COMPUTATIONAL PROBLEMS OF ELECTRICAL ENGINEERING

Vol. 2, No. 1, 2012

PROPERTIES OF PROBABILITY PRODUCTIVE DEPENDENCIES IN THE DATA ANALYSIS OF LARGE DATA VOLUMES

Oleksandr Pshenychnyi
Lviv Polytechnic National University
sasha.pshenychniy@gmail.com

Abstract: This paper describes findings in the area of aggregated associative dependencies detection. The work gives a method of building aggregated associative dependencies in large data volumes. This research can be applied to the wide range of data types and optimize data mining processes

Key words: aggregated associative dependency, data mining, large data volume, functional dependency.

Introduction

Nowadays computer systems operate with huge amount of data: there are plenty of terabyte storages and even a few petabyte-storages of structured data. Although hardware becomes more powerful every month, it isn't able to analyze these data storages.

This research discusses the problem of detection of associative relationships in large volumes of data, shows the results of investigation into the properties of associative dependencies and ways to optimize the search of such dependencies in large data volumes.

Data mining methods identifying dependencies and correlations are widely used in sociology, psychology, political science, physics, power industry, astronomy, computer science, and many other applied disciplines. The task of identifying the associative dependencies in polls was considered in [1]. This trend in data analysis is relatively old, but active research is still held in this area. For example, the research [2] describes a method of aggregate association rules of a construction based on simpler dependencies. This interest in the identification of data relationships can be explained by the increase in processing capacity of computer technologies and increasing volumes of collected data in many social areas. The data is collected in such amounts that its expert analysis is incomplete or may be even impossible. Modern computing hardware allows us to implement more intricate algorithms and to apply them to large amounts of data. It encourages researchers to develop such algorithms, and motivates the owners of large databases and data warehouses to develop data analysis software for the accumulated information.

Some branches of science and technology have already had powerful methods of data analysis, specifically designed to fit their needs and data structures. The most notable ones are the following software products: CLASSIFI (Department of Pathology, UT Southwestern Medical Center) [3], BiNGO (Department of Plant Systems Biology, VIB / Ghent University) [4] and EASE (National Institute of Allergy and Infectious Diseases) [5]. However, most research institutions can't afford to develop such systems and require a common method, applicable to the wide range of data types and structures.

Thus, the effective detection of associative dependencies in multi-attribute data is an actual problem of modern data analysis.

It should be noted that data mining is a very broad area of data analysis, and the detection of associative relationships is only a part of it.

The goal of this research is to investigate the properties of aggregated associative dependencies that will help to implement effective data dependency identification algorithms, applicable to various data types and structures.

Relations and the work area of existing methods of data analysis

Data mining technologies are mostly used to identify data dependencies of the types "if … then …" or "for … is true …". Such dependencies are represented by implications, production rules or associative rules.

Data mining includes the wide range of mathematical and algorithmic tools, such as neural networks, evolutional algorithms, decision trees etc. However, modern research is more and more accented on logical dependencies of the search in datasets. With their help the problems of classification, forecasting, creation of formal description of real objects and others are solved. [2]

It is possible to find an incredibly big amount of associative dependencies in large databases. It is impossible even to store all those dependencies in any data storage in the world, let alone analyzing them. Fortunately, it is not the goal of any business branch. Instead, there is a need for finding associative dependencies, which meet some predefined quality requirements (they are statistically substantiated).

The main problems of currently available data dependency detection methods are as follows:

they can work only with binary properties of objects;

they don't find dependencies with low support;

they aren't able to effectively process addition of new data to data source;

they are non-efficient in analyzing multi-attribute data;

they are not flexible enough to satisfy the needs of dependency quality filtering.

One of the ways to solve some of these problems is building aggregated associative rules. The research [2] offers to use a system of 4 parameters of an associative dependency, which describe the rule. In general, such complex of criteria is more flexible and useful than a single association intensity attribute, proposed in [1], but sufficiency of this system has not yet been proven.

Taken into consideration the aforesaid, the given research discusses the urgent scientific and technical problems of the detection of aggregated associative dependencies and the development of valuation methods for the analysis of large databases.

Problem Statement

The search of arbitrary associative rules $P\left(x\right) \to Q\left(x\right), x \in r\left(R\right)$ in ratio $r\left(R\right)$ is a very broad task, the solution to which is still to be found and this problem is not the object of our research. The paper proposes studying the properties of associative dependencies, in which conditional and resulting predicates look like:

$$\begin{aligned} &P = P_1^e \vee P_2^e \vee ... \vee P_h^e = \bigvee_{k=1}^{h} P_k^e, \\ &P_k^e = A_{i_1} \in \left\{a_{(i_1)(j_1)}\right\} \wedge A_{i_2} \in \overline{\left\{a_{(i_2)(j_2)}\right\}} \wedge A_{i_k} \in \left\{a_{(i_k)(j_k)}\right\}, \\ &\forall l = \overline{1..k} : A_{i_l} \in R, \forall m = 1..\left|\left\{a_{(i_l)(j_m)}\right\}\right| : a_{(i_l)(j_m)} \in dom\left(A_{i_l}\right) \\ &\forall i, j \in \left\{1..h\right\} : \arg\left(P_i^e\right) = \arg\left(P_j^e\right) \end{aligned} \quad (1)$$

$$F_I : \bigvee_{k=1}^{s} P_k^e \to \bigvee_{l=1}^{t} Q_l^e \quad (2)$$

The denotation $\arg\left(P\right)$ is used for the operator returning a set of attribute-arguments of a predicate P.

We assign to such associative dependencies a separate term – probabilistic productive dependency (PPD). Various sources use different markings, interpretations and explanations of this concept. For example, in [2] a term "intensity of association" is used, in [1] it is "degree of confidence". However, these notions concern to a greater extent the existing data, and not the data to be worked with currently. Moreover, considering the system as static, not being renewed with new knowledge, we lose the sense of using such techniques as Laplacian smoothing [6] and other methods of protection against noise and data uncertainty. Just for emphasising on the study of random processes, their dynamics and problems, the term "probability" has been added to the term of "dependency", which is being investigated. The second part of the term, namely "productive", does not need a special explanation, since the productive rule underlies a dependence.

Thus, the probabilistic productive dependence is a productional rule of a type Eq.(2) in the selection of the basic relation that holds for a significant number of objects of the selection. The threshold of significance should be determined in an expert way or based on calculations of the probability of a false selection of this relationship.

Let's write the Eqs.(1), (2) in terms of relational algebra:

$$\begin{aligned} &P^e\left(x\right) = \pi_{A_{i_1}A_{i_2}...A_{i_l}}\left(x\right) \in \left\{a_{(i_1)(j_1)}\right\} \times \overline{\left\{a_{(i_2)(j_2)}\right\}} \times ... \times \left\{a_{(i_k)(j_k)}\right\}, \\ &x \in r\left(R\right), \forall l = \overline{1..k} : A_{i_l} \in R, \forall m = 1..\left|\left\{a_{(i_l)(j_m)}\right\}\right| : \\ &a_{(i_l)(j_m)} \in dom\left(A_{i_l}\right) \end{aligned} \quad (3)$$

$$\begin{aligned} &P\left(x\right) = \pi_{A_{i_1}A_{i_2}...A_{i_l}}\left(x\right) \in \left\{a_{(i_1)(j_{1,1})}\right\} \times \left\{a_{(i_2)(j_{1,2})}\right\} \times ... \times \left\{a_{(i_k)(j_{1,k})}\right\} \cup \\ &\cup \left\{a_{(i_1)(j_{2,1})}\right\} \times \left\{a_{2,(i_2)(j_{2,2})}\right\} \times ... \times \left\{a_{(i_k)(j_{2,k})}\right\} \cup ... \cup \\ &\cup \left\{a_{(i_1)(j_{h,1})}\right\} \times \left\{a_{2,(i_2)(j_{h,2})}\right\} \times ... \times \left\{a_{(i_k)(j_{h,k})}\right\} \end{aligned} \quad (4)$$

In other words, PPD predicates can be presented as predicates defined on processions of a ratio $r\left(R\right)$, not only on a set of attributes.

The PPD threshold of significance can be defined on the basis of an arbitrary function of assessing the importance of a detected dependence. However, the most often used indicators are those of a level of support and a level of confidence. In the source [2] it was shown that these parameters are insufficient to describe adequately the dependencies of a subject area and it was proposed to use the additional indicators: a level of improvement and a full mutual information.

Let us consider these notions in detail:

Level of support is a characteristic of a selection predicate in a relation, calculated as the ratio of the number of objects that satisfy the predicate P to the total number of objects in the relation:

$$Sup\left(P\right) = \frac{\left|\sigma_P\left(r\right)\right|}{\left|r\right|} \quad (5)$$

When calculating the level of support for PPD, conditional and resulting predicate are combined by a sign of conjunction:

$$Sup\left(S \to T\right) = Sup\left(S \wedge T\right) = \frac{\left|\sigma_{S \wedge T}\left(r\right)\right|}{\left|r\right|} \quad (6)$$

Level of confidence – the ratio of the number of objects, for which there is such a PPD to the number of objects in the selection:

$$Conf\left(S \to T\right) = P\left(S \to T\right) = \frac{\left|\sigma_{S \wedge T}\left(r\right)\right|}{\left|\sigma_{S}\left(r\right)\right|} \qquad (7)$$

Using the concept of support level, the confidence level can be calculated as:

$$Conf\left(S \to T\right) = \frac{Sup\left(S \to T\right)}{Sup\left(S\right)} \qquad (8)$$

Level of improvement is calculated as the ratio of the confidence levels and the support levels of PPD:

$$\mathrm{Imp}\left(S \to T\right) = \frac{Conf\left(S \to T\right)}{Sup\left(T\right)} = \frac{Sup\left(S \wedge T\right)}{Sup\left(S\right) \cdot Sup\left(T\right)} \qquad (9)$$

Full mutual information is calculated in general as:

$$I_{X \leftrightarrow Y} = \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \log_2 \frac{p_{ij}}{p_i r_j}, \qquad (10)$$

where $p_{ij} = P\left(X = x_i \wedge Y = y_j\right)$ – the probability of the fact that X is in the condition x_i, and Y – in the condition y_j; $p_i = P\left(X = x_i\right)$ – the probability of the fact that X is in the condition x_i; $r_j = P\left(Y = y_j\right)$ – the probability of the fact that Y is in the condition y_j.

For associative rules the mutual information can be defined as:

$$I_{X \leftrightarrow Y} = \sum_{i=1}^{n}\sum_{j=1}^{m} Sup(x_i \to y_j)\log_2 Imp(x_i \to y_j) \quad (11)$$

Analogs of the rules of the functional dependencies derivation for PPD

As in the case with F-dependencies, the PPD set occurring in a given ratio can be represented by their certain subset, from which all the relevant PPD from the given ratio can be received by means of derivation rules. Since PPD is an extension of F-dependencies, we should consider the transformation of axioms of the functional dependencies F1-F6 derivation for PPD.

PPDs are characterized by many parameters, the most important of which are given in Eqs.(5)–(11). However, a parameter which is used the most often and is the easiest to understand is that of the confidence level. More sophisticated characteristics of dependencies are based on it, what has been found during the investigations.

Filtering by the level of support does not allow to derive PPDs from small partial dependencies – the level of support certainly increases when combining PPDs, and therefore it is impossible to make a cut-off of some groups of dependencies based on this parameter.

The level of improvement is a nonlinear characteristic. That makes it impossible to perform a clipping of PPD generation based on the set of existing dependencies.

The full mutual information is also nonlinear by the power of the selection of both PPD predicates.

Thus, there is no sense in building derivation rules for the parameters of the improvement level and the full mutual information, because they are calculated from the level of support and the level of confidence; in addition, the given characteristics nonlinearly depend on the number of processions, which corresponds to the number of dependencies.

A strict proof of this fact is not given, because it is intuitively understood from the above considerations and a formal proof is rather cumbersome.

We resolve the problem of associative dependencies filtering flexibility using two-step filtering. It allows using arbitrary criteria of PPD quality at the second stage.

So, let us consider the transformation of the functional dependencies derivation for PPD:

Reflexivity of the confidence level.

$Conf\left(s \in S \to s \in S\right) = 1$ for any ratio r(R).

Proof:

$$Conf\left(s \in S \to s \in S\right) = \frac{\left|\sigma_{s \in S \wedge s \in S}\right|}{\left|\sigma_{s \in S}\right|} = \frac{\left|\sigma_{s \in S}\right|}{\left|\sigma_{s \in S}\right|} = 1$$

Replenishment of the confidence level. If $Conf\left(s \in S \to t \in T\right) = p$, then $Conf\left(s \in S \wedge w \in D\left(W\right) \to t \in T\right) = p$, where $D\left(W\right)$ – the domain of the attribute W of a ratio $r\left(R\right)$.

Proof:

$$Conf\left(s \in S \wedge w \in D\left(W\right) \to t \in T\right) = \frac{\left|\sigma_{s \in S \wedge w \in D(W) \wedge t \in T}\left(R\right)\right|}{\left|\sigma_{s \in S \wedge w \in D(W)}\left(R\right)\right|} =$$

$$= \left|\forall x \in r : q = \pi_{W=w}\left(x\right) \in D\left(W\right) \Rightarrow w \in D\left(W\right)\right| =$$

$$= \frac{\left|\sigma_{s \in S \wedge t \in T}\left(R\right)\right|}{\left|\sigma_{s \in S}\left(R\right)\right|} = Conf\left(s \in S \to t \in T\right) = p$$

$$Conf\left(a \in \left\{a_1, a_2\right\} \to d \in \left\{d_1\right\}\right) = \frac{4}{8} = 0,5$$

$$Conf\left(a \in \left\{a_1, a_2\right\} \wedge b \in \left\{b_1, b_2, b_3, b_4\right\} \to d \in \left\{d_1\right\}\right) =$$

$$= \frac{\left|\sigma_{a \in \{a_1, a_2\} \wedge b \in \{b_1, b_2, b_3, b_4\} \wedge d \in \{d_1\}}\left(R\right)\right|}{\left|\sigma_{a \in \{a_1, a_2\} \wedge b \in \{b_1, b_2, b_3, b_4\}}\left(R\right)\right|} = \frac{4}{8} = 0,5$$

Table 1

Example 1

A	B	C	D
a_1	b_1	c_1	d_1
a_1	b_1	c_2	d_1
a_1	b_2	c_2	d_1
a_2	b_2	c_2	d_1
a_2	b_2	c_2	d_2
a_2	b_2	c_2	d_2
a_2	b_3	c_3	d_2
a_2	b_2	c_2	d_2
a_3	b_3	c_3	d_2

Additivity of the confidence level. If $Conf\left(s \in S \to t \in T\right) = p$ and $Conf\left(s \in S \to w \in W\right) = 1$, then $Conf\left(s \in S \to t \in T \wedge w \in W\right) = p$.

Proof:

$$Conf\left(s \in S \to t \in T \wedge w \in W\right) = \frac{\left|\sigma_{s \in S \wedge t \in T \wedge w \in W}\right|}{\left|\sigma_{s \in S}\right|} =$$

$$= \left|s \in s \to w \in W\right| = \frac{\left|\sigma_{s \in S \wedge t \in T}\right|}{\left|\sigma_{s \in S}\right|} = Conf\left(s \in S \to t \in T\right) = p$$

Example 2:

Considering PPD from the example 1:

$$Conf\left(A \in \left\{a_1\right\} \to B \in \left\{b_1\right\}\right) = \frac{2}{3}$$

$$Conf\left(A \in \left\{a_1\right\} \to D \in \left\{d_1\right\}\right) = 1$$

Thereof, it can be concluded that $Conf\left(A \in \left\{a_1\right\} \to B \in \left\{b_1\right\} \wedge D \in \left\{d_1\right\}\right) = \frac{2}{3}$. It is proved by the calculations with the Eq.(7):

$$Conf\left(A \in \left\{a_1\right\} \to B \in \left\{b_1\right\} \wedge D \in \left\{d_1\right\}\right) = \frac{\left|\sigma_{A \in \left\{a_1\right\} \wedge B \in \left\{b_1\right\} \wedge D \in \left\{d_1\right\}}\right|}{\left|\sigma_{A \in \left\{a_1\right\}}\right|} = \frac{2}{3}$$

Projectivity of the confidence level. If

$$Conf\left(s \in S \to t \in T \wedge w \in W\right) = p$$

and

$$Conf\left(s \in S \to w \in W\right) = 1,$$

then $Conf\left(s \in S \to t \in T\right) = p$.

Proof:

$$Conf\left(s \in S \to t \in T\right) = \frac{\left|\sigma_{s \in S \wedge t \in T}\right|}{\left|\sigma_{s \in S}\right|} =$$

$$= \frac{\left|\sigma_{s \in S \wedge t \in T \wedge w \in W}\right|}{\left|\sigma_{s \in S}\right|}\left|s \in S \to w \in W\right. =$$

$$= Conf\left(s \in S \to t \in T \wedge w \in W\right) = p$$

Example 3:

Let's consider the example, introduced in the previous paragraph, in a reverse variant:

From PPD

$Conf\left(A \in \left\{a_1\right\} \to B \in \left\{b_1\right\} \wedge D \in \left\{d_1\right\}\right) = \frac{2}{3}$ and $Conf\left(A \in \left\{a_1\right\} \to D \in \left\{d_1\right\}\right) = 1$ it can be concluded that $Conf\left(A \in \left\{a_1\right\} \to B \in \left\{b_1\right\}\right) = \frac{2}{3}$. The verification can be done with the help of the Eq.(7):

$$Conf\left(A \in \left\{a_1\right\} \to B \in \left\{b_1\right\}\right) = \frac{\left|\sigma_{A \in \left\{a_1\right\} \wedge B \in \left\{b_1\right\}}\right|}{\left|\sigma_{A \in \left\{a_1\right\}}\right|} = \frac{2}{3}$$

Transitivity of the confidence level. If

$$Conf\left(s \in S \to t \in T\right) = p$$

and

$$Conf\left(t \in T \to w \in W\right) = 1,$$

then

$$Conf\left(s \in S \to w \in W\right) \geq p.$$

Proof:

$$Conf\left(s \in S \to w \in W\right) = \frac{\left|\sigma_{s \in S \wedge w \in W}\right|}{\left|\sigma_{s \in S}\right|}$$

$$Conf\left(s \in S \to t \in T\right) = \frac{\left|\sigma_{s \in S \wedge t \in T}\right|}{\left|\sigma_{s \in S}\right|} = p$$

Thus, since $\left|\sigma_{s \in S}\right| \geq 0$, $\left|\sigma_{s \in S \wedge w \in W}\right| \geq 0$ i $\left|\sigma_{s \in S \wedge t \in T}\right| \geq 0$ (follows from the definition of a relational selection operation), then in order to prove inequality $Conf\left(s \in S \to w \in W\right) \geq Conf\left(s \in S \to t \in T\right)$ we need to prove that $\left|\sigma_{s \in S \wedge w \in W}\right| \geq \left|\sigma_{s \in S \wedge t \in T}\right|$.

Let's consider a variable-procession x of a ratio r(R), provided that $\pi_s\left(x\right) \in S$ and $\pi_t\left(x\right) \in T$. According to the condition $Conf\left(t \in T \to w \in W\right) = 1$, if $\pi_T\left(x\right) \in T$, then $\pi_w\left(x\right) \in W$. Thus

$$Conf\left(s \in S \to t \in T\right) = p \wedge Conf\left(t \in T \to w \in W\right) = 1:$$
$$\forall x \in r\left(R\right): \pi_s\left(x\right) \in S \wedge \pi_t\left(x\right) \in T \Rightarrow \pi_w\left(x\right) \in W$$

Hence we have got a number of consequences:

$$\sigma_{s \in S \wedge w \in W} \subseteq \sigma_{s \in S \wedge t \in T}$$
$$\left|\sigma_{s \in S \wedge w \in W}\right| \geq \left|\sigma_{s \in S \wedge t \in T}\right|$$
$$Conf\left(s \in S \to w \in W\right) \geq Conf\left(s \in S \to t \in T\right) = p$$
$$Conf\left(s \in S \to w \in W\right) \geq p$$

Hereby, the transitivity of the PPD confidence level has been proved.

Example 4:

Using the data of the example 2 the following PPD can be created:

$$Conf\left(C \in \left\{c_2\right\} \to B \in \left\{b_1\right\}\right) = \frac{1}{6}$$

$$Conf\left(B \in \left\{b_1\right\} \to D \in \left\{d_1\right\}\right) = 1$$

From the rule of the transitivity of the PPD confidence level we conclude that $Conf\left(C \in \left\{c_2\right\} \to D \in \left\{d_1\right\}\right) \geq \frac{1}{6}$. Let us verify this by calculating $Conf\left(C \in \left\{c_2\right\} \to D \in \left\{d_1\right\}\right)$ using the Eq.(7):

$$Conf\left(C \in \left\{c_2\right\} \to D \in \left\{d_1\right\}\right) = \frac{\left|\sigma_{C \in \{c_2\} \wedge D \in \{d_1\}}\right|}{\left|\sigma_{C \in \{c_2\}}\right|} = \frac{3}{6} = \frac{1}{2}$$

$$Conf\left(C \in \left\{c_2\right\} \to D \in \left\{d_1\right\}\right) =$$

$$= \frac{1}{2} \geq \frac{1}{6} = Conf\left(C \in \left\{c_2\right\} \to B \in \left\{b_1\right\}\right)$$

Thus, the transitivity of the PPD confidence level has been confirmed.

The transitivity of the PPD confidence level is a powerful rule for making various assumptions and proofs.

Pseudotransitivity of the confidence level. This axiom of an F-dependencies derivation does not have a direct alternative for PPD, under the condition that only one restriction is imposed on the source dependencies. We are going to prove this statement.

Let us consider the dependencies

$Conf\left(s \in S \to t \in T\right) = p$ and $Conf\left(t \in T \wedge q \in Q \to w \in W\right) = 1$, implying a limitation of genuineness on $Conf\left(t \in T \wedge q \in Q \to w \in W\right)$.

We denote $X = \sigma_{s \in S}\left(R\right)$, $Y = \sigma_{t \in T}\left(R\right)$, $Z = \sigma_{q \in Q}\left(R\right)$, $V = \sigma_{w \in W}\left(R\right)$. Then

$$Conf\left(t \in T \wedge q \in Q \to w \in W\right) = \frac{\left|Y \cap Z \cap V\right|}{\left|Y \cap Z\right|} = 1$$

$$\left|X \cap Z \cap V\right| = \left|X \cap Z\right|$$

$$Conf\left(s \in S \to t \in T\right) = \frac{\left|X \cap Y\right|}{\left|X\right|} = p$$

$$Conf\left(s \in S \wedge q \in Q \to w \in W\right) = \frac{\left|X \cap Z \cap V\right|}{\left|X \cap Z\right|} =$$

$$= \frac{\left|\left(\left(X \cap Y\right) \cup \left(X \setminus Y\right)\right) \cap Z \cap V\right|}{\left|\left(\left(X \cap Y\right) \cup \left(X \setminus Y\right)\right) \cap Z\right|} = \frac{\left|\left(X \cap Y \cap Z \cap V\right) \cup \left(\left(X \setminus Y\right) \cap Z \cap V\right)\right|}{\left|\left(X \cap Y \cap Z\right) \cup \left(\left(X \setminus Y\right) \cap Z\right)\right|} =$$

$$= \frac{\left|\left(X \cap Y \cap Z\right) \cup \left(\left(X \setminus Y\right) \cap Z \cap V\right)\right|}{\left|\left(X \cap Y \cap Z\right) \cup \left(\left(X \setminus Y\right) \cap Z\right)\right|} =$$

$$= \frac{\left|X \cap Y \cap Z\right| + \left|\left(X \setminus Y\right) \cap Z \cap V\right| - \left|X \cap Y \cap Z \cap \left(X \setminus Y\right) \cap Z \cap V\right|}{\left|X \cap Y \cap Z\right| + \left|\left(X \setminus Y\right) \cap Z\right|} =$$

$$= \frac{\left|X \cap Y \cap Z\right| + \left|\left(X \setminus Y\right) \cap Z \cap V\right|}{\left|X \cap Y \cap Z\right| + \left|\left(X \setminus Y\right) \cap Z\right|} = \left(1\right)$$

$$X \cap Y \cap Z \subset X \cap Z$$

$$\left(X \setminus Y\right) \cap Z \cap V \subset X \cap Z$$

$$\left(X \setminus Y\right) \cap Z \subset X \cap Z$$

Thereby, $X \cap Z$ is a universal set U of given expressions and the result of calculations Eq.(1) will not change, if we consider only the processions from $X \cap Z$. We set $Y' = Y \cap \left(X \cap Z\right)$, $V' = V \cap \left(X \cap Z\right)$. Then

$$\left(1\right) = \frac{\left|Y'\right| + \left|\neg Y' \cap V'\right|}{\left|Y'\right| + \left|\neg Y'\right|} = \frac{\left|Y'\right| + \left|\neg Y' \cap V'\right|}{\left|U\right|}$$

$$\left|\neg Y' \cap V'\right| \in \left[0; \left|\neg Y'\right|\right]$$

$$\frac{\left|Y'\right| + \left|\neg Y' \cap V'\right|}{\left|U\right|} \in \left[\frac{\left|Y'\right|}{\left|U\right|}; 1\right]$$

Let us return to the introduced settings: $Y' = Y \cap \left(X \cap Z\right)$. Initial conditions do not impose restrictions on the value of a given expression, so $\left|Y'\right| \in \left[0; \left|U\right|\right]$ and accordingly

$$\frac{\left|Y'\right| + \left|\neg Y' \cap V'\right|}{\left|U\right|} \in \left[0; 1\right]$$

$$Conf\left(s \in S \wedge q \in Q \to w \in W\right) \in \left[0; 1\right]$$

Thus, the dependencies $Conf\left(s \in S \to t \in T\right) = p$ i $Conf\left(t \in T \wedge q \in Q \to w \in W\right) = 1$ do not have any influence on the dependence $Conf\left(s \in S \wedge q \in Q \to w \in W\right)$.

Now let us consider the restrictions of the other dependence, combining $Conf\left(s \in S \to t \in T\right) = 1$ and $Conf\left(t \in T \wedge q \in Q \to w \in W\right) = p$. Using the settings given above, we obtain:

$$Conf\left(t \in T \wedge q \in Q \to w \in W\right) = \frac{\left|Y \cap Z \cap V\right|}{\left|Y \cap Z\right|} = p$$

$$Conf\left(s \in S \to t \in T\right) = \frac{\left|X \cap Y\right|}{\left|X\right|} = 1$$

$$\left|X \cap Y\right| = \left|X\right|$$

$$Conf\left(s \in S \wedge q \in Q \to w \in W\right) = \frac{\left|X \cap Z \cap V\right|}{\left|X \cap Z\right|}$$

In this case, the initial dependencies do not limit the resulting expression, and it is apparent just at the first step: a set V may not have mutual processions with X, immediately turning

$Conf\left(s \in S \wedge q \in Q \to w \in W\right)$ into zero. On the other hand, the variant $V \subset X \cap Z$ is also possible – then $Conf\left(s \in S \wedge q \in Q \to w \in W\right) = 1$. The function $Conf\left(s \in S \wedge q \in Q \to w \in W\right)$ linearly depends on $\left|X \cap Z \cap V\right|$, having a variable set V, and thus having a range of values $\left[0;1\right]$. So, the dependencies $Conf\left(s \in S \to t \in T\right) = 1$ and $Conf\left(t \in T \wedge q \in Q \to w \in W\right) = p$ do not limit the range of values $Conf\left(t \in T \wedge q \in Q \to w \in W\right)$.

Thus, it has been proved that the pseudotransitivity of functional dependencies does not have a direct analog among PPD.

As we can see from the above arguments, most axioms of the F-dependencies derivation may be transformed for PPD only with the significant limitations of the conditional part of one of the dependencies. In addition, some analogs do not give a precise formula for calculating the confidence level of a new dependence, but only limit it. Thus, the given set of PPD derivation rules is not complete. To ensure the completeness of derivation rules, it is necessary to introduce additional derivation rules, specific for PPD.

Operations on PPD

Factorization

We shall name the expansion of the dependence F_I $\pi_{A_{i_1}A_{i_2}\ldots A_{i_k}}\left(s\right) \in \left\{s_1, s_2, \ldots s_m\right\} \to \pi_{A_{j_1}A_{j_2}\ldots A_{j_l}}\left(s\right) \in \left\{t_1, t_2, \ldots t_n\right\}$ into the set of dependencies

$$\left\{\pi_{A_{i_1}A_{i_2}\ldots A_{i_k}}\left(s\right) = s_i \to \pi_{A_{j_1}A_{j_2}\ldots A_{j_l}}\left(s\right) = t_j\right\}, i = \overline{1..m}, j = \overline{1..n}$$

a factorization and denote it as $F_I\left[Fact\right]$.

$$F_I = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\pi_{A_{i_1}A_{i_2}\ldots A_{i_k}}\left(s\right) = s_i \to \pi_{A_{j_1}A_{j_2}\ldots A_{j_l}}\left(s\right) = t_j\right) \quad (12)$$

Merging

Merging of PPD $s \in S_1 \to t \in T_1$ і $s \in S_2 \to t \in T_2$ is a new PPD $s \in S_1 \cup S_2 \to t \in T_1 \cup T_2$ (PPD denotations in terms of relational algebra are used – Eq. (4)).

$$\begin{aligned}&\left(s \in S_1 \to t \in T_1\right) + \left(s \in S_2 \to t \in T_2\right) = \\ &= s \in S_1 \cup S_2 \to t \in T_1 \cup T_2\end{aligned} \quad (13)$$

Let us consider qualities of the PPD merging operation:

Commutativity

The operation of PPD merging has the quality of commutativity.

$$\begin{aligned}&\left(s \in S_1 \to t \in T_1\right) + \left(s \in S_2 \to t \in T_2\right) = \\ &= \left(s \in S_2 \to t \in T_2\right) + \left(s \in S_1 \to t \in T_1\right)\end{aligned} \quad (14)$$

Proof:

$$\begin{aligned}&\left(s \in S_1 \to t \in T_1\right) + \left(s \in S_2 \to t \in T_2\right) = \\ &= \left(s \in S_1 \cup S_2 \to t \in T_1 \cup T_2\right) = \\ &= \left(s \in S_2 \cup S_1 \to t \in T_2 \cup T_1\right) = \\ &= \left(s \in S_2 \to t \in T_2\right) + \left(s \in S_1 \to t \in T_1\right)\end{aligned}$$

Associativity

The operation of PPD merging has the quality of associativity.

$$\begin{aligned}&\left(s \in S_1 \to t \in T_1\right) + \left(\left(s \in S_2 \to t \in T_2\right) + \left(s \in S_3 \to t \in T_3\right)\right) = \\ &\left(\left(s \in S_1 \to t \in T_1\right) + \left(s \in S_2 \to t \in T_2\right)\right) + \left(s \in S_3 \to t \in T_3\right)\end{aligned} \quad (15)$$

Proof:

$$\begin{aligned}&\left(s \in S_1 \to t \in T_1\right) + \left(\left(s \in S_2 \to t \in T_2\right) + \left(s \in S_3 \to t \in T_3\right)\right) = \\ &= \left(s \in S_1 \to t \in T_1\right) + \left(s \in S_2 \cup S_3 \to t \in T_2 \cup T_3\right) = \\ &= s \in S_1 \cup S_2 \cup S_3 \to t \in T_1 \cup T_2 \cup T_3\end{aligned}$$

We expand the right part of the associativity expression:

$$\begin{aligned}&\left(\left(s \in S_1 \to t \in T_1\right) + \left(s \in S_2 \to t \in T_2\right)\right) + \left(s \in S_3 \to t \in T_3\right) = \\ &= \left(s \in S_1 \cup S_2 \to t \in T_1 \cup T_2\right) + \left(s \in S_3 \to t \in T_3\right) = \\ &= s \in S_1 \cup S_2 \cup S_3 \to t \in T_1 \cup T_2 \cup T_3 = \\ &= \left(s \in S_1 \to t \in T_1\right) + \left(\left(s \in S_2 \to t \in T_2\right) + \left(s \in S_3 \to t \in T_3\right)\right)\end{aligned}$$

Rules of PPD Derivation

As it has been shown, the implementation of the transformed rules of the functional dependencies derivation is not enough to ensure the complete set of PPD derivation rules. We should consider the derivation rules, specific for PPDs, which will allow us to develop efficient algorithms for the search of these dependencies in databases.

Aggregation of a definition area. If there are PPD $s \in S_1 \to t \in T_1$ and $s \in S_2 \to t \in T_2$ and values $\sigma_{s=s_i}\left(r\left(R\right)\right)$, $\sigma_{t=t_j}\left(r\left(R\right)\right)$, $\sigma_{s=s_i \wedge t=t_j}\left(r\left(R\right)\right)$ provided that $\bigcup\limits_i s_i = S_1 \cup S_2$, $\bigcup\limits_j s_j = T_1 \cup T_2$ then for the PPD

a) $s \in S_1 \cup S_2 \to t \in T_1 \cup T_2$

b) $s \in S_1 \cup S_2 \to t \in T_1 \cap T_2$

c) $s \in S_1 \cap S_2 \to t \in T_1 \cup T_2$

d) $s \in S_1 \cap S_2 \to t \in T_1 \cap T_2$

e) $s \in S_1 \to t \in T_1 \cup T_2$

f) $s \in S_1 \to t \in T_1 \cap T_2$

g) $s \in S_2 \to t \in T_1 \cup T_2$

h) $s \in S_2 \to t \in T_1 \cap T_2$

i) $s \in S_1 \cup S_2 \to t \in T_1$

j) $s \in S_1 \cap S_2 \to t \in T_1$

k) $s \in S_1 \cup S_2 \to t \in T_2$

l) $s \in S_1 \cap S_2 \to t \in T_2$

all the parameters from formulas (5)-(11) can be calculated.

Proof:

$$\sigma_{x \in X}(r) = \sum_{z \in X} \sigma_{x=z}(r) \tag{16}$$

Let us demonstrate the proving of the most complex (the first) of the given consequences. Others can be provided similarly and are not presented here for brevity sake.

$$\begin{aligned}
&Sup\left(s \in S_1 \cup S_2 \to t \in T_1 \cup T_2\right) = \\
&= Sup\left(s \in S_1 \cup S_2 \wedge t \in T_1 \cup T_2\right) = \frac{\left|\sigma_{s \in S_1 \cup S_2 \wedge t \in T_1 \cup T_2}(r)\right|}{|r|} = \\
&= \frac{\left|\sigma_{s \in S_1 \wedge t \in T_1 \cup T_2}(r)\right| + \left|\sigma_{s \in S_2 \wedge t \in T_1 \cup T_2}(r)\right| - \left|\sigma_{s \in S_1 \cap S_2 \wedge t \in T_1 \cup T_2}(r)\right|}{|r|} = \\
&= \frac{1}{|r|}\left(\left|\sigma_{s \in S_1 \wedge t \in T_1}(r)\right| + \left|\sigma_{s \in S_1 \wedge t \in T_2}(r)\right| - \left|\sigma_{s \in S_1 \wedge t \in T_1 \cap T_2}(r)\right| + \right. \\
&+ \left|\sigma_{s \in S_2 \wedge t \in T_1}(r)\right| + \left|\sigma_{s \in S_2 \wedge t \in T_2}(r)\right| - \left|\sigma_{s \in S_2 \wedge t \in T_1 \cap T_2}(r)\right| - \\
&\left. - \left|\sigma_{s \in S_1 \cap S_2 \wedge t \in T_1}(r)\right| - \left|\sigma_{s \in S_1 \cap S_2 \wedge t \in T_2}(r)\right| + \left|\sigma_{s \in S_1 \cap S_2 \wedge t \in T_1 \cap T_2}(r)\right|\right)
\end{aligned} \tag{17}$$

$$Conf\left(s \in S_1 \cup S_2 \to t \in T_1 \cup T_2\right) = \frac{\left|\sigma_{s \in S_1 \cup S_2 \wedge t \in T_1 \cup T_2}(r)\right|}{\left|\sigma_{s \in S_1 \cup S_2}(r)\right|} = \frac{|r| \cdot Sup\left(s \in S_1 \cup S_2 \to t \in T_1 \cup T_2\right)}{\left|\sigma_{s \in S_1}(r)\right| + \left|\sigma_{s \in S_2}(r)\right| - \left|\sigma_{s \in S_1 \cap S_2}(r)\right|} \tag{18}$$

$$\begin{aligned}
&\mathrm{Imp}\left(s \in S_1 \cup S_2 \to t \in T_1 \cup T_2\right) = \\
&= \frac{Conf\left(s \in S_1 \cup S_2 \to t \in T_1 \cup T_2\right)}{Sup\left(t \in T_1 \cup T_2\right)} = \\
&= \frac{|r| \cdot Conf\left(s \in S_1 \cup S_2 \to t \in T_1 \cup T_2\right)}{\left|\sigma_{t \in T_1}(r)\right| + \left|\sigma_{t \in T_2}(r)\right| - \left|\sigma_{t \in T_1 \cap T_2}(r)\right|}
\end{aligned} \tag{19}$$

$$I_{s \in S_1 \cup S_2 \leftrightarrow t \in T_1 \cup T_2} = \sum_{x \in S_1 \cup S_2} \sum_{y \in T_1 \cup T_2} Sup(s = x \to t = y) \log_2 Imp(s = x \to t = y) \tag{20}$$

The complexity of calculating the Eqs.(16)–(20) directly and linearly depends on the power of the sets S_1 and S_2. In addition, there are data structures (Fibonacci heap, binomial heap [7]), allowing us to calculate the merging and the intersection of sets in a sublinear time.

The implementation of PPD derivation algorithms for the rule of aggregation can use subtypes and partial cases of the aggregation rule. For example, if $S_1 \cap S_2 = \varnothing$, Eqs.(16)-(20) become very simple and are calculated with the asymptotic complexity $O(1)$.

These rules of derivation (the given rule of PPD aggregation includes 12 subordinate rules, presented as consequences) are particularly effective for small sets of values, such as data from sociological and psychological interviews, weather observations, traffic studies and others.

PPD Derivation Rules Completeness

We shall consider an arbitrary PPD F_I: $\bigvee_{k=1}^{h} P_k^e \to \bigvee_{l=1}^{g} Q_l^e$. It can be received by the aggregation of the PPD $P_h^e \to Q_g^e$ з $\bigvee_{k=1}^{h-1} P_k^e \to \bigvee_{l=1}^{g-1} Q_l^e$ if $h > 1$ and $g > 1$, з $P_h^e \to \bigvee_{l=1}^{g-1} Q_l^e$, if $h = 1$ and $g > 1$, з $\bigvee_{k=1}^{h-1} P_k^e \to Q_g^e$, if $h > 1$ and $g = 1$.

Thus, we receive the factorization

$$\bigvee_{k=1}^{h} P_k^e \to \bigvee_{l=1}^{g} Q_l^e = \bigoplus_{i=0..\max(h-1,g-1)} \left(P^e_{\max(h-i,1)} \to Q^e_{\max(g-i,1)}\right) \tag{21}$$

From Eq.(1)

$$P_k^e = A_{i_1} \in \left\{a_{(i_1)(j_1)}\right\} \wedge A_{i_2} \in \left\{a_{(i_2)(j_2)}\right\} \wedge A_{i_k} \in \left\{a_{(i_k)(j_k)}\right\}$$

So, the parameters of an arbitrary PPD can be calculated with the help of statistics

$$\sigma_{x=x_i}\left(r\left(R\right)\right), \sigma_{x=x_i \wedge y=y_j}\left(r\left(R\right)\right),$$

$\sigma_{x=x_i \wedge y=y_j \wedge z=z_k}\left(r\left(R\right)\right)$ and so on. Using these means for presenting all PPD relations, which don't contain more than *k* parts of predicates conditions, we need

$$O\left(\max_{(i_1,i_2,\ldots,i_k) \in Z^k}\left(\left|class\left(A_{i_1}\right)\right| \cdot \left|class\left(A_{i_2}\right)\right| \cdot \ldots \cdot \left|class\left(A_{i_k}\right)\right|\right)\right)$$

of memory, where $Z \subset R$ is a set of attributes, among which dependencies are searched, $\left|class\left(A_{i_j}\right)\right|$ is a number of classification areas with the attribute A_{i_j}. The simplest variant is: $class\left(A_{i_j}\right) = dom\left(A_{i_j}\right)$, but for numerical and measurement data it is often convenient to divide them into

subordinate areas. This division increases the informational content and simplifies their search.

Keeping the full statistics with arbitrary deepness of enclosure is usually impossible due to limitations of available memory in a computing system, but in practice dependencies with more than 3-4 parts of the conditional predicate are not used. Accordingly, the presentation of all required data is quite possible even for quite large data sets.

Conclusions

This paper shows the results of research of building and merging aggregated associative dependencies. The investigated class of associative dependencies is widely used in data mining methods. The application is found in the branches of computer sciences, power engineering, physics, sociology etc.

As a result of the research the rules of expanding the full set of aggregated PPD from their certain set have been obtained. This fact allows storing only the minimal coverage of a data set with a selected class of dependencies, but not all the available dependencies in the context. This form of data presentation allows us to easily modify them (removing or adding processions, and also changing the attribute values of existing processions). The detection of this PPD property gives an important advantage over many other methods of static data analysis: data being changed, it is not necessary to recalculateall the data statistics, but only to update the required parameters.

The article provides the proof of effective calculation of such characteristics as a level of support, a level of confidence, a level of improvement and a full mutual information of PPD. However, there are more parameters which can be efficiently computed with the help of the studied properties and the developed rules of PPD derivation. During further research it is planned to examine the necessary conditions for the criteria of quality, so that they might be effectively calculated by applying the rules of PPD derivation.

The application of the PPD derivation rules described in this paper can reduce the required memory capacity of a computer system to

$$O\left(\max_{(i_1,i_2,\ldots,i_k)\in Z^k}\left(\left|class\left(A_{i_1}\right)\right|\cdot\left|class\left(A_{i_2}\right)\right|\cdot\ldots\cdot\left|class\left(A_{i_k}\right)\right|\right)\right),$$

where k is the maximum number of attributes that appears in the conditional and the resulting part of desired PPDs. Usually there is no need for values $k>3$, but it may occure for some specific values of fixed attributes, and in this case it becomes possible to keep separate statistics for these attributes.

Thus, the derivation rules described in this paper allow us to store and find efficiently the PPDs in large data sets, based on its minimal coverage.

References:

1. Chesnokov S. Determinational analysis of socio-economic data.– M.: Nauka. – 1982.– 168p. [Rus]

2. Titova O. Methods of construction and evaluation of aggregate associative rules in intelligent databases. – Kharkiv – 2006.

3. Main site of the Pathology Department, UT Southwestern Medical Center . [Electronic resource] http://pathcuric1.swmed.edu/pathdb/classifi.html

4. Description of a utility BiNGO, site of the Ghent University, [Electronic resource] http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html.

5. Official site of the National Institute of Allergy and Infectious Diseases (NIAID), NIH, [Electronic resource] http://david.abcc.ncifcrf.gov/content.jsp?file=/ease/ease1.htm &type=1

6. Norwig P., Thun S. Materials of online lecture “Machine Learning” of Stanford University – 2011. [Electronic resource] https://www.ai-class.com/course/video/quizquestion/97.

7. Llc Books, Heaps: Heapsort, Binary Heap, Smoothsort, Soft Heap, Fibonacci Heap, Treap, Binomial Heap, Pairing Heap, Leftist Tree, Skew Heap. Memphis, Tennessee, General Books LLC. – 2010. – 74p.

ВЛАСТИВОСТІ АСОЦІАТИВНИХ ЗАЛЕЖНОСТЕЙ У АНАЛІЗІ ДАНИХ

О. Пшеничний

У статті наведено результати дослідження властивостей асоціативних залежностей та можливостей їх ефективного агрегування. Розрозблено метод виявлення асоціативних залежностей широкого класу у великих наборах даних.

Oleksandr Pshenychnyi – M.Sc., graduated in 2009 from Lviv Polytechnic National University in speciality “Intelligent Decision Support Systems”. Currently is a post-graduate of Lviv Polytechnic National University, works as a software developer for Eleks Software Ltd.

Scientific interests area: data analysis methods, optimization and effective computation, parallel programming, grid- and cloud-based calculations, distributed data processing.