Characteristics of cosymmetric association rules

Michal Burda, Marian Mindek, and Jana Šarmanová

Department of Computer Science, VŠB – Technical University of Ostrava
17. listopadu 15, 708 33 Ostrava–Poruba, Czech Republic
{michal.burda, marian.mindek, jana.sarmanova}@vsb.cz

Abstract. Association rules are essential data mining tool and as such has been well researched. Many new types of association rules based on both categorial or quantitative data have been founded ([8], [7], [2], [4]). Our work is directed to the theoretical features of association rules; especially, we study a specific class of association rules called *δ-cosymmetric rules*. We present here some interesting properties of such rules and provide a definition of rules expressing the significant difference in position, as an example. We show here that even the usual implicational rules are special cases of δ-cosymmetric rules.

Key words: Cosymmetric rules, association rules, typed relations, data mining

1 Preface

This paper is intended to motivate the rise of a new class of association rules called *δ-cosymmetric rules*. First of all, we describe here briefly the notions of the *Logic of typed relations* used to write the association rules down (for more information see [6]). After that, we provide some motivating examples of the representative cosymmetric rule types. We also study several features of the δ-cosymmetric rules and define the δ-cosymmetric rule of significant difference in position. The end of this paper is dedicated to some notes on how to mine the δ-cosymmetric rules.

2 Logic of typed relations

In [6], we have developed the *Probabilistic Logic of Typed Relations* (PLTR) suitable for the formal association rules representation. In this section we briefly and informally describe main notions of that logic to understand the meaning of its formulae.

The main notion of PLTR is *typed relation*. Typed relation can be simply viewed as a data table with finite number of columns and rows. Each column represents one *attribute* and a set of such attributes is a *type* of the relation.

A typed relation is similar to classical concept of mathematical relation. We can perform usual set operations as union (∪), intersection (∩) or difference

K. Richta, V. Snášel, J. Pokorný (Eds.): Dateso 2005, pp. 20–31, ISBN 80-01-03204-3.

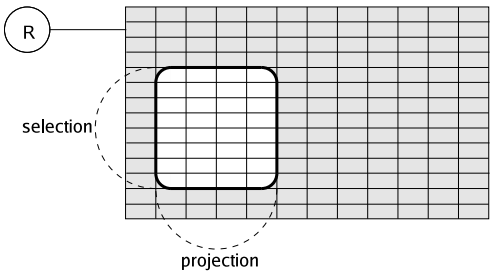

Fig. 1. Selection and projection on the relation R.

$(-)$. Furthermore, there exist two crucial relational operations: *selection* and *projection*. Selection is an unary operation of the form

$$R(c_1 \wedge c_2 \wedge \neg c_3)$$

where R is typed relation and $c_1 \wedge c_2 \wedge \neg c_3$ is a formula called *selection condition*. Selection is used to *select* only the rows satisfying the given condition. For example, when R is a data table (typed relation) of university students, the selection

$$R(\mathsf{age} > 25)$$

picks only the students older than 25. The projection on relation R is an unary operation of the form

$$R[A_1, A_2, \ldots, A_n].$$

The projection is used to take out only several columns (attributes) of the relation R. The choosed attributes are simply written in the comma-separated list in the brackets. The projection

$$R[\mathsf{name}, \mathsf{date\_of\_birth}]$$

results simply in the two-column data table with student's basic personal information. Obviously, we can combine selection and projection together to pick up an arbitrary sub-relation of the original typed relation, e.g.

$$R(\mathsf{age} > 25)[\mathsf{name}, \mathsf{date\_of\_birth}],$$

which results in a relation of basic personal information of students older than 25. (See also figure 1.) The rules written in PLTR use the relational operations described above to explicitly express a knowledge. For example,

$$R(\mathsf{age} > 65)[\mathsf{blood\_pressure}] >^{\star}_{mean} R(\mathsf{age} < 21)[\mathsf{blood\_pressure}]$$

tells that the blood pressure of people older than 65 is *in average* significantly higher than for people younger than 21. In the above rule we use a mapping $>^{\star}_{mean}$ to express the strong difference in the mean value between two "data columns". (See also figure 2.) The mapping $>^{\star}_{mean}$ is simply a function, which computes a truth value of the strong difference in mean from the given two typed relations.

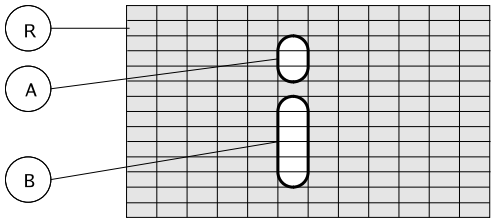

Fig. 2. Comparison of the two disjunctive sub-tables.

3 The δ-cosymmetric rules

There exist a wide variety of the association rule types. The best-known are the rules in the implicational form, which say that when the object satisfies some condition (called *antecedent*), it (very probably) gratifies some other condition (*succedent*), e.g.:

$$\mathsf{tequila} \wedge \mathsf{salt} \Rightarrow \mathsf{lemon}. \tag{1}$$

This rule simply says that customers who buy tequila and salt often buy lemons, too. However, there are many other rule types (e.g. associational, correlational etc. – see [1], [2], [5], [7], [8], [9]). It is not our goal to mention each of them. We preferable move the focus to the rules, which we later name δ-*cosymmetric*. Consider the subsequent rule from [4]:

$$\mathsf{sex} = \text{“female”} \;\Rightarrow\; \mathsf{wage}\text{: mean} = \$7.90/\text{hr (overall mean } \mathsf{wage} = \$9.02). \tag{2}$$

It indicates that the women's wage mean is significantly different to the rest of examined objects. That is, the rule says that women earn in average less than men. (The overall wage is in the rule for information only. To be statistically consistent, we must compare two disjoint sets of values, e.g. female againts male – see [4].) In general, the statistical test in the background of the rule *compares two sets of quantitative data* – women's wage against the wage of the remaining data table (in fact, against men's wage). We can apply the same mechanism and mine similar rules, e.g.:

$$\mathsf{non\text{-}smoker} \wedge \mathsf{wine\text{-}drinker} \;\Rightarrow\; \mathsf{life\text{-}expectancy} = 85 \quad (\text{overall} = 80). \tag{3}$$

Such rule says that people who drink wine and do not smoke live in average longer than the other people. One can see, we *compare* the life expectancy of people who don't smoke and drink wine against the rest of the data table. Such property is more visible when re-writing the original rules (2) and (3) (see also [4]) into PLTR:

$$R(\mathsf{sex} = \text{“female”})[\mathsf{wage}] <^{\star}_{mean} R(\mathsf{sex} \neq \text{“female”})[\mathsf{wage}] \tag{4}$$

and

$$R(\mathsf{non\text{-}smoker} \wedge \mathsf{wine\text{-}drinker})[\mathsf{life\text{-}expectancy}] >^{\star}_{mean} R(\neg(\mathsf{non\text{-}smoker} \wedge \mathsf{wine\text{-}drinker}))[\mathsf{life\text{-}expectancy}]. \tag{5}$$

Our research shows that many types of the associational rules can be transformed to the fashion of *comparing* "something" against "something else" (later in this paper, we mention some of them). Thus, it is natural to expect that such rules will have some equal properties and that it will behave similarly in alike situations. Therefore it is reasonable to identify the common features and use them in general definition of a new class of association rules. Later in this paper we try to do so and name the class of such association rules the *δ-cosymmetric rules*.

Moreover, it is obvious to contemplate rules of type (4) or (5) as formulae of PLTR. That is, one can treat the symbol $<^{\star}_{mean}$ as a predicate, whose truth value is the probability (quantity in interval $[0, 1]$). Such approach corresponds to the fact that the statistical test gets never the absolute truth – there is always a chance (non-zero probability) of a false result. In [6], we have developed a logic, whose truth values are probability intervals $i = \langle l, h\rangle$ where $0 \leq l \leq h \leq 1$.

3.1 Domain

The following subsections try to highlight some properties that are common in the class of association rules we want to name *δ-cosymmetric*. After that, we provide the first prototype definition of what δ-cosymmetric rule is.

We start with the domain of the δ-cosymmetric predicate $<^{\star}$. We can see, the rules of type (4) or (5) compare two typed relations. It is natural to expect that when $\langle A, B\rangle$ is comparable then $\langle B, A\rangle$ is comparable too.

Let $\mathcal{R}$ is a set of all typed relations. We may expect that each δ-cosymmetric predicate's domain D equals to the carthesian product of some set of typed relations:

$$\exists K \subseteq \mathcal{R} : D = K \times K.$$

This property tells us that for each typed relations $A, B \in K$, $\langle A, A\rangle$, $\langle A, B\rangle$ and $\langle B, A\rangle$ are comparable by the δ-cosymmetric predicate. That is, one can ask the truth value of the formulae $<^{\star}(A, A)$, $<^{\star}(A, B)$, $<^{\star}(B, A)$ for each $A, B \in K$.

3.2 Minimum difference

When mining the rules of type (4), it is useful to introduce an user-definable *minimum difference parameter* δ. (See also [4].) Its purpose is as follows: Finding conditions for which the means of some attribute are merely different does not lead to interesting information. If we were to discover, for example, a group of people with life expectancy five days more than the rest population, it may not be of interest to us even if it passes a statistical test.

The same concept can be used when comparing variances, probability or anything else. – The next thing common to each cosymmetric rule is the possibility to employ the minimum difference δ to it.

In the following, we will write the rule of the minimum difference δ the subsequent way:

$$R(C_1)[A] >^{\star}_{\delta} R(C_2)[A] \tag{6}$$

or prefixually:

$$>^{\star}_{\delta}\big(R(C_1)[A],\ R(C_2)[A]\big). \tag{7}$$

E.g. see the rule of the difference in wage of at least \$5:

$$R(\mathsf{sex} = \text{“female”})[\mathsf{wage}] <^{\star}_{mean;\$5} R(\mathsf{sex} \neq \text{“female”})[\mathsf{wage}]. \tag{8}$$

3.3 Non-symmetricity

In the following, we will need to define *the negation* of formula F. Suppose F is formula of PLTR (e.g. (4)) whose truth value is $i = \langle l, h\rangle$. (It stands for the fact that F is true with probability $p \in [l, h]$.) We define a truth value of formula's F negation (denoted $\neg F$) as $i' = \langle 1 - h,\ 1 - l\rangle$.

The third common feature of rules similar to (4) is its non-symmetricity. Suppose we are convinced of the validity of the rule $>^{\star}(A, B)$. What can we say about the truth value of the rule $>^{\star}(B, A)$? It is clear, if values of relation A are significantly higher than values of relation B, the contrary statement can't be true as well (so the formula (10) holds).

More generally, the truth value of a statement "objects of relation B are minimally over δ less than objects of relation A" equals to a negation of the statement "objects of relation A are minimally over $(-\delta)$ less than objects of the relation B". Formally written:

$$<^{\star}_{\delta}(B, A) \ \Leftrightarrow\ \neg\big(<^{\star}_{-\delta}(A, B)\big). \tag{9}$$

When $\delta = 0$ is omitted, it leads to

$$<^{\star}(B, A) \ \Leftrightarrow\ \neg\big(<^{\star}(A, B)\big). \tag{10}$$

3.4 Monotony

Let $>^{\star}_{\delta_1}\big(A, B\big) = \langle l_1, h_1\rangle$ and $>^{\star}_{\delta_2}\big(A, B\big) = \langle l_2, h_2\rangle$ where $>^{\star}$ is a predicate similar to the previously discussed. One can observe that the following holds all the time:

$$\big(\delta_1 < \delta_2\big) \Rightarrow \big((l_1 \geq l_2) \wedge (h_1 \geq h_2)\big). \tag{11}$$

Informally, this property says that the increase of the minimum difference δ leads to the reduction of the rule's probability.

3.5 Quasi-transitivity

We name *probable* the rule, which truth value $i = \langle l, h\rangle$ satisfies the condition $0,5 < l$. Let $<^{\star}_{\delta}(A, B) = \langle l_1, h_1\rangle$, $<^{\star}_{\delta}(B, C) = \langle l_2, h_2\rangle$ and $<^{\star}_{\delta}(A, C) = \langle l_3, h_3\rangle$. The last property of rules similar to (4) named *quasi-transitivity* tells the following:

$$\big((0,5 < l_1) \wedge (0,5 < l_2)\big) \Rightarrow \big(0,5 \leq l_3\big). \tag{12}$$

Informally, when some sub-table A is *probably* lower than B and B is *probably* lower than C, it implies that A is probably not higher than C.

Please note, we can't say that the probability of $A <^{\star} C$ is higher or equals to the maximum or minimum of the probabilities of $A <^{\star} B$ and $B <^{\star} C$, because such condition holds in fact very seldom.

3.6 The definition of δ-cosymmetric rules

Actually, we are still working on the precise definition of the δ-cosymmetric relationship predicate. We try to unhide the important properties of the rules similar to (4). The subsequent definition should be considered as the first prototype of our effort. As our knowledge about the rules increases, we will modify the definition to better pick up the reality.

Definition 1. *Let $\mathcal{R}$ be the set of all typed relations, V the set of all truth values, $K \subseteq \mathcal{R}$ and $D = K \times K$. We name $<^\star$* the cosymmetric predicate schema *if $<^\star$ is a set of relationship predicates $<^\star_\delta$: $D \to V$ (defined for each $\delta \in \mathbb{R}$) and if the following holds:*

1. *For each typed relations $A, B \in K$ and $\delta \in \mathbb{R}$ holds:*

$$<^\star_\delta (A, B) = \neg\big(<^\star_{-\delta} (B, A)\big),$$

2. *For each typed relations $A, B \in K$ and $\delta_1, \delta_2 \in \mathbb{R}$ and $i_1 = \langle l_1, h_1\rangle$, $i_2 = \langle l_2, h_2\rangle$ such that $<^\star_{\delta_1} (A, B) = i_1$, $<^\star_{\delta_2} (A, B) = i_2$ holds:*

$$\big(\delta_1 < \delta_2\big) \Rightarrow (l_1 \geq l_2) \wedge (h_1 \geq h_2),$$

3. *For each typed relations $A, B, C \in K$ and $\delta \in \mathbb{R}$ and $i_1 = \langle l_1, h_1\rangle$, $i_2 = \langle l_2, h_2\rangle$, $i_3 = \langle l_3, h_3\rangle$, such that $<^\star_\delta (A, B) = i_1$, $<^\star_\delta (B, C) = i_2$, $<^\star_\delta (A, C) = i_3$, holds:*

$$\big((0,5 < l_1) \wedge (0,5 < l_2)\big) \Rightarrow \big(0,5 \leq l_3\big).$$

The elements $<^\star_\delta$ of the set $<^\star$ are called δ-cosymmetric relationship predicates. *The set D is also called* the domain of the cosymmetric predicate schema.

4 Concrete δ-cosymmetric predicates

In the above section we have discussed several properties of a so-called cosymmetric rules. In this section, we provide an exemplary definitions of such rule type.

4.1 Cosymmetric rules of significant difference in position

The idea for cosymmetric rules of significant difference in position is subsequent. One may have data which are quantitative and may ask, for which subsets of data the focused quantitative attribute is rather higher or lower in contrast to the rest (c.f. rule (4) or (5)). In the other words, one may enquire for all hypotheses about the differences in position that are supported within data. We can determine the difference and measure the significancy with appropriate statistical test of hypotheses.

For such purpose we use the Aspin–Welch statistical test (see [3]), which is two-sample test on means. The test is similar to the common Student's t test. It

assumes the two random samples X and Y to be normally distributed (there is no need of equal variances) and it tests the zero hypothesis $H_0 : \mathsf{E}X - \mathsf{E}Y = \delta$ against the two-sided alternative hypothesis $H_A : \mathsf{E}X - \mathsf{E}Y \neq \delta$. The test statistic is

$$T = \frac{\bar{X} - \bar{Y} - \delta}{S}, \quad \text{where} \quad S = \sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}; \qquad \left(f = \frac{S^4}{\frac{S_X^4}{m^2(m-1)} + \frac{S_Y^4}{n^2(n-1)}} \right).$$

The hypothesis H_0 is rejected if $|T| \geq t_f(1 - \frac{\alpha}{2})$, where t_f is a distribution function of Student's distribution with f degrees of freedom.

Pursuant to the one-sided Aspin–Welch statistics, we can define the relationship predicate $<^\star_{AW;\delta}$ as follows.

Definition 2. *Predicate $<^\star_{AW;\delta}$ is a function where an interval of probability $i = \langle p, p \rangle$ is mapped the following way to each pair of typed relations $\langle X, Y \rangle$, which both are non-empty and both contain just one column.*

$$<^\star_{AW;\delta} (X, Y) = \langle p, p \rangle$$

for such p where $T = t_f(p)$ for T, f and t_f as above.

The usage example comes after. Suppose we have a data table D about patients suffering certain disease. Let such table contains categorial column sex and quantitative column pressure. One may be interested whether $D(\mathsf{sex} = \text{“male”})[\mathsf{pressure}]$ gives higher values than $D(\mathsf{sex} = \text{“female”})[\mathsf{pressure}]$. That is, one enquires the validity of the following rule:

$$D(\mathsf{sex} = \text{“male”})[\mathsf{pressure}] >_{AW;0} D(\mathsf{sex} = \text{“female”})[\mathsf{pressure}].$$

Now we can take a closer look at the Aspin–Welch predicate $<^\star_{AW;\delta}$ to see, whether it has all the properties enumerated in section 3.6.

Theorem 1. *The set of all Aspin–Welch relationship predicates $<^\star_{AW;\delta}$ $(\forall \delta)$ is cosymmetric predicate schema.*

Proof. (a) Non-symmetricity. We should check the equivalence (9). Suppose typed relations X, Y and value δ. Let $>^\star_{AW;\delta} (X, Y) = \langle p_1, p_1 \rangle$ and $>^\star_{AW;-\delta} (Y, X) = \langle p_2, p_2 \rangle$. We are going to show that $p_1 = 1 - p_2$. Computing the values of p_1 and p_2 means accordingly to the definition 2 computing the T characteristics. Thus,

$$T_1 = \frac{\bar{X} - \bar{Y} - \delta}{S} = t_f(p_1) \qquad \text{and} \qquad T_2 = \frac{\bar{Y} - \bar{X} - (-\delta)}{S} = t_f(p_2).$$

We see that $T_1 = -T_2$, so $t_f(p_1) = -t_f(p_2)$. It is commonly known that $t_f(p) = -t_f(1 - p)$, so $p_1 = 1 - p_2$.

(b) Monotony. It is commonly known that $t_f(p)$ is monotone, so when we increase δ, the value of characteristics T gets lower and so does the value of the resultant probability p.

(c) Quasi-transitivity. the validity of quasi-transitivity condition is evident from the fact that $\forall f \in \mathbb{N} : t_f(0, 5) = 0$.

4.2 Funded cosymmetric rules

We can go on and define various other δ-cosymmetric predicates similar to the definition of Aspin–Welch predicate. We don't have enough space for such definitions, so let us leastwise mention some possibilities.

We can define many other predicates for determining the significant difference in position. Such definitions could be based on various existing statistical tests – it is possible e.g. to employ the *rank tests* to achieve robust cosymmetric predicates etc. Similarly to the significant difference in position, we can define predicates deciding of the difference in variance (dispersion). For example, we can mine rules telling us whether the presence of some attribute puts there significant increase of dispersion of some other attribute etc.

We can employ the two-sample tests on binomial distribution to generate rules about discrete attributes. Generally said, almost every two-sample statistical test may be considered to be used in a definition of appropriate cosymmetric predicate.

Let's have a look on the *implicational rules* of type (1). We show that we can define δ-cosymmetric rules that are analogous to them. Before doing so, we should describe shortly the meaning of the implicational rules.

The GUHA method ([8], [7]) works with the so-called *generalized quantifiers*. These quantifiers form the base for the association rule creation. The rules are of the form $\varphi \sim \psi$, where φ and ψ are formulae and $\sim$ the generalized quantifier. The truth of the rule is determined from a *4-field table* (see table 1), which summarizes the amount of objects satisfying ceratin configurations.

Table 1. 4-field table of φ and ψ

	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

The a value denotes the number of objects satisfying both φ and ψ, b is the number of objects satisfying φ and not ψ etc.

The quantifier $\Rightarrow_{p,base}$ called also *the funded implication* is defined for $0 < p \leq 1$ and $base \geq 0$ as follows. The rule $\varphi \Rightarrow_{p,base} \psi$ is true if and only if (*iff*)

$$\frac{a}{a+b} \geq p \wedge a \geq Base.$$

More on such rules can be read from [8], [7] or [10]. The example of the rule based on the funded implication is (1).

Now, we provide a definition of a predicate that is similar to the quantifier of funded implication. After that, we show that it is δ-cosymmetric.

Definition 3. *Let A and B be the typed relations, each containing exactly one column with values from the set $\{0,1\}$ and let $\delta \in [-1,1]$. Let us denote $sum(A)$ the number of A's rows possessing "1". We define the* Funded relationship predicate $<^{\star}_{fnd;\delta}$ *as follows:*

$$>^{\star}_{fnd;\delta}(A,B) = \langle 1,1\rangle \qquad \textit{iff} \qquad \frac{sum(A)}{sum(A)+sum(B)} > \frac{1+\delta}{2},$$

$$>^{\star}_{fnd;\delta}(A,B) = \langle 0,5,0,5\rangle \qquad \textit{iff} \qquad \frac{sum(A)}{sum(A)+sum(B)} = \frac{1+\delta}{2},$$

$$>^{\star}_{fnd;\delta}(A,B) = \langle 0,0\rangle \qquad \textit{iff} \qquad \frac{sum(A)}{sum(A)+sum(B)} < \frac{1+\delta}{2}.$$

Theorem 2. *The set of all funded relationship predicates $<^{\star}_{fnd;\delta}$ $(\forall\delta)$ is cosymmetric predicate schema.*

Proof. (a) Non-symmetricity. We must prove that $\frac{a}{a+b} > \frac{1+\delta}{2}$ iff $\frac{b}{a+b} < \frac{1-\delta}{2}$.

$$\frac{a}{a+b} > \frac{1+\delta}{2} \quad\Leftrightarrow\quad \frac{2a}{a+b} - 1 > \delta \quad\Leftrightarrow\quad \frac{2a+2b-2b}{a+b} - 1 > \delta \quad\Leftrightarrow$$
$$\Leftrightarrow\quad 1 - \frac{2b}{a+b} > \delta \quad\Leftrightarrow\quad \frac{b}{a+b} < \frac{1-\delta}{2}.$$

(b) Monotony and *(c) Quasi-transitivity* are obvious.

If we omit the minimum support constraint in the definition of the funded implication, we get the same rules as with the funded δ-cosymmetric predicate. In the other words, the rule

$$\varphi \Rightarrow_{p,0} \psi$$

is true on data table R iff the following rule has truth value equal to $\langle 1,1\rangle$:

$$R(\psi)[\varphi] >^{\star}_{fnd;(2p-1)} R(\neg\psi)[\varphi].$$

As a result we can say that implicational GUHA rules are just special cases of δ-cosymmetric rules. This surprising result convinced us of the importance of the δ-cosymmetric rules research.

5 Schemes of δ-cosymmetric association rules

Consider the general pattern of a δ-cosymmetric rule:

$$R(C_1)[A] >^{\star} R(C_2)[A]. \tag{13}$$

When mining such rules, we can generate and test virtually every combination of C_1, C_2, A, but doing so makes not much sense. It is because the association rule mining process results often in a wide range of association rules and it is sometimes hard to be acquainted with it. Moreover, only several combinations of conditions C_1 and C_2 are easy to interpret. Consider the following rule – although it may be true, the analyst has probably no usage for it.

$$R(\mathsf{eyes} = \text{“blue”} \wedge \mathsf{sex} = \text{“male”})[\mathsf{fat}] >^{\star} R(\mathsf{age} > 30 \wedge \mathsf{wage} < \$200)[\mathsf{fat}] \tag{14}$$

In the following, we try to recognize the patterns of δ-cosymmetric rules of better interest than general pattern (13).

5.1 Scheme "one-against-the-rest"

The easiest pattern of interesting δ-cosymmetric rules is

$$R(C)[A] >^{\star} R(\neg C)[A]. \tag{15}$$

We take one condition C and compare values of some quantitative attribute A for two sub-tables where the first satisfies the given condition C and the second doesn't. Such rules express the condition at which the values of attribute A are "somehow" significantly higher (or lower) than the rest of the data table. This basic pattern we name *one-against-the-rest*.

A pattern similar to (15) is *conditional one-against-the-rest*:

$$R(C_1 \wedge C_2)[A] >^{\star} R(\neg C_1 \wedge C_2)[A]. \tag{16}$$

This pattern stands for "when considering only values fulfilling the condition C_2, the additional condition C_1 indicates the significant increase of value A (in the sense of $>^{\star}$)." That is, we fisrt restrict ourselves on data rows satisfying C_2 only and then we search simply the one-against-the-rest rules on them.

5.2 Scheme "one-on-one"

The pattern *one-on-one* is a little more tricky. It is good in situations, when we want to compare groups created accordingly to one categorial attribute. Suppose attribute B be categorial with domain $\{b_1, b_2, \ldots, b_n\}$. Let moreover attribute A be quantitative. The pattern one-on-one is as follows:

$$R(B = b_i)[A] >^{\star} R(B = b_j)[A] \qquad (\text{for } i \neq j). \tag{17}$$

The rule of such type means: "The objects with value b_i in attribute B involve significantly higher values of attribute A than objects with value b_j in attribute B." Generally, we can generate and test $\binom{n}{2}$ different hypotheses for a categorial attribute with n various values.

We can add an additional condition C to form *conditional one-on-one* pattern, too:

$$R(B = b_i \wedge C)[A] >^{\star} R(B = b_j \wedge C)[A] \qquad (\text{for } i \neq j). \tag{18}$$

6 Some notes of how to reduce the number of resultant rules

The large size of the association rule mining results is the common problem. Analyst hardly orientates himself or herself in a big list of mined rules. Therefore, we enumerate here some hints of how to prune the result from less-interesting rules and so to restrict the resultant δ-cosymmetric formulae to the reasonable amount.

1. *The significance level* – the basic restriction on eventual rules is stating the minimum probability of its validity – in the other words, one may set a number p_{min} and throw away every rule, which truth value is below that threshold. Significance level can be pre-set to any of the usual values as 0,95 or 0,99.
2. *Contradictory conditions* – the conditions appearing in the rule should be contradictory. That is, when considering the rule

$$R(C_1)[A] >^{\star} R(C_2)[A]$$

   then the formula $C_1 \wedge C_2$ should be contradiction. Rules satisfying that criterion are more easily interpretable and we avoid the uncorrect statistical comparing of non-disjunctive samples. (Compare with rule (14). Note also that the rules based on one-against-the-rest or one-on-one are all of contradictory conditions.)
3. *Minimum support* – minimum support is the best-known instrument for pruning away the non-interesting conditions from which the association rules are going to be formed. The minimum support criterion simply says that there must exist minimally *minsup* objects satisfying condition that appears in the rule. If not, such condition isn't used in the association rule generating process. The definition of the *minsup* value greatly improves the efficience of association rule mining algorithms (see [1], [9], [11] for more information). Minimum support should be set by expert only.
4. *Minimum difference* – setting the minimum difference δ is analogous to the stating of minimum rule probability. Doing so we express that we are interested in the rules, which confirm the dissimilarity to be at least of size δ. Minimum difference should be set by expert only.
5. *Easy-to-interpret rules only* – in section 5 we have shown that generating all possible rules makes no sense. One may to generate only the rules, which are easy to interpret. That is, we should generate rules conforming to the patterns discussed in section 5. A similar criterion on that topic is to use conditions in conjunctive form only.

7 Conclusion and future work

In this paper, we have introduced the new class of association rules – the δ-cosymmetric rules. We are the first who has shown, how to use the Probability logic of typed relations (PLTR, see [6]) to express rules of such type. This paper also shows the benefit of using PLTR as a language for writing the association rules in, too.

We have identified the basic properties of δ-cosymmetric rules and provided the definition of rules of significant difference in position, as an example. The second part of this paper was dedidacted to some notes on how to generate the δ-cosymmetric rules to obtain the interesting rules only.

This paper also presents two basic examples of concrete δ-cosymmetric rules: the Aspin–Welch predicate and the Funded predicate. The second is surprise for

us, since it shows that GUHA's implicational rules are just the special cases of more general δ-cosymmetric rules.

Our future work will address the deeper research of δ-cosymmetric rules. We will try to unhide more interesting features of that rule class. For example, our actual research shows that the cosymmetric rules can be used in the definition of a function that is *metric*. An interesting task will be undisputably the clustering using such metrics, etc.

We are also focused on finding the fast and efficient algorithm to mine the δ-cosymmetric rules. A lot of work was done in [4] by Aumann and Lindell. (However, they didn't know that they are mining cosymmetric rules – their algorithm should be slightly modified to comply the wide range of possible rule types.)

We are also interested in the methods of visualisation of δ-cosymmetric rules. The properties of δ-cosymmetric rules make rational to use the slightly modified Hasse's diagrams to visualize the rules mined according to the pattern "one-on-one" discussed above. We also work on employing the conceptual lattices to represent mined δ-cosymmetric rules.

References

1. Agrawal, R. Fast discovery of association rules. In *Advances in knowledge discovery and data mining* (1996), AAAI Press / MIT Press, pp. 307–328.
2. Agrawal, R., Imielinski, T., and Swami, A. Mining associations between sets of items in massive databases. In *ACM SIGMOD 1993 Int. Conference on Management of Data* (Washington D.C., 1993), pp. 207–216.
3. Anděl, J. *Statistické metody.* MATFYZPRESS, Praha, 1998.
4. Aumann, Y., and Lindell, Y. A statistical theory for quantitative association rules. In *Knowledge Discovery and Data Mining* (1999), pp. 261–270.
5. Berka, P. *Dobývání znalostí z databází.* Academia, Praha, 2003.
6. Burda, M., Hynar, M., and Šarmanová, J. Pravděpodobnostní logika typovaných relací. In *Znalosti, poster proceedings* (2005).
7. Hájek, P., and Havránek, T. *Mechanizing Hypothesis Formation.* Springer–Verlag, Berlin, 1978. Internet: `http://www.cs.cas.cz/~hajek/guhabook/` (May 2004).
8. Hájek, P., Havránek, T., and Chytil, M. K. *Metoda GUHA – automatická tvorba hypotéz.* Academia, Praha, 1983.
9. Han, J., and Kamber, M. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, USA, 2000.
10. Rauch, J. Asociační pravidla a matematická logika. In *Znalosti* (Brno, 2004), pp. 114–125.
11. Rauch, J., and Šimůnek, M. Alternative approach to mining association rules. In *FDM* (Japan, 2002), pp. 157–162.