A DATA MINING METHOD FOR THE SOLUTION OF FLUID-FLOW PROBLEM

Dr. Hua Nam Son, PhD, associate professor
Budapest Business School
(Hua.Nam.Son@pszfb.bgf.hu)

Dr. Gubán Miklós, PhD, college professor
Dennis Gabor College
(guban@gdf.hu)

Abstract

Discovering market baskets is an attractive topic of data mining theory. In most studies the researchers considered the purchases or the transactions of the customers as a set of items. In other words, the transactions are analyzed on conceptual level. In some recent studies quantitative approach was proposed. The researchers are interested not only in the items or products chosen by the customers, but also in the quantity of each item or product in the transactions. This study is a continuation of a previous research. The transactions are characterized as a set of quantified items. This allows us to use algebraic methods as efficient tools to establish a formal description of market basket model. The advantage of this approach is that we can have better insight to the natural order between transactions and the role of network structure on them. In this general model the well known results of lattice theory are applied. The explicit representation of frequent transactions or market baskets, as well as of the association rules are tested by a program. The results of testing show that the algorithms finding frequent transactions and confident association rules are efficient and can be used to explore frequent customers' baskets, the association rules between the products, as well as the natural associations between different flows of the economic activities.

Keywords: market basket, frequent items, association rule, lattice, service.
JEL: C18, C51, O31

Introduction

Generally, during the discovery of critical points of the service processes we have to deal with a large amount of data. The larger masses of data, the number of participants in the process, and the significant variations of product volumes should be processed after the assessment. For these reasons data mining method would be used to solve different problems in data analysis and information assessment. The data mining method in fact can be used with definite objective to solve the problems in the following areas:

- segmentation of the market participants,
- determination of the organizational dependencies

- detection of the frequent products, customers, processes, as well as
- exploration of the association rules between frequent products, customers and processes.

The approach hereby may be applied to solve the problems in frequent fluid flows management. The data mining methods are needed to solve the problems of classification that are suitable for the analysis on conceptual level as pointed out as a goal of the research. Therefore the previously porposed and partly presented method is quite useful. In a recent paper we introduced the quantity-based approach and the quantity-based methods of generating frequent market baskets, associations and the quantity-based methods of classifications.

The mathematical model

First we present a mathematical model that is suitable for solving the problem. Let us consider a finite set of products $P = \{p_1, \dots, p_i, \dots, p_n\}$ that in fact could be products or services.

1. Definition

Market basket: (see []) By market baskets (for short, MB) we mean those $\alpha = (\alpha[1], \dots, \alpha[i], \dots, \alpha[n])$ vectors, by which $\alpha[i] \in \aleph$. $\alpha[i]$ cav be refered to as the quantity of p_i in the basket α. The set of all MBs is denoted by Ω.

For $\alpha, \beta \in \Omega$ where $\alpha = (\alpha[1], \alpha[2], \dots, \alpha[n])$, $\beta = (\beta[1], \beta[2], \dots, \beta[n])$ we write $\alpha \le \beta$ if for all $i = 1,2,\dots,n$ we have $\alpha[i] \le \beta[i]$. $\langle \Omega, \le \rangle$ is a lattice with the natural partial order $\le$. For a set $A \subseteq \Omega$ we denote

$U(A) = \{\alpha \in \Omega \mid \forall \beta \in A : \beta \le \alpha\}$ and

$L(A) = \{\alpha \in \Omega \mid \forall \beta \in A : \alpha \le \beta\}$.

We denote also

$sup(A) = \{\alpha \in U(A) \mid \nexists \beta \in U(A) : \beta < \alpha\}$ and

$inf(A) = \{\alpha \in L(A) \mid \nexists \beta \in L(A) : \alpha < \beta\}$.

One should remark that $sup(A)$ and $inf(A)$ are single elements of Ω, namely $sup(A) = u \in \Omega$, where $u[i] = max\{\alpha[i] \mid \alpha \in A\}$ and $inf(A) = v \in \Omega$, where $v[i] = min\{\alpha[i] \mid \alpha \in A\}$.

2. Definition

Support: For a set $A \subseteq \Omega$ and $\alpha \in \Omega$ we denote by

$$supp_A(\alpha) = \frac{|\{\beta \in A \mid \alpha \leq \beta\}|}{|A|}$$

the support of α in A. In other words, $supp_A(\alpha)$ denotes the rate of all market baskets that exceeds the given threshold α (in the form of a sample market basket) to the whole A. The support of a market basket is a statistical index and naturally, the market baskets of more support are of more significance and attract the attention of the managers, as well as of the researchers.

3. Definition

Frequent Market Baskets. For a set $A \subseteq \Omega$, $\alpha \in \Omega$ and $0 \leq \varepsilon \leq 1$ we can say that α is ε-frequent MB, if

$$supp_A(\alpha) \geq \varepsilon .$$

The set of all ε-frequent MBs is denoted by Φ_A^{ε}. We have:

Example: Consider a set of items $P = \{a,b,c\}$ and a set of transactions $A = \{\alpha, \beta, \gamma, \delta\}$, in which $\alpha = (2,1,0)$, $\beta = (1,1,1)$, $\gamma = (1,0,1)$, $\delta = (2,2,0)$. One can see that for $\sigma = (1,1,0)$, $\eta = (1,2,0)$ we have $supp_A(\sigma) = \frac{3}{4}$ and $supp_A(\eta) = \frac{1}{4}$. For the threshold $\varepsilon = \frac{1}{2}$ the ε-frequent MBs of A are:

$$\Phi_A^{\frac{1}{2}} = \{(2,1,0),(1,0,1),(1,1,0),(2,0,0),(0,0,1),(0,1,0),(1,0,0),(0,0,0)\}.$$

If P is the finite set of products or services and $A \subseteq \Omega$ is the set of those transactions required by customers, then $\Phi_A^{\frac{1}{2}}$ is the set of all transactions that the transactions of more than 50% of customers are „stronger". It is evident that the determination of Φ_A^{ε} is an important task in the economic activities. In logistics, for example, we can only make good, optimal plans for the material supply or for logistical scheduling if $\Phi_A^{\varepsilon}$is determined well. In the area of marketing, the market analysis and the customer management would be more effective if the frequent items are succesfully chosen from the masses of products, and the associations between them are detected.

The above mentioned approach, which is proposed and described in more detail in [5, 6], significantly differs from those methods that have been studied and applied in other researches. Here in this study instead of the elementary products or services $P = \{p_1, \dots, p_i, \dots, p_n\}$ All $\alpha = (\alpha[1], \dots, \alpha[i], \dots, \alpha[n])$ transactions are examined. The analysis is not done on the conceptual level (for examples, on „bread, egg, …", or „railway transportation, truck transportation), but on the quantitative level („1 kg bread, 10 piece of egg, …", or „100 km railway transportation, 50 km truck transportation). The quantitative approach points out the real features of transactions, the natural relationships between transactions, and therefore enables us to study more thoroughly the structure of the set of transactions. In many cases the quantitative approach appears to be more effective.

Let us denote

$$\Phi_{A,k} = \{\alpha \in \Omega \mid \exists \alpha_1, \alpha_2, ..., \alpha_k \in A : \alpha \leq \{\alpha_1, \alpha_2, ..., \alpha_k\}\}$$

One can remark that if $k \leq l$ then $\Phi_{A,k} \supseteq \Phi_{A,l}$ and $\Phi_A^\varepsilon = \Phi_{A,k}$ where $k = \lceil \varepsilon \mid A \mid \rceil$ denotes the smallest integer that is greater or equal to $\varepsilon \mid A \mid$.

We have:

Theorem 1: For a set of items $P = \{p_1, p_2, ..., p_n\}$, a set of MBs $A \subseteq \Omega$ and a threshold $0 \leq \varepsilon \leq 1$ an MB $\alpha \in \Omega$ is ε-frequent iff there exist $\alpha_1, \alpha_2, ..., \alpha_k \in A$ such that $\alpha \in L(\{\alpha_1, \alpha_2, ..., \alpha_k\})$ where $k = \lceil \varepsilon \mid A \mid \rceil$.

Proof: If $\alpha_1, \alpha_2, ..., \alpha_k \in A$, $k = \lceil \varepsilon \mid A \mid \rceil$ exists, where $\alpha \in L(\{\alpha_1, \alpha_2, ..., \alpha_k\})$ then $\alpha \leq \alpha_i$ for all $i = 1,2,...,k$, i.e. $supp_A(\alpha) = \frac{|\{\beta \in A \mid \alpha \leq \beta\}|}{|A|} \geq \frac{k}{|A|} \geq \varepsilon$.

Vice versa, if $supp_A \geq \varepsilon$ then $|\{\beta \in A \mid \alpha \leq \beta\}| \geq \varepsilon.|A|$, i.e. $\alpha_1, \alpha_2, ..., \alpha_k \in A$, $k = \lceil \varepsilon \mid A \mid \rceil$ exists, where $\alpha \in L(\{\alpha_1, \alpha_2, ..., \alpha_k\})$. The proof is completed.

As a consequence of Theorem 1 we have:

Theorem 2: (Explicit representation of *large MBs*) For a set of items $P = \{p_1, p_2, ..., p_n\}$, a set of MBs $A \subseteq \Omega$ and a threshold $0 \leq \varepsilon \leq 1$ there exist $\alpha_1, \alpha_2, ..., \alpha_s \in \Omega$ where $s = \binom{|A|}{\lceil \varepsilon |A| \rceil}$ such that

$$\Phi_A^\varepsilon = \bigcup_{i=1}^{s} L(\alpha_i).$$

Proof: Let $\alpha_1, \alpha_2, ..., \alpha_s$ be the set of all $inf\{\beta_1, \beta_2, ..., \beta_k\}$ where $k = \lceil \varepsilon \mid A \mid \rceil$ and $\beta_i \in A$. By Theorem 1 we have

$$\alpha \in \Phi_A^\varepsilon \Leftrightarrow \alpha \leq inf(\{\beta_1, \beta_2, ..., \beta_k\})$$

for some $\{\beta_1, \beta_2, ..., \beta_k\} \subseteq A$, where $k = \lceil \varepsilon \mid A \mid \rceil$. This implies that $\Phi_A^\varepsilon = \bigcup_{i=1}^{s} L(\alpha_i)$. The proof is completed.

We should remark that $\alpha_i \leq \alpha_j$ iff $L(\alpha_i) \subseteq L(\alpha_j)$. For a set of MBs A and a given threshold ε the set of MBs $\alpha_1, \alpha_2, ..., \alpha_s$ for which

i. $\Phi_A^\varepsilon = \bigcup_{i=1}^{s} L(\alpha_i)$

ii. $\forall i, j : 0 \leq i, j \leq s$ we have $\alpha_i \not\leq \alpha_j$ and $\alpha_j \not\leq \alpha_i$

is called by *basic ε - frequent set of MBs* of A. It is easy to verify that for a given A, ε the basic ε - frequent set of MBs of A is unique, which we denote by S_A^ε. Since the determination of Φ_A^ε (the set of all ε -frequent set of MBs in A) is important, it is interesting to determine its basic ε - frequent set of MBs S_A^ε. We have:

Theorem 3: For a set of items P, a threshold $0 \le \varepsilon \le 1$ every set of MBs $A \subseteq \Omega$ has an unique basic ε - frequent set of MBs S_A^ε.

In other words, for a set of items P, a threshold $0 \le \varepsilon \le 1$ and for every set of MBs $A \subseteq \Omega$ there is a system of MBs $S_A^\varepsilon = \{\alpha_1, \alpha_2, \dots, \alpha_s\}$, $\alpha_i \in \Omega$ such that

$$\Phi_A^\varepsilon = \cup_{i=1}^s L(\alpha_i)$$

and

$$\forall i, j: 0 \le i, j \le s \; \alpha_i \nleq \alpha_j \text{ és } \alpha_j \nleq \alpha_i$$

Let us denote

$$\alpha \cup \beta = sup\{\alpha, \beta\}.$$

Association rules and confidency

In the following the confidence of the association rules are defined in the quantitative approach. In cross marketing, store layout (see, for example, [1]) the discovery of association rules with a given confidence is an important problem that should be solved efficiently. Here we recall the results of some previous studies (see [5, 6]) that enable us, to represent explicitly all confident association rules. More precisely, we show here a method that describes all β for a set of products P, a set of MBs $A \subseteq \Omega$, a threshold $0 \le \varepsilon \le 1$ and α such that $\alpha \to \beta$ association is ε - confident.

4. Definition

Association rule: For $\alpha, \beta \in \Omega$ we call $\alpha \to \beta$ an *association rule* of β to α. By the *confidence* of $\alpha \to \beta$ in a set of MBs A we understand the rate

$$conf_A(\alpha \to \beta) = \frac{supp_A(\alpha \cup \beta)}{supp_A(\alpha)}$$

5. Definition

Confident association rule: For a set of items P, a set of MBs $A \subseteq \Omega$ and a threshold $0 \le \varepsilon \le 1$ an association $\alpha \to \beta$ is ε*-confident* if $conf_A(\alpha \to \beta) \ge \varepsilon$. The set of all ε - confident associations of A is denoted by C_A^ε.

One can verify that

$$conf_A(\alpha \to \beta) = \frac{|U(\alpha \cup \beta) \cap A|}{|U(\alpha) \cap A|}$$

so an association $\alpha \to \beta$ is ε-*confident* if and only if $\frac{|U(\alpha \cup \beta) \cap A|}{|U(\alpha) \cap A|} \geq \varepsilon$.

The algorithms used in the model

3.1. Algorithm. Creation of all ε-frequent MBs of a given set of transactions A

```
$\Phi_A^\varepsilon := \emptyset$
$k := [\varepsilon|A|]$
**for** $\{B \subset A \,\big|\, |B| = k\}$
    $\Phi_A^\varepsilon := \Phi_A^\varepsilon \cup L(B)$
**endfor**
```

The 3.1 algorithm creates all ε-frequent MBs of a given set of transactions A for a given threshhold ε.

3.2. Algorithm. Creation of the basic ε - frequent set of MBs S_A^ε.

```
$S_A^\varepsilon := \emptyset$
$k \coloneqq [\varepsilon|A|]$
for $\{B \subset A \,\big|\, |B| = k\}$
    $P$ := false
    $i$ := 1
    while $i \leq |S_A^\varepsilon|$ do
        if $S_A^\varepsilon[i] \leq \inf(B)$ then
            $P$ := true
            if $S_A^\varepsilon[i] \neq \inf(B)$ then
                $S_A^\varepsilon = S_A^\varepsilon / S_A^\varepsilon[i]$
                $S_A^\varepsilon = S_A^\varepsilon \cup \inf(B)$
            else
                $i$ := $i$+1
            endif
        else
            if $S_A^\varepsilon[i] \geq \inf(B)$ then
                $P$ := true
                $i$ := $i$+1
            endif
            else
                $i$ := $i$+1
        endif
        if not $P$ then
            $S_A^\varepsilon = S_A^\varepsilon \cup \inf(B)$
        endif
    enddo
endfor
```

The output is S_A^ε.

The 3.2 algorithm for a given threshhold ε creates all ε-frequent MBs $S_A^\varepsilon$of a given set of transactions A.

3.3. Algorithm. Creation of all ε - confident association rules $\alpha \to \beta$ for given α .

```
$C := U(\alpha) \cap A = \{\gamma \in A | \alpha \leq \gamma\}$
$s := [\varepsilon|C|]$
for $\{B \subset A \,||B| \geq s\}$ do
    $D := \emptyset$
    if $\alpha \leq inf(B)$ then
        $D := D \cup inf(B)$
    endif
endfor
$U := \emptyset$
for $\alpha \in D$ do
    $U := U \cup L(\alpha)$
endfor
```

The output is $\bigcup_{i=1}^{k} L(\alpha_i)$.

The 3.3 algorithm for a given threshold ε produces all β for which $\alpha \to \beta$ is ε-confident.

Testing

After collecting and registering, the data are processed by a database managing system and stored in a database. For testing and evaluating the above algorithms we have developed an **MB examination** program which operation is based on these algorithms. The program processes the selected data and produces useful information, such as the support of some transactions in a set of transactions, the set of frequent transactions and the set of basic frequent transactions, as well as the set of all confident association rules. These information are valuable for the managers in the decision making process.

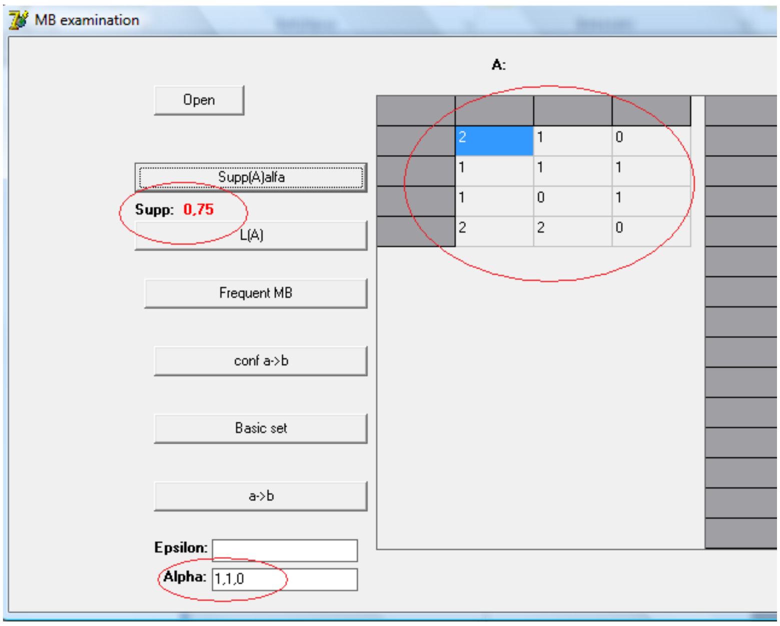
In the following by an example we show **MB examination** program's operation.

1. Example:

Let us consider a set of products $P = \{a, b, c\}$, a set of transactions $A = \{\alpha, \beta, \gamma, \delta\}$, where

$$\alpha = (2,1,0), \beta = (1,1,1), \gamma = (1,0,1), \delta = (2,2,0).$$

Suppose that $\sigma = (1,1,0), \eta = (1,2,0)$ are two MBs. The **MB examination** program gives as results $supp_A(\sigma)$=0,75; $supp_A(\eta)$=0,25, which we can see in the 1st Fig and 2nd Fig. The set of transactions and the input transaction are shown in the table A and in the field **Alpha,** respectively. The result is produced at **Supp.**



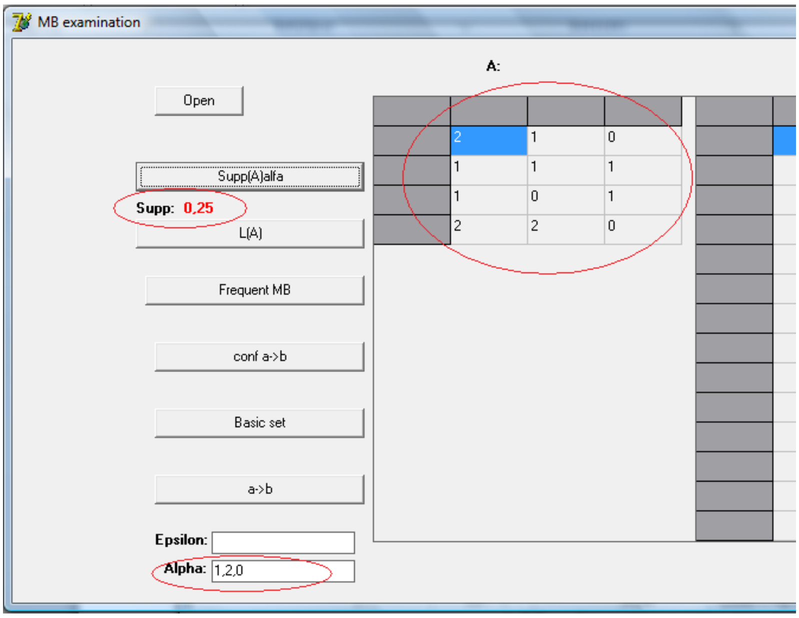
Supp: 0,75

2	1	0
1	1	1
1	0	1
2	2	0

Epsilon:

Alpha: 1,1,0

1. Fig. Computation of $supp_A(\sigma)$



Supp: 0,25

2	1	0
1	1	1
1	0	1
2	2	0

Epsilon:

Alpha: 1,2,0

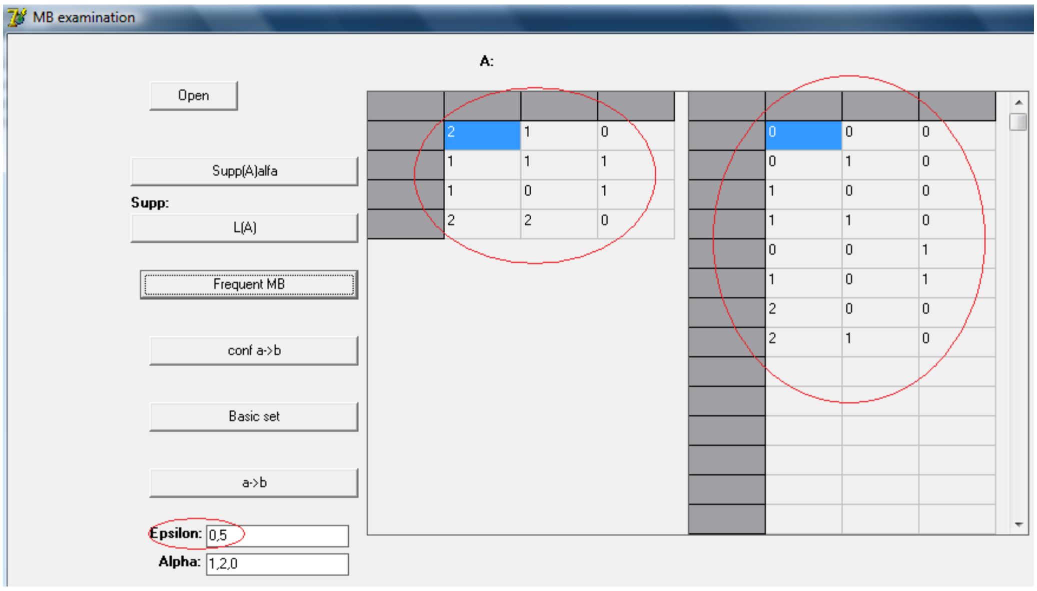
2. Fig. Computation of $supp_A(\eta)$

2. Example:

Let $\varepsilon = \frac{1}{2}$ be a threshold value. The set of all ε –frequent transactions of A could now be computed.

The set of transactions A and the threshold value is contained in the first column and in the **Epsilon** field in the Fig. 3, respectively. The result is snowen in the other column on the right. The **MB examination** program produces all ε –frequent transactions of A that are the following:

$$\Phi_A^{\frac{1}{2}} = \{(0,0,0),(0,1,0),(1,0,0),(1,1,0),(0,0,1),(1,0,1),(2,0,0),(2,1,0)\}$$



2	1	0
1	1	1
1	0	1
2	2	0

0	0	0
0	1	0
1	0	0
1	1	0
0	0	1
1	0	1
2	0	0
2	1	0

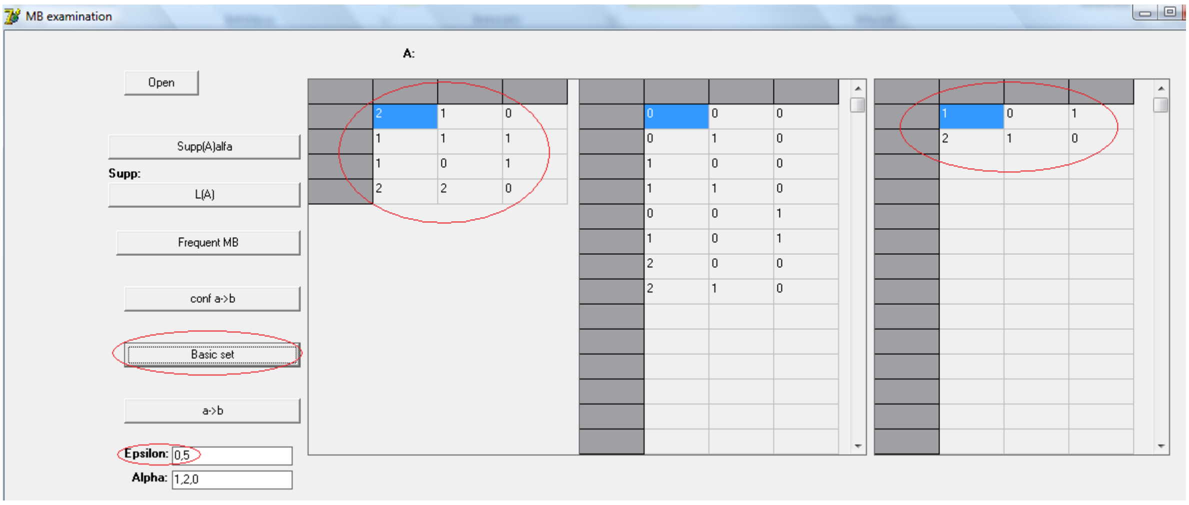
Epsilon: 0,5

Alpha: 1,2,0

3. Fig. Computation of $\Phi_A^{\frac{1}{2}}$

3. Example:

The set of all basic frequent transactions are computed in this example. The set of all input transactions can be found in the table on the left and the basic frequent transactions are snowen in the right table.



2	1	0
1	1	1
1	0	1
2	2	0

0	0	0
0	1	0
1	0	0
1	1	0
0	0	1
1	0	1
2	0	0
2	1	0

1	0	1
2	1	0

Epsilon: 0,5

Alpha: 1,2,0

4. Fig. Computation of basic frequent transactions

The basic frequent transactions are

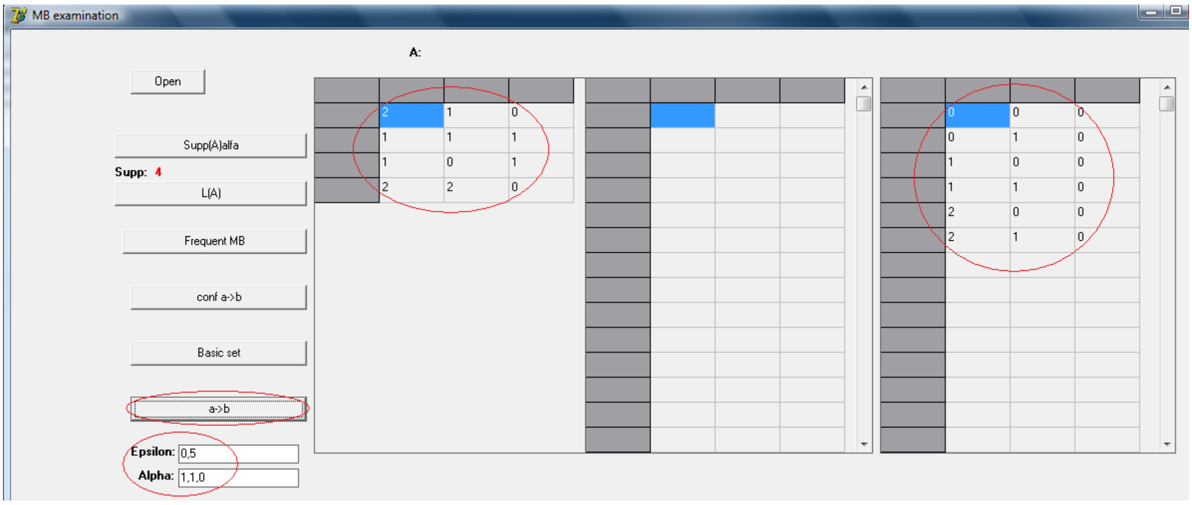
$$(1{,}0{,}1), (2{,}1{,}0),$$

This means that the set of all frequent transactions is

$$\Phi_A^{\frac{1}{2}} = L(1{,}0{,}1) \cup L(2{,}1{,}0).$$

4. Example:

Let A be the set of transactions as above. For a transaction $\sigma = (1{,}1{,}0)$ and $\varepsilon = \frac{1}{2}$ a threshold value let us find all η transaction for which $\sigma \to \eta$ is $\varepsilon -$confident association rules.



Supp: 4

Epsilon: 0,5

Alpha: 1,1,0

A:

2	1	0
1	1	1
1	0	1
2	2	0

0	0	0
0	1	0
1	0	0
1	1	0
2	0	0
2	1	0

5. Fig. Computation of $\sigma \to \eta$ $\varepsilon -$confident association rules.

The result is:

$$L(\alpha_1) \cup L(\alpha_2) = \{(0{,}0{,}0), (0{,}1{,}0), (1{,}0{,}0), (1{,}1{,}0), (2{,}0{,}0), (2{,}1{,}0)\}.$$

Effectiveness evaluation

In the case of large amount of data the complexity of the processing algorithms determines the effectiveness of the problem solving system. The complexity of above proposed algorithms are approximated in the following:

The complexity of 3.1 Algorithm is

$$O\left(\binom{|A|}{k} \cdot (m+1)^n\right) \approx |A|^k, \text{ where } k = [\varepsilon|A|].$$

The complexity of 3.2 Algorithm is

$$O\left(\binom{|A|}{k} \cdot m \cdot n\right).$$

and the complexity of 3.3 Algorithm is

$$O\left(\binom{|A|}{k} \cdot m \cdot n\right).$$

Conclusion

An appropriate mathematical model have been shown that could be used in data mining methods as an effective tool to solve the problems in different areas of economy, especially in data analysis, in optimization of logistical scheduling and in fluid flow analyzing. The results are essentially important for the managers in decision making. The theoretical results and the algorithms achieved in the previous studies are verified. The tests also prove that the approach is suitable for further theoretical researches and can be applied in various application areas of economy and production management.

References

1. R. Agrawal, R. Srikant: (1994) Fast algorithms for mining association rules. VLDB, 487-499.

2. Benczúr A. , Szabó Gy. I.: (2012) Functional Dependencies on Extended Relations Defined by Regular Landuages, Foundations of Information and Knowledge Systems, eds.: T. Lukasiewicz, A. Sali, FoIKS 2012 Kiel, Germany}, {LNCS 7153, pp. 384-404.

3. T. Brüggermann, P. Hedström, M. Josefsson: (2004) Data mining and Data Based Direct Marketing Activities, Book on Demand GmbH, Norderstedt, Germany.

4. Davey, B.A., Priestley, H. A.: (2002) Introduction to Lattices and Order, Cambridge University Press, ISBN 978-0-521-78451-1.

5. J. Demetrovics, Hua Nam Son, Akos Guban: (2011) An algebraic approach to market basket model: explicit representation of frequent market baskets and associations rules, CSIT 2011. Computer science and information technologies. Proceedings of the conference. Yerevan, pp. 170-173.

6. J. Demetrovics, Hua Nam Son, Akos Guban: (2011) An algebraic representation of frequent market baskets and association rules, Cybernetics and Information Technologies, Vol. 11, No 2, pp. 24-31.

7. Heikki Mannila, Hannu Toivonen: (1996) Discovering generalized episodes using minimal occurrences. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD' 96), AAAI Press, pp. 146 - 151.

8. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal: (1999) Discovering frequent closed itemsets for association rules. ICDT, pp. 398-416.

9. Ping-Yu Hsu, Yen-Liang Chen, Chun-Ching Ling: (2004) Algorithms for mining association rules in bag databases, Information Sciences, Volume 166, Issues 1-4, pp. 31-47.