A method for association rule quality evaluation based on information theory

D. Sitnikov[1], E. Titova[1] & O. Ryabov[2]
[1]*Kharkov State Academy of Culture, Ukraine*
[2]*National Institute of Advanced Industrial Science and Technology, Japan*

Abstract

The concept of patterns representing functional, logical and other dependencies in data lies in the basis of the Data Mining technology. One of the wide spread forms for representing discovered knowledge patterns is association rules. A method for evaluating an association rule from the viewpoint of information theory has been suggested, which allows us to calculate a generalized characteristic of associations (based on mutual information) with the help of the well known association rule parameters: Support, Confidence and Improvement. Using such a characteristic of associations complements the traditional association parameters and allows setting a linear order on the set of associations, which is useful for evaluating and filtering obtained dependencies. Besides we have carried out analysis of the dependence of the association rule self-descriptiveness on the standard parameters.

1 Introduction

A general definition of association rules has been suggested in [1]: Let $L = I_1, I_2, ..., I_m$ be a set of object features. Let T be a set of records. Each record t is represented by a binary vector $t[k] = 1$ if t contains the feature I_k and $t[k] = 0$ if t does not contain the feature I_k $(k = \overline{1,m})$. Let X be a subset including some features from L, i.e. $X \subseteq L$. We say that the record t satisfies X if $\forall I_k \subseteq X$, $t[k] = 1$. An association rule is an expression in the form $X \to Y$, where $X \subseteq L$, $Y \subseteq L$, at that $X \cap Y = \varnothing$.

WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060031

The association rule $X \to Y$ is supported in the set of records with the confidence level (briefly Conf) c if c% records in T that contain X also contain Y. The rule $X \to Y$ has a support (briefly Sup) of s in the set T if s% records in T contain $X \cup Y$.

To discover association rules a set of algorithms have been developed [1-4, 6, 7], which consist of the major two parts: finding all "large sets" (covers) and forming logic rules by splitting covers into subsets.

There are some problems in evaluating obtained association rules. It has turned out that the well known parameters Support and Confidence are not sufficient to completely characterize obtained association dependencies. Various additional and alternative parameters for evaluating association rules have been suggested in the last years: "interest", correlation, improvement (quite popular now) [3, 5].

Besides, it seems reasonable to develop a certain integral characteristic that would take into account several parameters. How to compare for example two association rules one of which has greater Support and Improvement and the other has a greater Confidence? Which rule is "better" and how much better? It is interesting to find such a characteristic for the analysis of an association, which firstly would be calculated with the help of the above three parameters and secondly would increase when associations are "stronger" and "decrease" when they are weaker.

1.1 Characteristics of association rules from the probabilistic viewpoint

Support, Confidence and Improvement are generally accepted characteristics of association rules [1-4]. Since association rules are probabilistic by nature it seems natural to consider them from the probabilistic viewpoint. Supp($X \to Y$) is defined as the ratio of the number of records in the database satisfying the rule $X \to Y$ to the total number of database records (it can also be defined as the number of records satisfying the given rule). From the probabilistic viewpoint support reflects the probability of the fact that an object (database record) has two features X and Y: P(XY). Conf($X \to Y$) (rule probability) is defined as the ration of the number of records satisfying the rule to the number of records satisfying the antecedent (left side of the rule). From the probabilistic viewpoint confidence reflects the conditional probability $P_X(Y) = \frac{P(XY)}{P(X)}$. If $X \to Y$ is an association rule then the following conditions should be met:

$$\mathrm{Supp}(X \to Y) = P(XY) \geq \min \mathrm{Supp};$$

$$\frac{\mathrm{Supp}(X \to Y)}{\mathrm{Supp}(X)} = P_X(Y) \geq \min \mathrm{Conf},$$

where minSupp, minConf are the minimal acceptable values of support and confidence.

Imp (improvement) is defined as follows:

$$\mathrm{Imp}(X \to Y) = \frac{P_X(Y)}{P(Y)} = \frac{\mathrm{Conf}(X \to Y)}{\mathrm{Supp}(Y)}.$$

Thus improvement characterizes an increase in the probability of the event Y under the condition that the event X has occurred as compared with the unconditional probability of Y.

Let us demonstrate that the above three parameters allow us to fully characterize (in some sense) a binary association. If both features can take on only values 0 and 1 we can write down the following probabilities P_{00}, P_{01}, P_{10}, P_{11} that give us all possible combinations of feature values and can be expressed in terms of Support, Confidence and Improvement.

First let us express Support, Confidence and Improvement of the rule $X \to Y$ in terms of the above probabilities:

$$\mathrm{Supp}(X \to Y) = P(XY) = P_{11};$$

$$\mathrm{Conf}(X \to Y) = \frac{\mathrm{Supp}(X \to Y)}{\mathrm{Supp}(X)} = \frac{P(XY)}{P(X)} = \frac{P_{11}}{P_{11} + P_{10}};$$

$$\mathrm{Imp}(X \to Y) = \frac{\mathrm{Supp}(X \to Y)}{\mathrm{Supp}(X)\mathrm{Supp}(Y)} = \frac{P(XY)}{P(X)P(Y)} = \frac{P_{11}}{(P_{11} + P_{10})(P_{11} + P_{01})}.$$

From these expressions the probability values can be obtained:

$$P_{11} = \mathrm{Supp}(X \to Y); \tag{1}$$

$$P_{10} = \frac{\mathrm{Supp}(X \to Y)}{\mathrm{Conf}(X \to Y)} - \mathrm{Supp}(X \to Y) = \mathrm{Supp}(X \to \overline{Y}); \tag{2}$$

$$P_{01} = \frac{\mathrm{Conf}(X \to Y)}{\mathrm{Imp}(X \to Y)} - \mathrm{Supp}(X \to Y) = \mathrm{Supp}(\overline{X} \to Y); \tag{3}$$

$$P_{00} = 1 - P_{11} - P_{10} - P_{01} =$$
$$= 1 + \mathrm{Supp}(X \to Y) - \frac{\mathrm{Supp}(X \to Y)}{\mathrm{Conf}(X \to Y)} - \frac{\mathrm{Conf}(X \to Y)}{\mathrm{Imp}(X \to Y)} = \mathrm{Supp}(\overline{X} \to \overline{Y}). \tag{4}$$

Thus in order to calculate P_{00}, P_{01}, P_{10}, P_{11} we have used all the parameters Supp, Conf, Imp and it can be easily seen that if any of these parameters is absent the probability values can not be obtained.

It should be noted that values of Supp, Conf and Imp can not be taken arbitrarily. They should satisfy the following conditions:

$$0 \le \mathrm{Supp}(\mathrm{X}\to\mathrm{Y}) \le 1. \tag{5}$$

$$\mathrm{Supp}(\mathrm{X}\to\mathrm{Y}) \le \mathrm{Conf}(\mathrm{X}\to\mathrm{Y}) \le 1. \tag{6}$$

$$\frac{\left(Conf(X\to Y)\right)^2}{Supp(X\to Y)\cdot Conf(X\to Y)+Conf(X\to Y)-Supp(X\to Y)} \le \mathrm{Imp}(\mathrm{X}\to\mathrm{Y}) \le \frac{\mathrm{Conf}(\mathrm{X}\to\mathrm{Y})}{\mathrm{Supp}(\mathrm{X}\to\mathrm{Y})}. \tag{7}$$

The above inequalities can be easily proven.

1.2 Evaluating an association from the viewpoint of information theory

We suggest a method for evaluating an association from the viewpoint of information theory. From this point of view the association rule X→Y can be considered as information on the event Y obtained as a result of receiving a message about the event X. The information "from event to event" is defined as $\mathrm{I}_{\mathrm{X}\to\mathrm{Y}} = \log_2 \frac{\mathrm{P(XY)}}{\mathrm{P(X)P(Y)}}$, which is nothing else but $\log_2(\mathrm{Imp}(\mathrm{X}\to\mathrm{Y}))$.

Mutual information is defined as follows:

$$\mathrm{I}_{\mathrm{X}\leftrightarrow\mathrm{Y}} = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{P}_{ij} \log_2 \frac{\mathrm{P}_{ij}}{\mathrm{p}_i \mathrm{p}_j},$$

where $\mathrm{P}_{ij} = \mathrm{P}((\mathrm{X}\sim \mathrm{x}_i)(\mathrm{Y}\sim \mathrm{y}_j))$ is the probability of the fact that X is in the state x_i and Y is in the state y_j;
$\mathrm{p}_i = \mathrm{P}(\mathrm{X}\sim \mathrm{x}_i)$ is the probability of the fact that X is in the state x_i;
$\mathrm{p}_j = \mathrm{P}(\mathrm{Y}\sim \mathrm{y}_j)$ is the probability of the fact that Y is in the state y_j.

In our case $P_{ij} \in \{P_{00}, P_{01}, P_{10}, P_{11}\}$.

The above considerations allow us to write down a formula for the mutual information of an association, which we call *association self-descriptiveness*:

$$\begin{aligned} I_{X\leftrightarrow Y} &= Supp(X\to Y)\cdot\log_2(\mathrm{Im}p(X\to Y)) \\ &+Supp(X\to \overline{Y})\cdot\log_2(\mathrm{Im}p(X\to \overline{Y})) + Supp(\overline{X}\to Y)\cdot\log_2(\mathrm{Im}p(\overline{X}\to Y)) \\ &+\mathrm{Supp}(\overline{\mathrm{X}}\to\overline{\mathrm{Y}})\cdot\log_2(\mathrm{Imp}(\overline{\mathrm{X}}\to\overline{\mathrm{Y}})). \end{aligned} \tag{8}$$

As we have demonstrated before, the three parameters Sup, Conf and Imp completely describe all the probabilities and therefore all the parameters in the

formula (8) can be expressed in the terms of Sup, Conf and Imp. The expressions for $\mathrm{Supp}(\overline{X} \to \overline{Y})$, $\mathrm{Supp}(\overline{X} \to Y)$ and $\mathrm{Supp}(X \to \overline{Y})$ have been obtained in the formulas (1)-(4). Let us express $\mathrm{Imp}(\overline{X} \to \overline{Y})$, $\mathrm{Imp}(\overline{X} \to Y)$ and $\mathrm{Imp}(X \to \overline{Y})$ in terms of the above association rule characteristics:

$$\mathrm{Imp}(\overline{X} \to Y) = \frac{\mathrm{Supp}(\overline{X} \to Y)}{\mathrm{Supp}(\overline{X})\mathrm{Supp}(Y)} = \frac{P_{01}}{(P_{01} + P_{00})(P_{01} + P_{11})}; \quad (9)$$

$$\mathrm{Imp}(X \to \overline{Y}) = \frac{\mathrm{Supp}(X \to \overline{Y})}{\mathrm{Supp}(X)\mathrm{Supp}(\overline{Y})} = \frac{P_{10}}{(P_{10} + P_{11})(P_{10} + P_{00})}; \quad (10)$$

$$\mathrm{Imp}(\overline{X} \to \overline{Y}) = \frac{\mathrm{Supp}(\overline{X} \to \overline{Y})}{\mathrm{Supp}(\overline{X})\mathrm{Supp}(\overline{Y})} = \frac{P_{00}}{(P_{00} + P_{01})(P_{00} + P_{10})}. \quad (11).$$

Then substitute the expressions for P_{01}, P_{00}, P_{10}, P_{11} from (1)-(4) in (9)-(11) and perform necessary algebraic transformations to get:

$$\mathrm{Imp}(\overline{X} \to Y) = \frac{\mathrm{Conf}(X \to Y) - \mathrm{Supp}(X \to Y) \cdot \mathrm{Imp}(X \to Y)}{\mathrm{Conf}(X \to Y) - \mathrm{Supp}(X \to Y)}; \quad (12)$$

$$\mathrm{Imp}(X \to \overline{Y}) = \frac{\mathrm{Imp}(X \to Y)\,(1 - \mathrm{Conf}(X \to Y))}{\mathrm{Imp}(X \to Y) - \mathrm{Conf}(X \to Y)}; \quad (13)$$

$$\mathrm{Imp}(\overline{X} \to \overline{Y}) = \frac{1 + \mathrm{Supp}(X \to Y) - \dfrac{\mathrm{Supp}(X \to Y)}{\mathrm{Conf}(X \to Y)} - \dfrac{\mathrm{Conf}(X \to Y)}{\mathrm{Imp}(X \to Y)}}{\left(1 - \dfrac{\mathrm{Supp}(X \to Y)}{\mathrm{Conf}(X \to Y)}\right)\left(1 - \dfrac{\mathrm{Conf}(X \to Y)}{\mathrm{Imp}(X \to Y)}\right)}. \quad (14)$$

Now we can write down a formula for mutual information, which is expressed in terms of the well known association characteristic: Supp, Conf and Imp.

$$I_{X \leftrightarrow Y} = Supp(X \to Y) \cdot \log_2(\mathrm{Im}p(X \to Y))$$
$$+\left(\frac{Supp(X \to Y)}{Conf(X \to Y)} - Supp(X \to Y)\right) \cdot \log_2\left(\frac{\mathrm{Im}p(X \to Y)\,(1 - Conf(X \to Y))}{\mathrm{Im}p(X \to Y) - Conf(X \to Y)}\right)$$
$$+\left(\frac{Conf(X \to Y)}{\mathrm{Im}p(X \to Y)} - Supp(X \to Y)\right)$$
$$\times \log_2\left(\frac{Conf(X \to Y) - Supp(X \to Y) \cdot \mathrm{Im}p(X \to Y)}{Conf(X \to Y) - Supp(X \to Y)}\right)$$

$$+\left(1+Supp(X\to Y)-\frac{Supp(X\to Y)}{Conf(X\to Y)}-\frac{Conf(X\to Y)}{\mathrm{Im}p(X\to Y)}\right) \tag{15}$$

$$\times\log_2\left(\frac{1+\mathrm{Supp}(\mathrm{X}\to\mathrm{Y})-\dfrac{\mathrm{Supp}(\mathrm{X}\to\mathrm{Y})}{\mathrm{Conf}(\mathrm{X}\to\mathrm{Y})}-\dfrac{\mathrm{Conf}(\mathrm{X}\to\mathrm{Y})}{\mathrm{Imp}(\mathrm{X}\to\mathrm{Y})}}{\left(1-\dfrac{\mathrm{Supp}(\mathrm{X}\to\mathrm{Y})}{\mathrm{Conf}(\mathrm{X}\to\mathrm{Y})}\right)\left(1-\dfrac{\mathrm{Conf}(\mathrm{X}\to\mathrm{Y})}{\mathrm{Imp}(\mathrm{X}\to\mathrm{Y})}\right)}\right).$$

It should be noted that the expressions in logarithms can not be less than zero, which is due to the inequalities (5-7). If zeros are encountered then the corresponding limit should be used to make calculations $\left(\lim_{\mathrm{P}\to 0}\mathrm{P}\log\mathrm{P}=0\right)$.

It should be noted that the given value (association self-descriptiveness) can be calculated for any probability values, which makes it possible to use it a threshold for filtering association rules (i.e. for cutting off the rules whose self-descriptiveness is less than a given one). If X and Y are independent in the sense that there is no association between them, the rule self-descriptiveness is zero.

The maximum of this parameter $I_{X\leftrightarrow Y}=1$ is reached if P_{00}=0,5; P_{01}=0; P_{10}=0; P_{11}=0,5, i.e. if Supp=0,5; Conf = 1; Imp=2. It is obvious that in real world databases it is hardly possible to obtain such parameters. Our tests have shown that normally rules whose self-descriptiveness is greater than 0.2 can be interesting, but the problem of defining numerical thresholds for associations in various types of databases and for different tasks requires much more research.

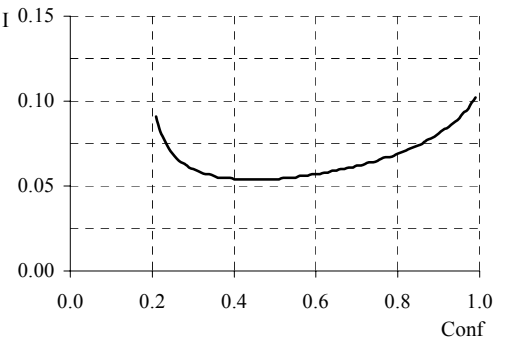

Figure 1.

1.3 Dependence of the association self-descriptiveness on the association rule parameters

Let us consider how Supp, Conf and Imp influence the self-descriptiveness of an association. Traditionally it is known that the greater Supp, Conf and Imp the

better association we have. In order to analyze how the self-descriptiveness of an association depends on the standard characteristics note first that the formula (15) defines a differentiable function with three parameters: Supp, Conf, Imp. Let us draw a graph of this function when Supp and Imp are constant (Supp=0.1; Imp=2), changing Conf in accordance with the inequality (6) (fig.1).

It is obvious that the graph has a minimum. In order to find the value of Conf corresponding to the function minimum the function should be differentiated:

$$\frac{\partial I_{X\leftrightarrow Y}}{\partial Conf} = \frac{Supp}{Conf^2} \times \log_2\left(\frac{Conf\cdot \mathrm{Im}p + Supp\cdot Conf\cdot \mathrm{Im}p - Supp\cdot \mathrm{Im}p - Conf^2}{Conf\cdot \mathrm{Im}p + Supp\cdot Conf\cdot \mathrm{Im}p - Supp\cdot \mathrm{Im}p - Conf^2\cdot \mathrm{Im}p}\right) + \frac{1}{\mathrm{Imp}}\cdot \log_2\left(\frac{\mathrm{Conf}\cdot \mathrm{Imp} + \mathrm{Supp}\cdot \mathrm{Conf}\cdot \mathrm{Imp} - \mathrm{Supp}\cdot \mathrm{Imp}^2 - \mathrm{Conf}^2}{\mathrm{Conf}\cdot \mathrm{Imp} + \mathrm{Supp}\cdot \mathrm{Conf}\cdot \mathrm{Imp} - \mathrm{Supp}\cdot \mathrm{Imp} - \mathrm{Conf}^2}\right).$$

The derivative turns to zero when $\mathrm{Conf} = \sqrt{\mathrm{Supp}\cdot \mathrm{Imp}}$. For example, if Supp=0.1 and Imp=2 the function reaches is minimum if Conf≈0.447.

Thus the rule with a greater confidence level does not necessarily have a greater self-descriptiveness (i.e. it does not necessarily carry more information from X to Y). For example, when Supp=0.3 and Imp=2 a rule with Conf=0.6 carries more information ($I_{X\leftrightarrow Y}$ ≈0.4) than a rule with Conf=0.78 ($I_{X\leftrightarrow Y}$ ≈0.3), although the confidence level of the latter is 1.3 times higher.

Consider now the influence of Supp on the self-descriptiveness of an association. On differentiating the mutual information function with Conf and Imp being constant we get:

$$\frac{\partial I_{X\leftrightarrow Y}}{\partial Supp} = \frac{1}{Conf} \times \log_2\left(\frac{\mathrm{Im}p\,(1-Conf)(Conf-Supp)}{Conf\cdot \mathrm{Im}p + Supp\cdot Conf\cdot \mathrm{Im}p - Supp\cdot \mathrm{Im}p - Conf^2}\right) - \log_2\left(\frac{(1-\mathrm{Conf})(\mathrm{Conf}-\mathrm{Supp}\cdot \mathrm{Imp})}{\mathrm{Conf}\cdot \mathrm{Imp} + \mathrm{Supp}\cdot \mathrm{Conf}\cdot \mathrm{Imp} - \mathrm{Supp}\cdot \mathrm{Imp} - \mathrm{Conf}^2}\right).$$

If Imp>1, $\frac{\partial \mathrm{I}_{\mathrm{X}\leftrightarrow \mathrm{Y}}}{\partial \mathrm{Supp}} > 0$.

The graph of this function with Conf=0.8; Imp=1.5 illustrates the fact that in this case there are no minimums (fig. 2).

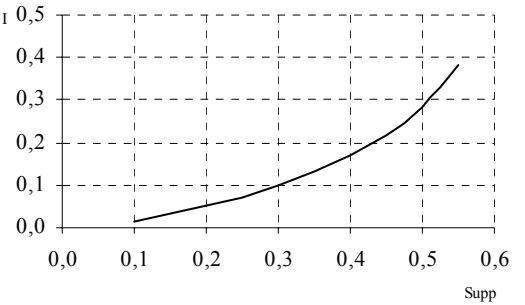

Figure 2.

At last let us investigate how Imp influences the self-descriptiveness of an association.

$$\frac{\partial I_{X\leftrightarrow Y}}{\partial \mathrm{Im}p} = \frac{Conf}{\mathrm{Im}p^{\,2}}$$
$$\times \log_2\left(\frac{Conf\cdot \mathrm{Im}p + Supp\cdot Conf\cdot \mathrm{Im}p - Supp\cdot \mathrm{Im}p - Conf^2}{\mathrm{Im}p - Conf}\right)$$
$$-\frac{\mathrm{Conf}}{\mathrm{Imp}}\cdot \log_2(\mathrm{Conf} - \mathrm{Supp}\cdot \mathrm{Imp}).$$

This derivative turns to zero when Imp=1, which means that in this case the features are independent in the sense that no information is carried from X to Y (mutual information is zero). An example of the above dependence is shown in fig. 3. Imp changes within the limits defined by the inequality (7).

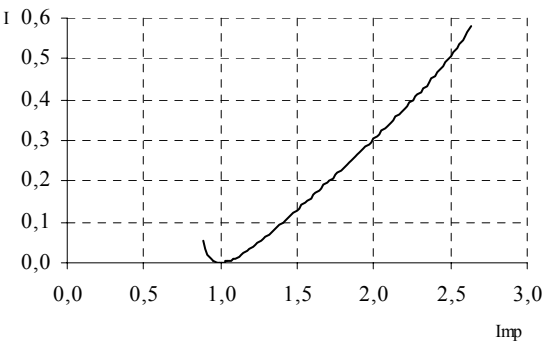

Figure 3.

It should be noted here that Supp and Imp are "symmetric" parameters as far as the "direction" of a rule is concerned, i.e. Supp(X→Y)=Supp(Y→X) and Imp(X→Y)=Imp(Y→X). Therefore when we consider these parameters we can talk about support and confidence of an association and not of an association rule (sometimes these terms are not distinguished). The same is true for the self-descriptiveness of a rule, which is "symmetric" as well. From the viewpoint of information theory it is natural since mutual information from X to Y and from Y to X is always the same. This explains the fact that when Supp and Imp grow the self-descriptiveness grows as well.

Conf is not a "symmetric" parameter, i.e. generally speaking Conf(X→Y)≠Conf(Y→X). Confidence characterizes the rule itself and the rule "direction" is important.

2 Conclusion

The suggested method for evaluating associations from the viewpoint of information theory is based on calculating an integral characteristic (mutual information), which allows us to compare associations with different values of support, confidence and improvement. Using such a characteristic does not suppose rejecting the use of the standard parameters but complements them and allows taking into account the values of these parameters. The self-descriptiveness of a rule (mutual information) can be used to filter the associations discovered in a database.

The analysis of the dependence of the association self-descriptiveness on the standard characteristics (support, confidence, and improvement) has demonstrated that this dependence is not obvious (it is not always true that the higher the levels of support, confidence and improvement are, the greater the self-descriptiveness is). Thus rules with "bad" values of Sup, Conf or Imp can also be interesting from the viewpoint of the information transferred. Of course, finding associations with "bad" (small) values of support, confidence and improvement is associated with additional expenses (time, memory) and problems associated with the necessity of finding such rules require additional research for particular tasks.

It should be noted that the above considerations are true for binary associations between two features that take on only Boolean values. The cases of more than two Boolean features and arbitrary discrete features are yet to be investigated.

References

[1] *R.Agrawal, T.Imielinski, A.Swami.* Mining association rules between sets of items in large databases // Proc. of the ACM SIGMOD Conference. – Washington DC, USA, May 1993. – P. 207-216.

[2] *R.Agrawal, R.Srikant*. Fast algorithms for mining association rules // Proc. of the 20th VLDB Conference Santiago. – Chile, September 1994. – P. 487-499.

[3] *R.Srikant, R.Agrawal*. Mining generalized association rules // Proc. of the 21st VLDB Conference Zurich. – Switzerland, September 1995. – P. 407-419.
[4] *A.Amir, R.Feldman, R.Kashi.* A new and versatile method for association generation // Information Systems. – 1997. – Vol. 22, № 6/7. – P. 333-347.
[5] *Аджиев В.* MineSet – визуальный инструмент аналитика // Открытые системы. – 1997. – №3. – С. 72-77.
[6] *Hannu Toivonen.* Sampling Large Databases for Association Rules // Proc. of the 22nd International Conference on Very Large Databases. – Mumbai, India, 1996. – P. 134-145.
[7] *Ashoka Savasere, Edward Omiecinski, Shamkant B. Navathe.* An Efficient Algorithm for Mining Association Rules in Large Databases // Proc. of the 21st International Conference on Very Large Databases. – Zurich, Switzerland, 1995. – P. 432-444.