low} \times \mathcal{S}_{low} = \{(16;21\rangle, (21;22\rangle, (22;23\rangle\ (23;24\rangle\} \times \{\langle 50;70), \langle 70;80), \langle 80;90)\}\ .$$

The 4ft-Miner procedure is accompanied by the *LMDataSource* module which makes possible to define the set $AC(B\text{MI} \uparrow\uparrow D\text{iastolic}, \Rightarrow_{0.75,30})$ 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 $B\text{MI}(\alpha) \Rightarrow_{p,B} D\text{iastolic}(\beta) \in AC(BMI \uparrow\uparrow Diastolic, \Rightarrow_{0.75,30})$ if and only if it holds $p \geq 0.75 \wedge B \geq 30$ and there is an $i \in \{1, \ldots, R\ \}$ such that $\alpha \subseteq \mathcal{A}_i$ and $\beta \subseteq \mathcal{S}_i$. An example is in Fig. 5, four rectangles are used. Very informally speaking, we can see $AC(BMI \uparrow\uparrow Diastolic, \Rightarrow_{0.75,30})$ as a union

$$\mathcal{A}_{low} \times \mathcal{S}_{low} \cup \mathcal{A}_{below\ avrg} \times \mathcal{S}_{below\ avrg} \cup \mathcal{A}_{above\ avrg} \times \mathcal{S}_{above\ avrg} \cup \mathcal{A}_{high} \times \mathcal{S}_{high}$$

where *avrg* abbreviates *average*.

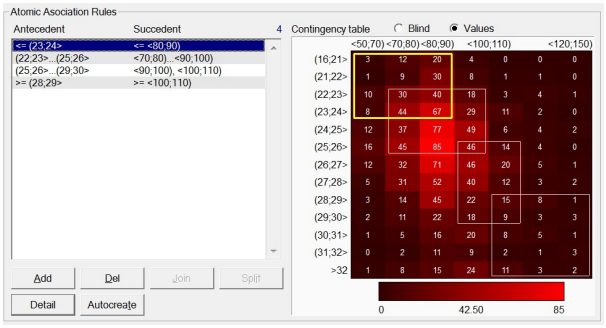

Fig. 5. Definition of $AC(BMI \uparrow\uparrow Syst, \Rightarrow_{0.7,80})$

The rule $B\text{MI}(16;22\rangle \wedge A\text{lcohol}(occasionally) \Rightarrow_{0.87,34} D\text{iastolic}\langle 70;90)$ is an example of an agreed consequence. This rule does not logically follow from an atomic consequence $B\text{MI}(16;22\rangle \Rightarrow_{0.87,34} D\text{iastolic}\langle 70;90)$ but it is possible to agree that it says nothing new to the rule $B\text{MI}(16;22\rangle \Rightarrow_{0.87,34} D\text{iastolic}\langle 70;90)$.

The rule $B\text{MI}(16;22\rangle \wedge M\_\text{Status}(married) \Rightarrow_{0.78,31} D\text{iastolic}\langle 80;100)$ is an example of a logical consequence of an agreed consequence. This is because $B\text{MI}(16;22\rangle \Rightarrow_{0.78,31} D\text{iastolic}\langle 80;90)$ is an atomic consequence, $B\text{MI}(16;22\rangle \wedge M\_\text{Status}(married) \Rightarrow_{0.78,31} D\text{iastolic}\langle 80;90)$ is its agreed consequence and $B\text{MI}(16;22\rangle \wedge M\_\text{Status}(married) \Rightarrow_{0.78,31} D\text{iastolic}\langle 80;100)$ is a logical consequence of $B\text{MI}(16;22\rangle \wedge M\_\text{Status}(married) \Rightarrow_{0.78,31} D\text{iastolic}\langle 80;90)$.

This way, a set $Cons(BMI \uparrow\uparrow Diastolic, \Rightarrow_{0.75,30})$ is produced. Let us note that deduction rules $\frac{\varphi \Rightarrow_{p,B} \psi}{\varphi' \Rightarrow_{p,B} \psi'}$ (see end of section 3.1) play a crucial role in these

considerations. Additional examples of these considerations are in [8, 12]. Similar considerations are valid for additional 4ft-quantifiers.

We can summarise: The input of the 4ft-Miner is a triple $\langle ANT, \approx, SUC\rangle$ and a data matrix $\mathcal{M}$. The triple $\langle ANT, \approx, SUC\rangle$ defines a set $\mathcal{S}(ANT, \approx, SUC)$ of rules. Output of the 4ft-Miner is a set $True(\mathcal{S}(ANT, \approx, SUC), \mathcal{M})$ of all rules from $\mathcal{S}(ANT, \approx, SUC)$ which are true in $\mathcal{M}$. We have outlined that there is a procedure $CONS$ input of which is an SI-formula Γ and a 4ft-quantifier $\approx$ and output of $CONS$ is a set $Cons(\Gamma, \approx)$ of all rules belonging to $\mathcal{S}(ANT, \approx, SUC)$ which can be considered as consequences of Γ. In addition, we have introduced a set $Cons(BMI \uparrow\uparrow Diastolic, \Rightarrow_{0.75,30})$.

7 Language of Conclusions

The goal of this section is to introduce formulas of the language $\mathcal{L}_{Concl}$ which represent conclusions of analysis. The conclusions are formulated on the basis of a comparison of a set $True(\mathcal{S}(ANT, \approx, SUC), \mathcal{M})$ resulting from the run of the procedure 4ft-Miner and sets $Cons(\Gamma, \approx)$ for relevant SI-formulas Γ. An SI-formula Γ is relevant if it is used in the analytical question or if it can be formulated from the attributes which occur in $True(\mathcal{S}(ANT, \approx, SUC), \mathcal{M})$.

When dealing with a particular SI-formula Γ, we are interested in relations of sets of rules $True(\mathcal{S}(ANT, \approx, SUC), \mathcal{M})$ and $Cons(\Gamma, \approx)$. The conclusions can be formulated on the basis of their intersection $True(\mathcal{S}(ANT, \approx, SUC), \mathcal{M}) \cap Cons(\Gamma, \approx)$ and difference $True(\mathcal{S}(ANT, \approx, SUC), \mathcal{M}) \setminus Cons(\Gamma, \approx)$. These can be also produced by the 4ft-Miner procedure. In addition, it is possible to sort and filter rules of these sets in various ways. Relations of $True(\mathcal{S}(ANT, \approx, SUC), \mathcal{M})$ and $Cons(\Gamma, \approx)$ can be also investigated by SQL tools. This can lead to variety of conclusions on result of application of the procedure 4ft-Miner.

We outline four of them. We use the analytical question Θ_2 introduced in section 4 as $\Theta_2 = [Entry : B\text{MI} \uparrow\uparrow D\text{iastolic} \not\rightarrow P\text{ersonal} \approx^? M\text{easurement}]$. A formula $AQ\_RAR(\Theta_2)$ of language $\mathcal{L}_{RAR}$ is introduced in sections 5.1 and 5.2. It defines a set $\mathcal{S}(AQ\_RAR(\Theta_2))$ of association rules we consider relevant to the question Θ_2. The formula $AQ\_RAR(\Theta_2)$ is specified as $\langle ANT_{\Theta_2}, \Rightarrow_{0.75,30}, SUC_{\Theta_2}\rangle$ where $\mathcal{B}(ANT_{\Theta_2}) = \mathcal{B}(Personal)$ and $\mathcal{B}(SUC_{\Theta_2}) = \mathcal{B}(Measurement)$. The procedure 4ft-Miner produces the set $True(\mathcal{S}(ANT_{\Theta_2}, \Rightarrow_{0.75,30}, SUC_{\Theta_2}), Entry)$ of all relevant association rules true in the data matrix *Entry*. We denote it as $True(\Theta_2, Entry)$.

We denote a difference $True(\Theta_2, Entry) \setminus Cons(BMI \uparrow\uparrow Diastolic, \Rightarrow_{0.75,30})$ of sets $True(\Theta_2, Entry)$ and $Cons(BMI \uparrow\uparrow Diastolic, \Rightarrow_{0.75,30})$ as Dif_Tr_Con. We assume that both the set $Cons(BMI \uparrow\uparrow Diastolic, \Rightarrow_{0.75,30})$ and the set $True(\Theta_2, \Rightarrow_{0.75,30}, Entry)$ are not empty. Then one or more from the following possibilities P1, P2, P3, P4 can occur:

P1: $True(\Theta_2, \Rightarrow_{0.75,30}, Entry) = Cons(BMI \uparrow\uparrow Diastolic, \Rightarrow_{0.75,30})$. This means that the conclusion can be: *In the data matrix* Entry*, all potentially interesting relations between combinations of values of attributes of group Personal*

on one side and combinations of values of attributes of group Measurement on the other side are consequences of BMI ↑↑ Diastolic.

P2: There is a rule $\varphi \Rightarrow_{p,B} \psi \in Dif\_Tr\_Con$ (where $p \geq 0.75$ and $B \geq 30$) such that attribute *BMI* does not occur in φ or attribute *Diastolic* does not occur in ψ. The rule $Education(apprentice) \wedge Responsibility(independent) \Rightarrow_{0.78,54} Diastolic\langle 80;100\rangle$ is an example. This means that the conclusion can be: *In the data matrix* Entry*, there are the following interesting true rules which are not consequences of BMI ↑↑ Diastolic:* a list of rules $\varphi \Rightarrow_{p,B} \psi \in Dif\_Tr\_Con$ satisfying that attribute *BMI* does not occur in φ or attribute *Diastolic* does not occur in ψ *follows.*

P3: There is a rule $\varphi \Rightarrow_{p,B} \psi \in Dif\_Tr\_Con$ (where $p \geq 0.75$ and $B \geq 30$) such that attribute *BMI* occurs in φ and attribute *Diastolic* occurs in ψ. This rule is then true in data matrix *Entry* and it is not a consequence BMI ↑↑ Diastolic, thus we can consider it as an exception to BMI ↑↑ Diastolic. This means that the conclusion can be: *In the data matrix* Entry*, there are the following interesting true rules which can be considered as exceptions from BMI ↑↑ Diastolic:* a list of rules $\varphi \Rightarrow_{p,B} \psi \in Dif\_Tr\_Con$ such that attribute *BMI* occurs in φ and attribute *Diastolic* occurs in ψ *follows.*

This approach to exceptions differs from that in [15]. Let us note that we can modify the definition of $Cons(BMI \uparrow\uparrow Diastolic, \Rightarrow_{0.75,30})$ such that the rule $BMI(21;22\rangle \wedge M\_Status(married) \Rightarrow_{0.77,36} Diastolic\langle 80;100\rangle$ which is true in *Entry* does not belong to $Cons(BMI \uparrow\uparrow Diastolic, \Rightarrow_{0.75,30})$.

P4: There is an additional SI-formula Γ' such that there are enough rules in $Cons(\Gamma', \Rightarrow_{0.75,30}) \cap True(\Theta_2, \Rightarrow_{0.75,30}, Entry)$. An example is the SI-formula BMI ↑↑ Systolic. We can define a set $Cons(BMI \uparrow\uparrow Systolic, \Rightarrow_{0.75,30})$ such that there are 182 rules $\varphi \Rightarrow_{p,B} \psi$ true in *Entry* where $p \geq 0.75$, $B \geq 30$, attribute *BMI* occurs in φ and attribute *Systolic* occurs in ψ and all of these rules belong to $Cons(BMI \uparrow\uparrow Systolic, \Rightarrow_{0.75,30})$. This means that the conclusion can be (here we consider $BMI \uparrow\uparrow Systolic$ as an unknown item of domain knowledge): *In the data matrix* Entry*, there are lot of rules which can be considered as consequences of yet unknown item of knowledge BMI ↑↑ Systolic. It is reasonable to investigate BMI ↑↑ Systolic as a working hypothesis.*

Let us note that additional conclusions can be formulated on the basis of a difference $Cons(\Gamma, \approx) \setminus True(\mathcal{S}(ANT, \approx, SUC), \mathcal{M})$ of sets $Cons(\Gamma, \approx)$ and $True(\mathcal{S}(ANT, \approx, SUC), \mathcal{M})$.

8 Related work

The FOFRADAR framework deals with association rules of the form $\varphi \approx \psi$ where φ and ψ are general Boolean attributes derived from columns of analysed data matrices and $\approx$ stands for a general condition concerning a contingency table of φ and ψ. A crucial feature of this approach is dealing with basic Boolean attributes of the form $A(\alpha)$ where α is a *subset of categories*. This makes possible to deal with rules like BMI(low) $\Rightarrow_{p,B}$ Diastolic(low) where "*low*" stands for a suitable subsets of categories. This way, dealing with items of domain knowledge

like BMI $\uparrow\uparrow$ Diastolic can be converted to dealing with suitable association rules when suitable deduction rules $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ are applied, see section 6 and also section 5.2.

An additional important feature of rules $\varphi \approx \psi$ is a possibility to deal with disjunction of basic Boolean attributes. This is especially important when dealing with attributes with low frequencies. Examples of such attributes of data matrix *Entry* are *Hyperlipidemia*(*yes*) – with frequency 54, *Diabetes*(*yes*) – 30, *Ictus*(*yes*) – 2, *Infarction*(*yes*) – 34. All these frequencies are very low and thus it is crucial to deal with disjunction of these attributes instead of with conjunctions, the frequencies of conjunctions are almost null. In addition, disjunctions *Hyperlipidemia*(*yes*) $\vee$ *Diabetes*(*yes*), *Hyperlipidemia*(*yes*) $\vee$ *Ictus*(*yes*), ... can be interpreted as "patient has problems". An example of dealing with such conjunctions is in [12]. Disjunctions of basic Boolean attributes are also involved in the FOFRADAR approach, deduction rules $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ are very important in this case.

"Classical" association rules introduced in [1] deal neither with basic Boolean attributes $A(\alpha)$ nor disjunctions of basic Boolean attributes. Thus the approach to dealing with domain knowledge used in FOFRADAR cannot be fully applied when dealing with "classical" association rules. Let us note that the procedure 4FT introduced in [6] mines even for more general rules than the 4ft-Miner procedure used in this paper.

An approach to use ontologies expressing a hierarchy to interpret a resulting set of "classical" association rules is described in [4]. Very informally speaking, this is based on partitioning a set of resulting rules to sets of known rules, novel rules, missing rules and contradictory rules depending on their relation to an ontology in question. This is similar to assigning a set of consequences to an SI-formula as described in section 6. However, domain knowledge expressed by SI-formulas differs from domain knowledge expressed by a given ontology.

The goal of FOFRADAR is to be a formal description of the data mining process with association rules. Thus, using ontologies in FOFRADAR approach is a challenge. Additional challenge is related to inclusion of suitable means of inductive databases languages [13] as well as the additional approaches to deal with domain knowledge in data mining with "classical" association rules, see e.g. [2, 5].

9 Conclusions

We have presented new considerations on language $\mathcal{L}_{Concl}$. Formulas of this language correspond to conclusions of data mining process with association rules. $\mathcal{L}_{Concl}$ is one of languages of the formal framework FOFRADAR the goal of which is to formally describe the whole process of data mining with association rules. FOFRADAR is intended as a theoretical basis for automation of data mining process with association rules. The results presented here will be used together with former results [7, 8, 11, 12] to start experiments with automation of data mining process with association rules.

The FOFRADAR framework deals with association rules which are substantially more general than "classical" association rules defined in [1] and used in mainstream applications of association rules. There are various approaches to deal with domain knowledge in "classical" association rules data mining which can be included to FOFRADAR. However, this is rather a long process requiring additional effort.

References

1. Agrawal, R., Imielinski, T., Swami, A.: (1993) Mining Associations between Sets of Items in Large Databases. In: Buneman, P., Jajodia, S. (eds.) Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207–216. ACM Press, Fort Collins
2. Delgado, M. et al.: (1998) Mining association rules with improved semantics in medical databases, *Artificial Intelligence in Medicine*, **21**(1–3), 2001, pp. 241–245
3. Hájek, P., Havránek, T.: (1978) Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory. Springer, Berlin Heidelberg New York
4. Mansingh, G., Osei-Bryson, K.-M., Reichgelt. H.: (2011) Using ontologies to facilitate post-processing of association rules byh domain experts. Information Sciences, **181**(3), pp. 419–434
5. Ordonez, C., Ezquerra, N., Santana, C.A.: (2006) Constraining and Summarizing Association Rules in Medical Data, *Knowledge and Information Systems (KAIS)*, **9**(3), pp. 259–283.
6. Ralbovský, M., Kuchař, T.: (2009) Using Disjunctions in Association Mining. In: Perner, P. (ed.) Advances in Data Mining, Springer, Berlin, pp. 339–351
7. Rauch, J.: (2012) Formalizing Data Mining with Association Rules. In: IEEE International Conference on Granular Computing. Los Alamitos : IEEE Computer Society, 2012, pp. 485–490.
8. Rauch, J.: (2012) Domain Knowledge and Data Mining with Association Rules - a Logical Point of View. In: Foundations of Intelligent Systems. Springer, Berlin, pp. 11–20
9. Rauch, J.: (2012) EverMiner: consideration on knowledge driven permanent data mining process. International Journal of Data Mining, Modelling and Management, **4** (3), pp. 224–243
10. Rauch, J.: (2013) Observational Calculi and Association Rules. Springer, Berlin
11. Rauch, J., Šimůnek, M.: (2005) An Alternative Approach to Mining Association Rules. In: Lin, T. Y. et al. (eds) Data Mining: Foundations, Methods, and Applications, Springer, Berlin, pp. 219–238
12. Rauch, J., Šimůnek, M.: (2011) Applying Domain Knowledge in AssociationRules Mining Process - First Experience. In: Kryszkiewicz, M. et al.: (eds) Foundations of Intelligent Systems, Springer, Berlin, pp. 113–122
13. Romei, A., Turini, F.: (2011) Inductive database languages: requirements and examples. Knowledge and Information Systems, **26**(3), pp. 351–384
14. Šimůnek, M., Rauch J.: EverMiner Towards Fully Automated KDD Process. In: Funatsu, K., Hasegava, K. (eds.) New Fundamental Technologies in Data Mining, pp. 221-240. InTech, Rijeka (2011)
15. Suzuki, E. (2004) Discovering interesting exception rules with rule pair. In J. Fuernkranz (Ed.), Proceedings of the ECML/PKDD Workshop on Advances in Inductive Rule Learning pp. 163–178