Observational Calculi, Classes of Association Rules and F-property

Jan Rauch
Faculty of Informatics and Statistics,* University of Economics, Prague
nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic
rauch@vse.cz

Abstract

An overview of several classes of association rules is given. It is shown that these classes have theoretically interesting and practically useful properties. The class of association rules with the F-property is introduced. It is shown that association rules from this class have properties similar to properties of the Fisher's test. Results concerning the association rules with the F-property are presented.

1 Introduction

This paper tries to contribute to foundations of granular computing. The paper deals with observational calculi (OC for short) which were defined and studied in [2] as a theoretical basis for the GUHA method of mechanizing hypotheses formation. GUHA is an original Czech method of exploratory data analysis developed since the 1960s, that is closely related to granular computing; see Section 2.

The most simple OC is *predicate observational calculus* which is a modification of classical predicate calculus (gene-ralized quantifiers are added and only finite structures as models are allowed). Formulas of OC correspond to various assertions concerning analyzed data matrices. There are also the assertions corresponding to statistical hypotheses tests. OC were further modified and studied to get the additional results which help to better understand association rules. We mean not only "classical" association rules defined in [1] as a pattern describing market baskets. The association rule we are interested in is a relation of two Boolean attributes derived in a general way from columns of analyzed data matrix; see Section 3.

Both theoretically interesting and practically important results about deduction rules concerning association rules, definability of association rules in predicate calculus, and about association rules with missing information were achieved. These results are related to classes of association rules. There are classes of implicational rules, double implicational rules, etc.; see Sections 4 and 5.

Goal of this paper is to present results concerning class of association rules with F-property. This class contains important association rules corresponding to Fisher's test and χ^2 test; see Section 6. Some of these results were published in last 30 years under various names, some of them are new. Overview of these results is in Section 7. Results concerning relation of classes of Σ-equivalence rules and rules with F-property are in Section 8. Several concluding remarks are in Section 9.

2 GUHA and Granular Computing

Aim of the GUHA method is to offer all interesting facts following from the analyzed data to the given problem. GUHA is realized by the GUHA-procedures. The GUHA-procedure is a computer program, the input of which consists of the analyzed data and of a simple definition of relevant (i.e. potentially interesting) patterns. The procedure automatically generates each particular pattern and tests if it is true in the analyzed data. The output of the procedure consists of all prime patterns. The pattern is prime if it is true in the analyzed data and if it does not immediately follow from the other, more simple output patterns [2].

Several GUHA procedures were implemented; see e.g. [3, 4, 16]. The most important GUHA procedure is the procedure ASSOC [2] which mines for association rules. The association rules the procedure ASSOC mines for are more general than the classical association rules defined in [1]. Procedure ASSOC mines among other for association rules corresponding to statistical hypothesis tests. There are several implementations of the procedure ASSOC; see e.g. [3, 4, 6, 8, 16]. Namely the last two implementations have significantly more tools for a definition and tuning of a set of rules to be automatically generated and tested than the usual apriori algorithm has.

The association rule is the expression $\varphi \approx \psi$ where φ and ψ are Boolean attributes. It means that φ and ψ are

*The work described here has been supported by Grant No. 201/05/0325 of the Czech Science Foundation and by Grant No. ME913 of Ministry of Education, Youth and Sports of the Czech Republic

0-7695-3032-X/07 $25.00 © 2007 IEEE
DOI 10.1109/GrC.2007.88

associated in the way given by the symbol $\approx$ called the *4ft-quantifier*. The association rule $\varphi \approx \psi$ concerns the analyzed data matrix $\mathcal{M}$. The rule $\varphi \approx \psi$ is *true in the data matrix* $\mathcal{M}$ if the condition given by the 4ft-quantifier $\approx$ is satisfied in the four-fold contingency table of φ and ψ in $\mathcal{M}$, otherwise $\varphi \approx \psi$ is *false in the data matrix* $\mathcal{M}$.

The four-fold contingency table $4ft(\varphi, \psi, \mathcal{M})$ (the *4ft table* for short) of φ and ψ in the data matrix $\mathcal{M}$ is the quadruple $\langle a, b, c, d \rangle$ of natural numbers such that 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.

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

An example of the data matrix with the columns - the attributes $A, B, C, \dots$ is the data matrix $\mathcal{M}$ in Fig. 1. There are also examples of the basic Boolean attributes $A(2)$ and $B(5, 6)$ in Fig. 1.

object	A	B	C	$\dots$	$A(2)$	$B(5,6)$
o_1	2	7	16	$\dots$	1	0
o_2	7	5	4	$\dots$	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
o_n	1	6	5	$\dots$	0	1

Figure 1. Data matrix $\mathcal{M}$ and basic Boolean attributes $A(2)$, $B(5, 6)$

All implementations of the ASSOC procedure are based on the *representation of the analyzed data by strings of bits* [16]. The apriori approach is not used. Each category of each attribute is represented by a string of bits, there is one bit for each row of the data matrix. If the category a of the attribute A is represented by the string ω_a of bits then the i-th bit in ω_a is "1" if and only if it is $A(o_i) = a$ where $A(o_i)$ is the value of the attribute A for the i-th row o_i of the data matrix $\mathcal{M}$.

This representation has an important effect - it is possible to easy define and automatically generate various both basic and derived Boolean attributes. Let us suppose that the attribute AGE has categories $1, \dots, 100$. The set of relevant basic Boolean attributes $AGE\langle\alpha; \alpha + 9\rangle$ can be then easy generated. It consists of the basic Boolean attributes $AGE\langle 1; 10\rangle, AGE\langle 2; 11\rangle, \dots, AGE\langle 91; 100\rangle$. Such set of the basic Boolean attributes is usually called *sliding window*.

There are lot of possibilities how to define sets of relevant basic Boolean attributes and consequently sets of relevant association rules. Bit string approach makes possible to generate and verify them in a very efficient way [16].

An additional and very important effect of the fact that GUHA is not realized by apriori concerns interesting association rules that are true even if their confidence and support are extremely low. An example of such rule concerns the 4ft quantifier $\Rightarrow^{+}_{p,B}$; see section 3 and [18]. It is possible to skip uninteresting rules with low confidence already in the phase of generation and verification, it is not necessary to filter very large set of rules with extremely low confidence and/or support.

The bit string approach is used in additional five GUHA procedures implemented in the LISp-Miner system [17]. These procedures mine for various patterns verified using one or two contingency tables. General contingency tables concerning not only Boolean attributes are used.

The particular categories, both basic and derived Boolean attributes and patterns the GUHA procedures mine for can be understood as granules. The power of the GUHA methods implemented using bit string representation of analyzed data is in ability to easy define and verify large sets of patterns (i.e. the granules composed from hierarchical system of more simple granules). Formulas of observational calculi correspond to patterns the GUHA procedures deal with and thus logic of observational calculi is a kind of logic of granules.

3 Association Rules

The association rule is the expression $\varphi \approx \psi$ where φ and ψ are Boolean attributes and $\approx$ is the *4ft-quantifier*. The rule $\varphi \approx \psi$ is true in the analyzed data matrix $\mathcal{M}$ if the condition related to the 4ft-quantifier $\approx$ is satisfied in the 4ft table $4ft(\varphi, \psi, \mathcal{M})$ of φ and ψ in $\mathcal{M}$ see Tab. 1 and Section 2. Some important 4ft-quantifiers are presented below. We use a, b, c, d see Tab. 1. In addition we use $r = a + b$, $k = a + c$ and $n = a + b + c + d$.

The 4ft-quantifier $\Rightarrow_{p,B}$ of *founded implication* is for

$0 < p \leq 1$ and $B > 0$ defined in [2] by the condition $\frac{a}{a+b} \geq p \wedge a \geq B$. The 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 objects of $\mathcal{M}$ satisfying both φ and ψ.

The 4ft-quantifier $\Rightarrow^{!}_{p,\alpha,B}$ of *lower critical implication* is for $0 < p \leq 1$, $0 < \alpha < 0.5$, and $B > 0$ defined in [2] by $\sum_{i=a}^{a+b} \binom{a+b}{i} p^i (1-p)^{a+b-i} \leq \alpha \wedge a \geq Base$. The rule $\varphi \Rightarrow^{!}_{p,\alpha,Base} \psi$ corresponds to the statistical test (on the level α) of the null hypothesis $H_0 : P(\psi|\varphi) \leq p$ against the alternative one $H_1 : P(\psi|\varphi) > p$. Here $P(\psi|\varphi)$ is the conditional probability of the validity of ψ under the condition φ.

The 4ft-quantifier $\Leftrightarrow_{p,B}$ of *founded double implication* is for $0 < p \leq 1$ and $B > 0$ defined in [5] by the condition $\frac{a}{a+b+c} \geq p \wedge a \geq B$. The rule $\varphi \Leftrightarrow_{p,B} \psi$ means that at least $100p$ per cent of objects satisfying φ or ψ satisfy both φ and ψ and that there are at least B objects of $\mathcal{M}$ satisfying both φ and ψ.

The 4ft-quantifier $\equiv_{p,B}$ of *founded equivalence* is for $0 < p \leq 1$ and $B > 0$ defined in [5] by the condition $\frac{a+d}{n} \geq p \wedge a \geq B$. The rule $\varphi \equiv_{p,B} \psi$ means that φ and ψ have the same value (either *true* or *false*) for at least $100p$ per cent of all objects of $\mathcal{M}$ and that there are at least B objects satisfying both φ and ψ.

The Fisher's quantifier $\sim_{\alpha,B}$ is for $0 < \alpha < 0.5$ and $B > 0$ defined in [2] by $\sum_{i=a}^{\min(r,k)} \frac{\binom{k}{i}\binom{n-k}{r-i}}{\binom{r}{n}} \leq \alpha \wedge ad > bc \wedge a \geq B$. The rule $\varphi \sim_{\alpha,B} \psi$ corresponds to the statistical test (on the level α) of the null hypothesis of independence of φ and ψ against the alternative one of the positive dependence.

The 4ft-quantifier $\Rightarrow^{+}_{p,B}$ of *above average dependence* is for $0 < p$ and $B > 0$ defined in [16] by the condition $\frac{a}{a+b} \geq (1+p)\frac{a+c}{n} \wedge a \geq B$. The rule $\varphi \Rightarrow^{+}_{p,B} \psi$ means that among the objects satisfying φ is at least $100p$ per cent more objects satisfying ψ than among all objects and that there are at least B objects satisfying both φ and ψ.

We write $\approx (a,b,c,d) = 1$ if the condition corresponding to the 4ft-quantifier $\approx$ is satisfied for the quadruple $\langle a,b,c,d \rangle$ of integer non-negative numbers, otherwise we write $\approx (a,b,c,d) = 0$.

4 Classes of Association Rules

Classes of association rules are defined by classes of 4ft quantifiers. An example is the class of *implicational rules*. The association rule $\varphi \approx \psi$ belongs to the *class of implicational rules* if the 4ft quantifier $\approx$ belongs to the *class of implicational quantifiers*. If the 4ft quantifier $\approx$ belongs to the *class of implicational quantifiers* then we say that $\approx$ is the *implicational 4ft quantifier* etc.

Several important classes of 4ft quantifiers are defined by *Truth Preservation Conditions*. We call these quantifiers $\mathcal{TPC}$ quantifiers. Each class Ω of $\mathcal{TPC}$ quantifiers is defined according to this scheme:

The 4ft quantifier $\approx$ belongs to the class Ω if $\approx (a,b,c,d) = 1 \wedge TPC_{\Omega}$ implies $\approx (a',b',c',d') = 1$ for all the 4ft tables $\langle a,b,c,d \rangle$, $\langle a',b',c',d' \rangle$ satisfying TPC_{Ω}. Here TPC_{Ω} is the true preservation condition for the class Ω. There are the following classes of $\mathcal{TPC}$ quantifiers:

Implicational quantifiers defined in [2] by $TPC_{\Rightarrow}$: $a' \geq a \wedge b' \leq b$ (examples: $\Rightarrow_{p,B}$ and $\Rightarrow^{!}_{p,\alpha,B}$).

Σ-*double implicational quantifiers* defined in [14] by $TPC_{\Sigma,\Leftrightarrow}$: $a' \geq a \wedge b' + c' \leq b + c$ (example: $\Leftrightarrow_{p,B}$).

Double implicational quantifiers defined in [14] by $TPC_{\Leftrightarrow}$: $a' \geq a \wedge b' \leq b \wedge c' \leq c$ (examples: all Σ-double implicational), Note that there are also double implicational quantifiers that are not Σ-double implicational but they seem not too much interesting.

Σ-*equivalence quantifiers* defined in [14] by $TPC_{\Sigma,\equiv}$: $a' + d' \geq a + d \wedge b' + c' \leq b + c$ (examples: $\equiv_{p,B}$).

Equivalence quantifiers defined by the condition $TPC_{\equiv}$: $a' \geq a \wedge b' \leq b \wedge c' \leq c \wedge d' \geq d$ (examples: $\sim_{\alpha,B}$, $\Rightarrow^{+}_{p,B}$ and moreover all the above defined classes are own subclasses of this class). The class of equivalence quantifiers is defined in [2] as the class of *association quantifiers*. We us the name *equivalence quantifiers* to avoid confusion with the association rules defined in [1]. It is proved moreover in [12] that the association rules defined in [1] are even not association in the sense of [2].

There are additional classes of 4ft-quantifiers, e.g., *pure double implicational*, *typical double implicational*; see [12]. The symmetrical quantifiers are defined in [2] by the condition $\approx (a,b,c,d) = \approx (a,c,b,d)$. It means that the symmetrical association rule $\varphi \approx \psi$ is true if and only if the association rule $\psi \approx \varphi$ is true. The 4ft-quantifier $\approx$ is symmetrical iff $\approx (a,b,c,d) = \approx (a,c,b,d)$. The 4ft-quantifiers $\Leftrightarrow_{p,B}$, $\equiv_{p,B}$, $\sim_{\alpha,B}$, and $\Rightarrow^{+}_{p,B}$ are symmetrical. The class of the 4ft quantifiers with *F-property* is introduced in Section 6.

5 Results on Classes of Association Rules

The first group of results concerns **deduction rules** of the form $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ where both $\varphi \approx \psi$ and $\varphi' \approx \psi'$ are association rules. This deduction rule is correct if it is true for each data matrix $\mathcal{M}$: If $\varphi \approx \psi$ is true in $\mathcal{M}$ then also $\varphi' \approx \psi'$ is true in $\mathcal{M}$. It is shown in [14] that there is an important subclass of *interesting implicational quantifiers* and a formula $\Gamma(\varphi,\psi,\varphi',\psi')$ of propositional calculus such that the deduction rule $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ with interesting implicational quantifier is correct if and only if the formula $\Gamma(\varphi,\psi,\varphi',\psi')$ is a tautology. There are similar results for Σ-double implicational and Σ-equivalence quantifiers.

The second group of results concerns **missing information**. An example of the data matrix $\mathcal{M}^X$ with missing information is in Fig. 2 (cf. Fig. 1 in section 2). The symbol X we interpret as ”the corresponding value is not known”.

object	A	B	C	$\dots$	$A(2)$	$B(5,6)$
o_1	2	7	16	$\dots$	1	0
o_2	X	X	4	$\dots$	0	X
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
o_n	1	6	5	$\dots$	0	1

Figure 2. Data matrix $\mathcal{M}^X$ with missing information

One of ways how to deal with missing information is done by the principle of the *secured X-extension* [2]. It means that the value $A(\alpha)(o, \mathcal{M}^X)$ of basic Boolean attribute $A(\alpha)$ for the object o in the data matrix $\mathcal{M}^X$ is $1(0)$ if it is sure $1(0)$ respectively in all possible completions of $\mathcal{M}^X$. The completion of $\mathcal{M}^X$ is each data matrix $\mathcal{M}$ that is a result of completion of each X by one of possible values of the corresponding attribute - the column of $\mathcal{M}^X$ (i.e., A, B, C, $\dots$ in Fig. 2). The values of the Boolean attributes derived from the basic Boolean attributes are defined by the truth tables of the connectives $\neg$, $\wedge$ and $\vee$ extended by the principle of secured X-extension [2]; see Fig. 3.

	$\neg$
1	0
X	X
0	1

$\wedge$	1	X	0
1	1	X	0
X	X	X	0
0	0	0	0

$\vee$	1	X	0
1	1	1	1
X	1	X	X
0	1	X	0

Figure 3. Extended truth tables of $\vee$, $\wedge$ and $\neg$

The principle of the *secured X-extension* is applied also to the association rule $\varphi \approx \psi$. The association rule is true in the data matrix $\mathcal{M}^{\mathcal{X}}$ with missing information if this association rule is true in all the possible completions $\mathcal{M}$ of $\mathcal{M}^X$. The core of the problem is that we have to evaluate the condition related to the 4ft-quantifier in the 4ft table $4ft(\varphi,\psi, \mathcal{M})$ for each completion $\mathcal{M}$ of $\mathcal{M}^X$.

This problem can be solved by the secured completion of nine-fold table $9ft(\varphi, \psi, \mathcal{M}^X)$ of φ and ψ in $\mathcal{M}^X$; see Table 2. Here $f_{1,1}$ is the number of rows o of $\mathcal{M}^X$ such that both $\varphi(o, \mathcal{M}^X) = 1$ and $\psi(o, \mathcal{M}^X) = 1$, $f_{1,X}$ is the number of rows o of $\mathcal{M}^X$ such that both $\varphi(o, \mathcal{M}^X) = 1$ and $\psi(o, \mathcal{M}^X) = X$, etc.

Table 2. Nine-fold table $9ft(\varphi, \psi, \mathcal{M}^X)$

$\mathcal{M}^X$	ψ	ψ_X	$\neg\psi$
φ	$f_{1,1}$	$f_{1,X}$	$f_{1,0}$
φ_X	$f_{X,1}$	$f_{X,X}$	$f_{X,0}$
$\neg\varphi$	$f_{0,1}$	$f_{0,X}$	$f_{0,0}$

It is important that for lot of 4ft quantifiers $\approx$ we can for each nine-fold table $9ft(\varphi, \psi, \mathcal{M}^X)$ compute the quadruple $\langle a_{\approx}, b_{\approx}, c_{\approx}, d_{\approx}\rangle$ such that $\varphi \approx \psi$ is true in all the possible completions $\mathcal{M}$ of $\mathcal{M}^X$ iff $\approx (a_{\approx}, b_{\approx}, c_{\approx}, d_{\approx}) = 1$ [2, 15]. The quadruple $\langle a_{\approx}, b_{\approx}, c_{\approx}, d_{\approx}\rangle$ is called the *secured completion of* $9ft(\varphi, \psi, \mathcal{M}^X)$ *for* $\approx$. The secured completion $\langle a_{\Rightarrow^*}, b_{\Rightarrow^*}, c_{\Rightarrow^*}, d_{\Rightarrow^*}\rangle$ of $9ft(\varphi, \psi, \mathcal{M}^X)$ for implicational quantifier $\Rightarrow^*$ is $\langle f_{1,1}, f_{0,1} + f_{1,X} + f_{X,0} + f_{X,X}, 0, 0\rangle$ [2]. There are similar results for Σ-double implicational and Σ-equivalence quantifiers [15] and for quantifiers with F-property; see section 7.

The third group of results concerns **tables of critical frequencies** which can help to avoid complex computation related to some 4ft-quantifiers defined using statistical hypothesis tests. Examples of such 4ft-quantifiers are $\Rightarrow^!_{p,\alpha,B}$ defined by $\sum_{i=a}^{a+b} \binom{a+b}{i} p^i (1-p)^{a+b-i} \leq \alpha \wedge a \geq B$ and the Fisher's quantifier $\sim_{\alpha,B}$ defined by the condition $\sum_{i=a}^{\min(r,k)} \frac{\binom{k}{i}\binom{n-k}{r-i}}{\binom{r}{n}} \leq \alpha \;\wedge\; ad > bc \;\wedge\; a \geq B$.

There is a non-negative non-decreasing function $Tb_{\Rightarrow^*}$ for each implicational quantifier $\Rightarrow^*$ such that for each integer $a \geq 0$ there is $Tb_{\Rightarrow^*}(a) \in \{0, 1, 2, \dots\} \cup \{\infty\}$ such that for $b \geq 0$ it is $\Rightarrow^* (a, b, c, d) = 1$ iff $b < Tb_{\Rightarrow^*}(a)$ [2]. The function $Tb_{\Rightarrow^*}(a)$ is called the *table of maximal* b for implicational quantifier $\Rightarrow^*$. Similar tables can be found for Σ-double implicational and Σ-equivalence quantifiers [2, 11] and for quantifiers with F-property; see Section 7.

The **fourth group of results** concerns **relation of association rules to classical predicate calculus**. Observational calculi are defined in [2] by modifications of classical predicate calculi; see also section 1. Thus there is a natural question if some association rules can be defined by formulas of classical predicate calculus (i.e calculus with only classical quantifiers $\forall$ and $\exists$). There is a reasonable criterion of definability of association rules in classical predicate calculus with equality that can be further simplified for implicational, Σ-double implicational and Σ-equivalence quantifiers [13].

6 F-property

Definition of the class of association rules with F-property was inspired by considerations on dealing with missing information for Fisher's quantifier. Especially the task of finding a secured completion $\langle a_F, b_F, c_F, d_F\rangle$ of nine-fold table $9ft(\varphi, \psi, \mathcal{M}^X)$ of φ and ψ in data matrix $\mathcal{M}^X$ with missing information for the Fisher's quantifier was solved in [9]. It is proved that it is $a_F = f_{1,1}$, $b_F = f_{1,0} + f_{1,X} + f_{X,0} + F_1$, $c_F = f_{0,1} + f_{0,X} + f_{X,0} + F_2$,

$d_F = f_{0,0}$ where $F_1 + F_2 = f_{X,X}$ such that $|b_F - c_F|$ is minimal. The proof is rather complicated.

There is a natural question if there are some additional 4ft-quantifiers that have the same secured completion. Two additional 4ft quantifiers are defined in [2]: The quantifier $\sim^S_{\delta,B}$ of *simple deviation* is defined for $\delta > 0$ and $B > 0$ by the condition $ad > e^\delta bc \,\wedge\, a \geq B$. The association rule $\varphi \sim^S_{\delta,B} \psi$ can be interpreted as *the logarithmic interaction of φ and ψ is estimated to be greater than δ.*

The χ^2 quantifier $\sim^\chi_{\alpha,B}$ is defined for $0 < \alpha < 0.5$ and $B > 0$ by $ad > bc \wedge \frac{n(ad>bc)}{(a+c)(b+d)(a+b)(a+c)} \geq \chi^2_\alpha \wedge a \geq B$. Here χ^2_α is $(1-\alpha)$ quantile of the χ^2 distribution function. The association rule $\varphi \sim^\chi_{\alpha,Base} \psi$ can be derived from the statistical χ^2 test (on the level of α) of the null hypothesis H0: φ and ψ are independent against the alternative H1: the logarithmic interaction of φ and ψ is positive. It is proved in [2] that both $\sim^S_{\delta,B}$ and $\chi^\chi_{\alpha,B}$ have the same secured completion $\langle a_F, b_F, c_F, d_F\rangle$ as the Fisher's quantifier.

Still the question remains if there is a whole class of 4ft-quantifiers that have the same secured completion $\langle a_F, b_F, c_F, d_F\rangle$. The class of 4ft-quantifiers with F-property is defined in [10] (see also [15]) such that the 4ft quantifier $\approx$ has the *F-property* if it satisfies:

1. If $\approx (a,b,c,d) = 1$ and $b \geq c-1 \geq 0$ then $\approx (a, b+1, c-1, d) = 1$.

2. If $\approx (a,b,c,d) = 1$ and $c \geq b-1 \geq 0$ then $\approx (a, b-1, c+1, d) = 1$.

It is also proved in [10] that each equivalence quantifier $\approx$ has the same secured completion $\langle a_F, b_F, c_F, d_F\rangle$ as the Fisher's quantifier if and only if $\approx$ has the F-property. There is also a direct proof in [10] that the Fisher's quantifier, the quantifier of simple deviation and the χ^2 quantifier have the F-property. It is an alternative proof that these quantifiers have the above defined secured completion $\langle a_F, b_F, c_F, d_F\rangle$.

7 Results on F-property

7.1 Deduction Rules and F-property

We are interested in the deduction rules of the form

$$\frac{\varphi \approx_F \psi}{\varphi' \approx_F \psi'}$$

where both $\varphi \approx_F \psi$ and $\varphi' \approx_F \psi'$ are association rules and $\approx_F$ is the 4ft-quantifier with the F-property. Remember that there are interesting results concerning correctness of analogous rules with various 4ft-quantifiers; see Section 5.

There is a criterion of correctness of the rule $\frac{\varphi \approx_F \psi}{\varphi' \approx_F \psi'}$ however it is not too much simple. It is moreover known only for association rules $\varphi \approx_F \psi$ such that the Boolean attributes φ and ψ are derived from the data matrix with the Boolean columns $P_1, P_2, \ldots, P_N$. It means that φ and ψ are derived from predicates $P_1, P_2, \ldots, P_N$ and not from basic Boolean attributes $A(\alpha), B(\beta), \ldots$ etc. The example of such the rule is $P_1 \wedge P_2 \approx_F P_3 \vee P_2$.

We are going to present this criterion in a bit informal way, the formal one is out of the scope of this paper. The criterion is based on propositional formulas associated with Boolean attributes φ and ψ. If P_i is one of Boolean attributes $P_1, P_2, \ldots, P_N$ then the propositional formula $\mathcal{P}(P_i)$ related to $\mathcal{P}(P_i)$ is defined as $\mathcal{P}(P_i) = p_i$ where p_i is a propositional variable. If τ and ω are Boolean attributes derived from $P_1, P_2, \ldots, P_N$ then the propositional formula $\mathcal{P}(\tau \wedge \omega)$ related to $\tau \wedge \omega$ is defined as $\mathcal{P}(\tau) \wedge \mathcal{P}(\omega)$ and analogously for Boolean connectives $\vee$ and $\neg$. Thus it is $\mathcal{P}(P_1 \wedge P_2) = p_1 \wedge p_2$, $\mathcal{P}(P_3 \vee \neg P_4) = p_3 \vee \neg p_4$ etc.

The presented criterion concerns a specific subclass of quantifiers with the F-property. We need notions of strong symmetrical quantifiers and the F^+ property. We say that the 4ft-quantifier $\approx$ is *strong symmetrical* iff

$$\approx (a,b,c,d) \;=\; \approx (a,c,b,d) = \;\approx (d,b,c,a)\,.$$

The 4ft-quantifier $\approx$ *has F^+ property* if it has the F property and if it satisfies the following conditions:

- It is $\approx (0,b,c,d) = 0$ for each 4ft-table $\langle 0,b,c,d\rangle$ and it is $\approx (a,b,c,0) = 0$ for each 4ft-table $\langle a,b,c,0\rangle$.

- There are 4ft-tables $\langle a,b,c,d\rangle$ and $\langle a,b',c',d\rangle$ such that $a+b+c+d = a+b'+c'+d$, $\approx (a,b,c,d) = 1$ and $\approx (a,b',c',d) = 0$.

The criterion of correctness is proved in [10] and it can be a bit informally formulated as follows. Let $\approx_F$ be the strong symmetrical equivalence quantifier with the F^+ property. Then the deduction rule $\frac{\varphi_1 \approx_F \psi_1}{\varphi_2 \approx_F \psi_2}$ where φ_1, φ_2, ψ_1, and ψ_2 are built from $P_1, P_2, \ldots, P_N$ is correct if and only if at least one of the following conditions a) – g) is satisfied. (The symbol $\rightarrow$ denotes the propositional connective of implication.)

a) The following formulas are tautologies
$\mathcal{P}(\varphi_1) \wedge \mathcal{P}(\psi_1) \rightarrow \mathcal{P}(\varphi_2) \wedge \mathcal{P}(\psi_2)$,
$\mathcal{P}(\varphi_2) \wedge \neg\mathcal{P}(\psi_2) \rightarrow \mathcal{P}(\varphi_1) \wedge \neg\mathcal{P}(\psi_1)$,
$\neg\mathcal{P}(\varphi_2) \wedge \mathcal{P}(\psi_2) \rightarrow \neg\mathcal{P}(\varphi_1) \wedge \mathcal{P}(\psi_1)$,
$\neg\mathcal{P}(\varphi_1) \wedge \neg\mathcal{P}(\psi_1) \rightarrow \neg\mathcal{P}(\varphi_2) \wedge \neg\mathcal{P}(\psi_2)$.

b) At least one of formulas $\mathcal{P}(\varphi_1) \rightarrow \neg\mathcal{P}(\psi_1)$
and $\neg\mathcal{P}(\varphi_1) \rightarrow\; \mathcal{P}(\psi_1)$ are tautologies.

c) The following formulas are tautologies
$\mathcal{P}(\varphi_1) \wedge \mathcal{P}(\psi_1) \rightarrow \mathcal{P}(\varphi_2) \wedge \mathcal{P}(\psi_2)$
$\mathcal{P}(\varphi_2) \wedge \neg\mathcal{P}(\psi_2) \rightarrow \neg\mathcal{P}(\varphi_1) \wedge \mathcal{P}(\psi_1)$
$\neg\mathcal{P}(\varphi_2) \wedge \mathcal{P}(\psi_2) \rightarrow \mathcal{P}(\varphi_1) \wedge \neg\mathcal{P}(\psi_1)$
$\neg\mathcal{P}(\varphi_1) \wedge \neg\mathcal{P}(\psi_1) \rightarrow \neg\mathcal{P}(\varphi_2) \wedge \neg\mathcal{P}(\psi_2)$.

d) The following formulas are tautologies
$\mathcal{P}(\varphi_1) \wedge \mathcal{P}(\psi_1) \rightarrow \neg\mathcal{P}(\varphi_2) \wedge \neg\mathcal{P}(\psi_2)$,
$\mathcal{P}(\varphi_2) \wedge \neg\mathcal{P}(\psi_2) \rightarrow \mathcal{P}(\varphi_1) \wedge \neg\mathcal{P}(\psi_1)$,
$\neg\mathcal{P}(\varphi_2) \wedge \mathcal{P}(\psi_2) \rightarrow \neg\mathcal{P}(\varphi_1) \wedge \mathcal{P}(\psi_1)$,
$\neg\mathcal{P}(\varphi_1) \wedge \neg\mathcal{P}(\psi_1) \rightarrow \mathcal{P}(\varphi_2) \wedge \mathcal{P}(\psi_2)$.

e) The following formulas are tautologies
$\mathcal{P}(\varphi_1) \wedge \mathcal{P}(\psi_1) \rightarrow \neg\mathcal{P}(\varphi_2) \wedge \neg\mathcal{P}(\psi_2)$,
$\mathcal{P}(\varphi_2) \wedge \neg\mathcal{P}(\psi_2) \rightarrow \neg\mathcal{P}(\varphi_1) \wedge \mathcal{P}(\psi_1)$,
$\neg\mathcal{P}(\varphi_2) \wedge \mathcal{P}(\psi_2) \rightarrow \mathcal{P}(\varphi_1) \wedge \neg\mathcal{P}(\psi_1)$,
$\neg\mathcal{P}(\varphi_1) \wedge \neg\mathcal{P}(\psi_1) \rightarrow \mathcal{P}(\varphi_2) \wedge \mathcal{P}(\psi_2)$.

f) Both the following two formulas are tautologies
$\mathcal{P}(\varphi_1) \wedge \mathcal{P}(\psi_1) \rightarrow \mathcal{P}(\varphi_2) \wedge \mathcal{P}(\psi_2)$,
$\neg\mathcal{P}(\varphi_1) \wedge \neg\mathcal{P}(\psi_1) \rightarrow \neg\mathcal{P}(\varphi_2) \wedge \neg\mathcal{P}(\psi_2)$,
and at least one of formulas $\mathcal{P}(\varphi_2) \rightarrow \mathcal{P}(\psi_2)$,
$\neg\mathcal{P}(\varphi_2) \rightarrow \neg\mathcal{P}(\psi_2)$ are tautologies.

g) Both the following two formulas are tautologies
$\mathcal{P}(\varphi_1) \wedge \mathcal{P}(\psi_1) \rightarrow \neg\mathcal{P}(\varphi_2) \wedge \neg\mathcal{P}(\psi_2)$,
$\neg\mathcal{P}(\varphi_1) \wedge \neg\mathcal{P}(\psi_1) \rightarrow \mathcal{P}(\varphi_2) \wedge \mathcal{P}(\psi_2)$,
and at least one of formulas $\mathcal{P}(\varphi_2) \rightarrow \mathcal{P}(\psi_2)$,
$\neg\mathcal{P}(\varphi_2) \rightarrow \neg\mathcal{P}(\psi_2)$ are tautologies.

We already know that the Fisher's quantifier, the quantifier of simple deviation and the χ^2 quantifier are equivalence quantifiers. It is proved in [10] that these quantifiers are also strong symmetrical and that they have F^+ property.

7.2 Tables of Critical Frequencies and F-property

It is proved in [10] that for the symmetrical 4ft-quantifier $\approx$ with the F-property there is a function $T_\approx$ that assigns to each triple $\langle a, d, n\rangle$ of natural numbers satisfying $a+d \leq n$ the number $T_\approx(a, d, n)$ such that for each $b \geq 0$ and $c \geq 0$ where $a+b+c+d = n$ it is

$$\approx (a, b, c, d) = 1 \text{ iff } |b - c| \geq T_\approx(a, d, n)\ .$$

The function $T_\approx(a, d, n)$ can be used in the same way as the function $Tb_{\Rightarrow^*}(a)$ called table of maximal b for implicational quantifier $\Rightarrow^*$, see Section 5. The function $T_\approx(a, d, n)$ is called *table of minimal* $|b - c|$.

7.3 Symmetry and F-property

There was a conjecture that having the same secured completion as the Fisher's quantifier is for equivalence quantifiers the same as being symmetrical. However it can be easy verified that the 4ft-quantifier $\approx^A$ defined such that $\approx^A (a, b, c, d) = 1$ if and only if
$b = c = 0 \vee b = 1 \wedge c = 0 \vee b = 0 \wedge c = 1 \vee b = 2 \wedge c = 0$ is equivalence, has the F-property (i.e. it has the same secured completion as the Fisher's quantifier) and it is not symmetrical.

Let us define the 4ft-quantifier $\approx^B$ [10] such that it is $\approx^B (a, b, c, d) = 1$ iff $b \leq 2 \wedge c \leq 1 \vee b \leq 1 \wedge c \leq 2$. It can be easy verified that $\approx^B$ is equivalence and symmetrical but it has not the F-property and thus it is equivalence and symmetrical but it has not the same secured completion as the Fisher's quantifier.

Thus the above mentioned conjecture is denied.

8 F-property and Σ-equivalence Rules

There are the following facts:

- The Fisher's quantifier, the quantifier of simple deviation and the χ^2 quantifier belong to the class of equivalence quantifiers; see [2].
- If $\approx$ is the Fisher's quantifier, the quantifier of simple deviation or the χ^2 quantifier then the rule $\varphi \approx \psi$ expresses a tendency of φ and ψ to have the same value (either *true* or *false*).
- The Σ-*equivalence quantifiers* are defined by $TPC_{\Sigma,\equiv}$ $a'+d' \geq a+d \wedge b'+c' \leq b+c$, thus they also express a tendency of φ and ψ to have the same values. An example is the quantifier $\equiv_{p,B}$ defined by the condition $\frac{a+d}{n} \geq p \wedge a \geq B$; see Section 3.

Thus it is natural to ask the following questions:

- Do the Fisher's quantifier, the quantifier of simple deviation and the χ^2 quantifier belong to the class of Σ-equivalence quantifiers?
- What is the relation of the class of quantifiers with the F-property to the class of Σ-equivalence quantifiers?

It is easy to prove that each Σ-equivalence quantifier has the F-property. Let us suppose that $\approx$ is the Σ-equivalence quantifier. We have to prove (see Section 6):

1. If $\approx (a, b, c, d) = 1$ and $b \geq c - 1 \geq 0$ then
$\approx (a, b+1, c-1, d) = 1$.
2. If $\approx (a, b, c, d) = 1$ and $c \geq b - 1 \geq 0$ then
$\approx (a, b-1, c+1, d) = 1$.

We prove the condition 1), the proof of the condition 2) is analogous. Let us suppose $\approx (a, b, c, d) = 1$ and $b \geq c - 1 \geq 0$, we have to show $\approx (a, b+1, c-1, d) = 1$. However, it follows from the truth preservation condition $TPC_{\Sigma,\equiv}$: $a'+d' \geq a+d \wedge b'+c' \leq b+c$ for Σ-equivalence quantifiers. The $TPC_{\Sigma,\equiv}$ says that if $\approx (a, b, c, d) = 1$ and $a'+d' \geq a+d \wedge b'+c' \leq b+c$ then $\approx (a', b', c', d') = 1$. In our case we have $a' = a, b' = b+1, c' = c-1, d' = d$

and thus $a' + d' = a + d \wedge b' + c' = b + c$ and according to $TPC_{\Sigma,\equiv}$ it is $\approx (a, b+1, c-1, d) = 1$. This finishes the proof.

The question remains if there are some 4ft-quantifiers with the F-property that are not Σ-equivalence. It is proved in [12] that the Fisher's quantifier, the quantifier of simple deviation and the χ^2 quantifier are examples of important quantifiers with the F-property that do not belong to the class of Σ-equivalence quantifiers. The proof is out of the scope of this paper. It is based on the notion of the $AD01$-property. We say that the 4ft quantifier $\approx$ has the $AD01$-*property* if there are $\langle a, b, c, d\rangle$ and $\langle a, b', c', d\rangle$ such that $b + c = b' + c' \wedge \approx (a, b, c, d) = 1 \wedge \approx (a, b', c', d) = 0$. There is a theorem saying that if the quantifier $\approx$ has the $AD01$-property then it is not Σ-equivalence. It is also proved in [12] that Fisher's quantifier, the quantifier of simple deviation and the χ^2 quantifier have the $AD01$-property. Thus these quantifiers are not Σ-equivalence.

9 Conclusions and Further Work

Association rules can be defined as general relations of two Boolean attributes derived in various ways from columns of analyzed data matrix. Several classes of association rules related to types of these relations is introduced. Overview of the theoretically interesting and practically important results related to these classes is given.

A new class of association rules with F-property is described in a detailed way. This class contains important association rules corresponding to the Fisher's test and to the χ^2 test. Relation of the class of rules with the F-property to the class of Σ-equivalence rules is studied.

There are tens of additional measures for association patterns described in literature; see e.g. [7]. These measures can be understood as the additional 4ft-quantifiers. We suppose to study relations of these additional 4ft-quantifiers to known classes of association rules.

References

[1] Aggraval R et al. (1996) Fast Discovery of Association Rules. In: Fayyad, U. M. et al.(Eds.) Advances in Knowledge Discovery and Data Mining. AAAI Press

[2] Hájek P, Havránek T (1978) Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory. Springer, Berlin Heidelberg New York

[3] Hájek P (guest ed.) (1978) International Journal of Man-Machine Studies, special issue on GUHA, 10

[4] Hájek P (guest ed.) (1981) International Journal of Man-Machine Studies, second special issue on GUHA, 15

[5] Hájek P, Havránek T, Chytil M (1983) GUHA Method. Academia, Prague (in Czech)

[6] Hájek P, Sochorová A, Zvárová J (1995) GUHA for personal computers. Computational Statistics & Data Analysis, 19

[7] Pang-Ning T, Kumar V, Srivastava J (2004) Selecting the Right Objective Measure for Association Analysis. Information Systems, 29

[8] Ralbovský M, Kuchař T (2007) Using Disjunctions in Association Mining. In: Perner P (Ed.): Advances in Data Mining - Theoretical Aspects and Applications. Springer, Berlin Heidelberg New York

[9] Rauch J (1975) Ein Beitrag zu der GUHA Methode in der dreiwertigen Logik. Kybernetika, 11

[10] Rauch J (1986) Logical Foundations of Hypothesis Formation from Databases. Mathematical Institute of the Czechoslovak Academy of Sciences, Prague, Dissertation (in Czech)

[11] Rauch J (1998) Classes of Four-Fold Table Quantifiers. In: Zytkow J, Quafafou M (Eds.): Principles of Data Mining and Knowledge Discovery. Springer, Berlin Heidelberg New York

[12] Rauch J (1998) Contribution to Logical Foundations of KDD. University of Economics Prague, Assoc. Prof. Thesis (in Czech)

[13] Rauch J (2004) Definability of Association Rules and Tables of Critical Frequencies. In: Lin T Y et al. (Eds.): Foundations of Data Mining. Brighton, IEEE Computer Society

[14] Rauch J (2005) Logic of Association Rules. Applied Intelligence, 22

[15] Rauch J (2005) Classes of Association Rules, an Overview. In: Lin T Y, Xie Y (Eds.). Foundation of semantic Oriented Data and Web Mining. Houston, IEEE Computer Society

[16] 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 Heidelberg New York

[17] Rauch J, Šimůnek M (2005) GUHA Method and Granular Computing. In: Hu, X et al. (Eds.) Proceedings of IEEE conference Granular Computing. Beijing, IEEE Computer Society

[18] Rauch J, Šimůnek M: Semantic Web Presentation of Analytical Reports from Data Mining – Preliminary Considerations. In Proceedings of Web Intelligence 07