DISCOVERING BINARY AND FUZZY ASSOCIATION RULES FROM DATABASE

Hoang Kiem - Do Phuc
College of Natural Sciences

(Received on June 10^{th} 1999)

ABSTRACT *: In this paper, we propose a method for discovering the binary and fuzzy association rules from database. We study a new kind of association rules: fuzzy association rules. We use the rough set theory of Z.Pawlak and fuzzy set theory of L.A. Zadeh for defining some concepts and developing our method based on these new concepts. We also introduce an application for discovering the fuzzy rules from a database in supermarket. In this application, we study the quantity of purchased items instead of 'to be purchased or-not to be purchased' information of a particular item as the traditional approach did.*

Keywords: binary association rules, binary descriptor vector, binary information mapping, binary information system, binary large descriptor set, fuzzy association rules, fuzzy descriptor vector, fuzzy information mapping, fuzzy information system, fuzzy large descriptor set.

1. INTRODUCTION

Problem of association rule discovery is one of the key problems in knowledge discovery from large database. There are many research works related to this problem [1][2][9][12]. In these works, the association rules have the format "There are n% of customers who buy book, buy video tape". We propose a method that can help us to discover the fuzzy association rules such as "There are m% of customers who buy MANY books, buy FEW video tape". We employ Z.Pawlak's idea in rough set theory [14] for defining the binary information system, and some concepts on binary information system. After that, we employ L.A. Zadeh's idea in fuzzy set theory [11] for changing the above concepts to fuzzy concepts. These new definitions are used for discovering the binary large descriptor set, fuzzy large descriptor set, binary association rules and fuzzy association rules. We propose a method for discovering these things from database. Our method works with the information system in the memory instead of the database in disk, so the efficiency of mining process is improved significantly. We also discuss a real life application in discovering the binary and fuzzy association rules from a supermarket database. This paper is organized as follows: 1) Introduction; 2) Formal definitions on binary information system 3) An algorithm for discovering binary large descriptor sets and binary association rules from binary information system 4) Formal definitions on fuzzy information system 5) An application to a database in supermarket 6) Upgrading the proposed algorithm for discovering fuzzy association rules 7) Using a cube database for discovering the relations among many fuzzy information systems 8) Conclusions and future works.

2. FORMAL DEFINITIONS ON BINARY INFORMATION SYSTEM

2.1 Binary information system

Let $O=\{o_1,...,o_n\}$ be a finite set of n objects, $D=\{d_1,...,d_m\}$ be a finite set of m descriptors, $B =\{0,1\}$. A binary information system is defined as $S_B = (O,D,B,\chi)$ where χ is a mapping χ: OxD--- > B, χ (o,d)=1 if object o has descriptor d and χ (o,d)=0 if not.

2.2. Binary information mapping

Given a binary information system $S_B = (O,D,B,\chi)$. Let P(O) be the power set of O, let P(D) be the power set of D, binary information mappings ρ_B and λ_B are defined as follows:

$$\rho_B : P(D) \text{ ------ } > P(O) \text{ and } \lambda_B : P(O) \text{ ---- } > P(D) \quad (1)$$

- Given $S \subset D$, $\rho_B (S) = \{ o \in O \mid \forall d \in S , \chi(o,d) = 1 \}$
- Given $X \subset O$, $\lambda_B(X) = \{ d \in D \mid \forall o \in X , \chi(o,d) = 1 \}$

Some properties of $\rho_B(S)$ and $\lambda_B(X)$ can be stated as:

i) $\forall$ S1,S2 $\in$ P(D), S1 $\subseteq$ S2 $\Rightarrow$ $\rho_B(S2) \subseteq \rho_B(S1)$

ii) $\forall$ X1,X2 $\in$ P(O), X1 $\subseteq$ X2 $\Rightarrow$ $\lambda_B(X2) \subseteq \lambda_B(X1)$

2.3. Binary large descriptor set

Given a binary information system $S_B = (O,D,B,\chi)$ and a threshold $\upsilon \in [0,1]$. Let S be a subset of D, S is a binary large descriptor set with threshold υ if:

$$\text{Card}(\rho_B(S)) >= \upsilon * \text{Card}(O) \quad (2)$$

Let L_B be a set of all binary large descriptor sets discovered from S_B, we have the following property:

i) $\forall S \in L_B , T \subset S \Rightarrow T \in L_B$

We denote $L_{B,h}$ as a subset of L_B, if $X \in L_{B,h}$, card(X) = h (h is a positive integer).

2.4. Binary large association rules and confidence factor

Given a binary information system $S_B = (O,D,B,\chi)$ and a threshold $\upsilon \in [0,1]$. Let L be an element of L_B , X and Y be the subsets of L where $L= X \cup Y$ and $X \cap Y=\{\}$.We define a binary association rule between descriptor set X and Y as an information mapping : X --- >Y. The confidence factor of this rule is denoted as CF_B (X --- >Y) and calculated by:

$$CF_B (X \text{ --- } >Y) = \text{Card}(\rho_B (X) \cap \rho_B (Y)) / \text{Card}(\rho_B (X)) \quad (3)$$

We denote $R_{B,\beta}$ as the set of all binary large association rules which are discovered from S_B where $CF_B(r)>=\beta$, $\forall r \in R_{B,\beta}$.

2.5. Binary descriptor vectors and some operations on these vectors

Given a binary information system $S_B = (O,D,B,\chi)$ where $O=\{o_1,...,o_n\}$ is a finite set of n objects and $D=\{d_1,...,d_m\}$ is a finite set of m descriptors, we define some concepts related to our algorithm as follows:

2.5.1 Binary descriptor vector

A binary descriptor vector $v_B(X)=(X_1,...,X_n)$ where $X \subset D$ is a vector with n components. Each component X_i takes a value in B. Let VS_B be a set of all binary descriptor vectors of S_B. If card(X)=1, X is a descriptor of S_B and $X_j = \chi(o_j,X)$. On V_B, we define two operations as:

2.5.2 Binary descriptor vector product

Given $X_1, X_2 \subset D$, let $v_B(X_1)=(X_{11},...,X_{1n})$ and $v_B(d_2)=(X_{21},...,X_{2n})$ be the elements of VS_B. The binary descriptor vector product of $v_B(X_1)$ and $v_B(X_2)$ is denoted as:

$$v_B(X_3)=v_B(X_1)\ \Theta_B\ v_B(X_2) \qquad (4)$$

where $v_B(X_3)=(X_{31},...,X_{3n})$, $X_{3j} = \min(X_{1j,}\ X_{2j})$, j=1,...,n, $X_3 = X_1 \cup X_2 \subset D$. From vector $v_B(X_3)$, we know all objects having descriptor set X_1 and X_2. We use $v_B(X_1)$ for representing $\rho_B(Xd_1)$; $v_B(X_2)$ for representing $\rho_B(X_2)$; and $v_B(X_3)$ for representing $\rho_B(X_3)$.

2.5.3. Support of the binary descriptor vectors

Given $X_1 \subset D$, a support of $v_B(X_1)$ denoted as $sup_B(v_B(X_1))$ is defined as:

$$sup_B(v_B(X_1)) = \{\ o \in O \mid \forall d \in X_1,\ \chi(o,d) = 1\ \} \qquad (5)$$

It is easy to see $card(sup_B(v_B(X_1))) = card(\rho_B(X_1))$

2.5.4. Calculating card(ρ_B(S))

Let $S=\{s_1,...,s_k\}$ be a subset of D where s_j is a descriptor of S_B, j =1,...,k.. Each s_j corresponds to a binary descriptor vector $v_B(\{s_j\})$ of VS_B. The cardinality of $\rho_B(S)$ is calculated by:

$$card(\rho_B(S)) = card(sup_B(v_B(\{s_1\})\ \Theta_B\ ....\ \Theta_B\ v_B(\{s_k\}))) \qquad (6)$$

We denote $VS_{B,h}$ as a subset of VS_B containing only vector $v_B(X)$ where $X \subset D$ and card(X) = h, h is a given threshold.

3. AN ALGORITHM FOR DISCOVERING BINARY LARGE DESCRIPTOR SETS AND BINARY ASSOCIATION RULES FROM BINARY INFORMATION SYSTEM

In this section, we develop the idea of Apriori-Tid algorithm presented in [12] and propose an algorithm for discovering binary large descriptor sets and binary association rules from binary information system. Our proposed algorithm works with bits in memory and does not work with the database on disk, so we can improve the speed of mining rule process. Given a database and two threshold MINSUPP, MINCONF for the min support and min confidence of the association rule in data mining literature, the Apriori-Tid algorithm has two phases:

Phase 1: Discover the large descriptor sets based on a given MINSUP threshold.

Phase 2: Build up the association rules based on a given MINCONF threshold.

Given a binary information matrix S_B = (O,D,B,χ) where O={o_1,...,o_n} is a finite set of n objects, and D={d_1,...,d_m} is a finite set of m descriptors, B={0,1} and thresholds υ, $\beta \in [0,1]$ where υ is MINSUPP and β is the MINCONF. Our proposed algorithm is as follows:

3.1. Discover the binary large descriptor set L_B

```
Answer ={ }
Generate $L_{B,1}$ from $S_B$;
For ( k=2; $L_{B,k}$ <> { } ; k++) do
    Generate $L_{B,k}$ from $L_{B,k-1}$
Answer = $\cup_k L_{B,k}$
Return Answer;
```

3.1.1 Generating $L_{B,1}$

```
$L_{B,1}$ = { }
For ( i=1; i <= m; i++) do
    If (card(sup$_B$(v$_B$({$d_i$}))) > $\upsilon$ * card(O) then begin
        SaveLargeSet({ $d_i$ }, $L_{B,1}$);
        SaveDescriptorVector(v$_B$({$d_i$ }, $VS_{B,1}$);
    End;
Answer = $L_{B,1}$
Return Answer;
```

where m = card(D).

3.1.2. Create $L_{B,k}$ from $L_{B,k-1}$

Based on the property $\forall$ S $\in L_B$, T$\subset$ S $\Rightarrow$ T$\in L_B$, we generate $L_{B,k}$ from $L_{B,k-1}$. The procedure is as follows:

```
Create a matrix which rows and columns are the elements of $L_{B,k-1}$.
$L_{B,k}$ = { }
For(each X $\in L_{B,k-1}$) do
    For (each Y $\in L_{B,k-1}$ and X <> Y) do begin
        T = X$\cup$Y;
        If (card(sup$_B$(v$_B$(T)) > $\upsilon$*card(O) ) and Card(T) = k ) then begin
            SaveLargeSet(T, $L_{B,k}$);
            SaveDescriptorVector(v$_B$(T), $VS_{B,k}$);
        end;
```

```
end;
Answer = $L_{B,k}$;
Return Answer;
```

SaveLargeSet(T, $L_{B,k}$) is a function for saving a binary large descriptor set T to $L_{B,k}$.

SaveDescriptorVector(v_B(T), $VS_{B,k}$) is a function for saving a binary descriptor vector v_B(T)to $VS_{B,k}$. Based on (4), we can calculate rapidly $sup_B(v_B(T))$ at k^{th} step of the above loop from the elements of $VS_{B,k-1}$.

3.2. Discover the binary association rules from $R_{B,\beta}$

```
$R_{B,\beta}$ = { };
For ( each L $\in L_B$ ) do begin
    For ( each X,Y $\in$ L and $X \cup Y$ = L and $X \cap Y$ ={ } ) do begin
        if ( $CF_B$ (X --- >Y) >= $\beta$ ) then SaveRule(X --- >Y, $R_{B,\beta}$);
        if ( $CF_B$ (Y --- >X) >= $\beta$ ) then SaveRule(Y --- >X, $R_{B,\beta}$);
    End
End
Answer = $R_{B,\beta}$;
```

SaveRule(X --- >Y, $R_{B,\beta}$) is a function for saving the binary association rule to $R_{B,\beta}$.

3.3. An example

Given a binary information system S_B = (O,D,B,χ) as defined in table 1 and υ = 0.5, β=0.5. From S_B, we have the following binary descriptor vectors of V_B.

$v_B(\{d_1\}) = (1,0,1,0)$; $v_B(\{d_2\}) = (0,1,1,1)$;

$v_B(\{d_3\}) = (1,1,1,0)$; $v_B(\{d_4\}) = (1,0,0,0)$;

$v_B(\{d_5\}) = (0,1,1,1)$.

Table 1: A binary information system

	d_1	d_2	d_3	d_4	d_5
o_1	1	0	1	1	0
o_2	0	1	1	0	1
o_3	1	1	1	0	1
o_4	0	1	0	0	1

3.3.1. Creating $L_{B,1}$

Calculate card($sup_B(v_B(\{d_i\})$), i = 1,...,5, we have:

card($sup_B(v_B(\{d_1\})$) = 2; card($sup_B(v_B(\{d_2\})$) = 3;

card($sup_B(v_B(\{d_3\})$) = 3; card($sup_B(v_B(\{d_4\})$) = 1; card($sup_B(v_B(\{d_5\})$) = 3.

With MINCONF=0.5, we have $L_{B,1}$= { {d_1}, {d_2}, {d_3}, {d_5}}.

In $VS_{B,1}$, we have the following descriptor vectors:

$v_B(\{d_1\}) = (1,0,1,0)$; $v_B(\{d_2\}) = (0,1,1,1)$; $v_B(\{d_3\}) = (1,1,1,0)$; $v_B(\{d_5\}) = (0,1,1,1)$.

3.3.2. Creating $L_{B,2}$ from $L_{B,1}$

From $L_{B,1}$, we build a matrix expressing the mapping f: $L_{B,1}$x $L_{B,1}$---- > R (real number set).

	{d_1}	{d_2}	{d_3}	{d_5}
{d_1}	2	1	2	1
{d_2}	1	2	2	3
{d_3}	2	2	2	2
{d_5}	1	3	2	3

Given $(X,Y)\in L_{B,1}xL_{B,1}$ and X<>Y, T=X$\cup$Y, the value of f(X,Y) is card($sup_B(v_B(T))$). We select T=X$\cup$Y which card($sup_B(v_B(X) \Theta_B v_B(Y))$) = card($sup_B(v_B(T))$ >= MINSUP* card(O) and card(T)=2. When we calculate $sup_B(v_B(T))$, T=X$\cup$Y we use $v_B(X)$ which was saved in $VS_{B,1}$ for increasing the speed of calculation. Finally, we have $L_{B,2}$ as follows:

$L_{B,2}$ = { {d_1 ,d_3}, {d_2,d_3}, {d_2,d_5} , {d_3,d_5}}.

In $VS_{B,2}$, we have the following descriptor vectors:

$v_B(\{d_1,d_3\}) = (1,0,1,0)$; $v_B(\{d_2,d_3\}) =(0,1,1,0)$;

$v_B(\{d_2,d_5\}) = (0,1,1,1)$; $v_B(\{d_3,d_5\}) =(0,1,1,0)$

3.3.3 Creating $L_{B,3}$ from $L_{B,2}$

From $L_{B,2}$, we build matrix expressing the mapping f: $L_{B,2}$x $L_{B,2}$---- > R (real number set).

	{d_1 ,d_3}	{d_2,d_3}	{d_2,d_5}	{d_3,d_5}
{d_1 ,d_3}	2	1	1	1
{d_2,d_3}	1	2	2	2
{d_2,d_5}	1	2	3	2
{d_3,d_5}	1	2	2	2

Given $(X,Y)\in L_{B,2}xL_{B,2}$ and X<>Y, T=X$\cup$Y, the value of f(X,Y) is card($sup_B(v_B(T))$). We select T=X$\cup$Y which card($sup_B(v_B(X) \Theta_B v_B(Y))$) = card($sup_B(v_B(T))$ >= MINSUP* card(O) and card(T)=3. When we calculate $sup_B(v_B(T))$, T=X$\cup$Y we use $v_B(X)$ which was saved in $VS_{B,2}$ for increasing the speed of calculation. Finally, we have $L_{B,3}$ ={ {d_2 ,d_3, d_5}} and $VS_{B,3}$ containing only one descriptor vector $v_B(\{d_2,d_3 , d_5\}) = (0,1,1,0)$

3.3.4 Creating $L_{B,4}$ from $L_{B,3}$

From $L_{B,3}$, we build matrix expressing the mapping f: $L_{B,3}$x $L_{B,3}$---- > R (real number set).

	$\{d_2,d_3,d_5\}$
$\{d_2,d_3,d_5\}$	2

We have $L_{B,4}$ = { }. Stop the procedure.

Finally, we have:

$L_B = L_{B,1} \cup L_{B,2} \cup L_{B,3} = \{\{d_1\},\{d_2\}, \{d_3\}, \{d_5\}, \{d_1,d_3\}, \{d_2,d_3\},\{d_2,d_5\}, \{d_3,d_5\},\{d_1,d_3,d_5\}\}$

If MINCONF = 50%, we may have some binary association rules as follows:

$\{d_1\}$ ---- > $\{d_3\}$, $\{d_3\}$ ---- > $\{d_1\}$, $\{d_3,d_5\}$--- > $\{d_2\}$

4. FORMAL DEFINITIONS ON FUZZY INFORMATION SYSTEM

4.1 Fuzzy information system

Let $O=\{o_1,\ldots,o_n\}$ be a finite set of n objects and $D=\{d_1,\ldots,d_m\}$ be a finite set of m descriptors, let F=[0,1] be a subset of real number set. A fuzzy information system is a $S_F = (O,D,F,\mu)$. Mapping μ is defined as μ: OxD--- > F, where $\mu(o,d) \in F$ expresses the degree of object o having descriptor d..

4.2. Fuzzy information mappings

Given a fuzzy information system $S_F = (O,D,B,\mu)$ and a threshold $\tau \in F$, we define the fuzzy information mappings ρ_F and λ_F as follows:

ρ_F: P(D) ------ > P(O) and λ_F: P(O) ---- > P(D)

- Given $S \subset D$, $\rho_F(S) = \{ o \in O \mid \forall d \in S , \mu(o,d) >= \tau \}$
- Given $X \subset O$, $\lambda_F(X) = \{ d \in D \mid \forall o \in X , \mu(o,d) >= \tau \}$

Some properties of $\rho_F(S)$ and $\lambda_F(X)$ can be stated as follows:

i) $\forall$ S1,S2 $\in$ P(D), S1 $\subseteq$ S2 $\Rightarrow$ $\rho_F(S2) \subseteq \rho_F(S1)$

ii) $\forall$ X1,X2 $\in$ P(O), X1 $\subseteq$ X2 $\Rightarrow$ $\lambda_F(X2) \subseteq \lambda_F(X1)$

4.3. Fuzzy large descriptor sets

Given a fuzzy information matrix $S_F = (O,D,F,\mu)$ and a threshold $\upsilon \in T$. Let $S \subset D$, S is a fuzzy large descriptor set with threshold υ if:

$$Card(\rho_F(S)) >= \upsilon * Card(O) \qquad (7)$$

Let L_F be a set of all fuzzy large set discovered from S_F, we have the following property:

i) $\forall S \in L_F, T \subset S \Rightarrow T \in L_F$

We denote $L_{F,h}$ as a subset of $L_{F,}$ if $X \in L_{F,h}$, card(X) = h (h is an positive integer)..

4.4. Fuzzy association rules and confidence factor

Given a fuzzy information matrix $S_F = (O,D,F,\mu)$ and a threshold $\upsilon \in T$. Let S be an element of L_F , X and Y be the subsets of S where $S= X \cup Y$ and $X \cap Y = \{\}$.

We define a fuzzy association rule between descriptor sets X and Y and denoted as X--- >Y. The confidence factor of this rule is calculated by:

$$CF_F (X\text{---}>Y) = Card(\rho_F (X) \cap \rho_F (Y)) / Card(\rho_F (X)) \qquad (8)$$

We denote $R_{F,\beta}$ as the set of all fuzzy large association rules r discovered from S_F where $CF_F(r)>=\beta$.

4.5. Fuzzy descriptor vectors and some operations on these vectors

Given a fuzzy information system $S_F = (O,D,F,\mu)$ where $O=\{o_1,...,o_n\}$ is a finite set of n objects and $D=\{d_1,...,d_m\}$ is a finite set of m descriptors, we define some concepts related to our algorithm.

4.5.1 Fuzzy descriptor vectors

Let X be a subset of D, we define a fuzzy descriptor vector $v_F(X)$ for representing X. A fuzzy descriptor vector $v_F(X)=(X_1,...,X_n)$ is a vector with n components, each component X_i takes a value in F. Let VS_F be a set of all fuzzy descriptor vectors of S_F. If card(X)=1, X is an element of D, where $X_j=\mu(o_j, X)$. On $VS_{F,}$ we define two operations as follows:

4.5.2. Fuzzy vector product

Let $v_F(d_1)=(d_{11}, . . . , d_{1n})$ and $v_F(d_2) =(d_{21}, . . . , d_{2n})$ be the elements of VS_F. The fuzzy vector product of $v_F(d_1)$ and $v_F(d_2)$ is a vector $v_F(d_3)=(d_{31}, . . . , d_{3n})$ of VS_F where $d_{3j} = \min(d_{1j,}\ d_{2j})$, j = 1,...,n. This fuzzy vector product is denoted as $v_F(d_3) = v_F(d_1) \quad \Theta_F$ $v_F(d_2)$.From vector $v_F(d_3)$, we can know all objects having descriptor d_1 greater than a given threshold. We use $v_F(d_1)$ for representing $\rho_B(\{d_1\})$; $v_F(d_2)$ for representing $\rho_B(\{d_2\})$; and $v_F(d_3)$ for representing $\rho_B(\{d_1, d_2\})$. It is easy to discover this property:

Let Z be a subset of D and $Z=X\cup Y$ and $X\cap Y=\{\}$. If we know $v_F(X)$ and $v_F(Y)$, v_F (Z) can be calculated by $v_F(Z)=v_F(X) \ \Theta_F \ v_F(Y)$.

4.5.3. Support of the fuzzy descriptor vector

Let $v(d_1)=(d_{11}, . . . , d_{1n})$ be a element of VS_F and a given threshold $\alpha \in T$. A support of fuzzy vector $v(d_1)$ is defined as:

$$sup_F(v_F(d_1)) = \{ o_j \in O : \mu(o_j,d_1) >= \alpha \} \text{, for all } j =1,...,n. \qquad (9)$$

where card($sup_F(v_F(d_1))$)) is the number of objects o having $\mu(o_j,d_1) >=\alpha$.

4.5.4. Calculating cardinality of $\rho_F(S)$

Let $S=\{s_1,...,s_k\}$ be a subset of D, where s_j , j =1,...,k is an element of D. Each s_j corresponds to a vector $v(s_j)$ of V_F. The cardinality of $\rho_F(S)$ is calculated by:

$$card(\rho_F(S)) = card(sup_F(v_F(d))) \qquad (10)$$

where $v_F(d) = v_F(s_1)\ \Theta_F \ldots \Theta_F\ v_F(s_k)$.

We denote $VS_{F,h}$ as a subset of VS_F containing only vector $v_F(d)$ where $d \in P(D)$ and card(d) = h.

5. AN APPLICATION TO A DATABASE IN SUPERMARKET

Suppose that, we have a table where each row is a transaction and each column is an item. A typical table of this kind is shown in table 2.

Table 2: Table of transaction and items

	d1	d2	d3	d4	d5	d6
o1	9	30	12	0	0	0
o2	22	56	15	0	0	0
o3	14	23	17	41	21	31
o4	17	45	14	61	15	16
o5	0	0	1	71	18	19
o6	0	0	1	31	15	15

The element $b_{i,j}$ of this table is the quantity of the purchased item d_j in transaction o_i. Let O be the set of transaction and D be the set of items. We define a mapping χ as χ: OxD --- > { 0, 1 } where $\chi(o,d)=1$ if transaction o has item d. From table 2, we build up χ as shown in table 3.

With threshold $\upsilon = 0.6$ (60%), set {d1,d2,d3} and {d4,d5,d6} are binary large descriptor sets and binary association rule such as:{d1,d2} --- > {d3},

Table 3: Binary information matrix

	d1	d2	d3	d4	d5	d6
o1	1	1	1	0	0	0
o2	1	1	1	0	0	0
o3	1	1	1	1	1	1
o4	1	1	1	1	1	1
o5	0	0	1	1	1	1
o6	0	0	1	1	1	1

In binary information matrix, each $\chi(o,d)$ is equal to 1 or 0. This representation does not consider the quantity of purchased items in transaction because if transaction o1 has 12 items of d3 and transaction o5 has only 1 items of d3 but $\chi(o1,d3)=\chi(o5,d3) = 1$. If d3 is an item that has high price, the selling of a small quantity of d3 and large quantity of d3 has big difference. We use three fuzzy sets MANY, AVERAGE, FEW [7,8] for fuzzifying each item. The membership functions of these fuzzy sets are appropriated to item d are μ^{d}_{MANY}, μ^{d}_{AVER}; μ^{d}_{FEW}. Let $\mu^{d1}_{MANY,\ldots,}$ μ^{dn}_{MANY} be n membership functions of n fuzzy sets MANY appropriated to n items. We fuzzify the information table in table 2 and the result is shown in table 4.

Table 4: A fuzzy information matrix with n membership function μ_{MANY}.

	d1	d2	d3	d4	d5	d6
o1	0.8	0.9	0.8	0	0	0
o2	0.8	0.8	0.8	0	0	0
o3	0.9	0.8	0.7	0.6	0.6	0.3
o4	0.6	0.6	0.9	0.6	0.7	0.5
o5	0	0	0.1	0.7	0.8	0.4
o6	0	0	0.1	0.6	0.7	0.5

where μ_{MANY}: OxD --- > [0,1].With $\upsilon = 0.5(50\%)$; $\tau = 0.6(60\%)$, we can discover from the fuzzy information matrix the following large descriptor sets for the concept MANY of n items as {d1,d2,d3},{d4, d5}. Set { d1,d2,d3} means the fuzzy association rule:{d1,d2} --- > {d3}, the $CF_F(*)$ of this rule is equal to 1 or 100%. It means that There are 100% customers who buy MANY {d1,d2}, buy MANY {d3}.

6. UPGRADING THE PROPOSED ALGORITHM FOR DISCOVERING FUZZY ASSOCIATION RULES

Given a fuzzy information matrix $S_F = (O,D,B,\mu)$ where $O=\{o_1,...,o_n\}$ is a finite set of n objects, and $D=\{d_1,...,d_m\}$ is a finite set of m descriptors, $B=\{0,1\}$ and thresholds υ, $\beta\in[0,1]$ where υ is MINSUPP and β is the MINCONF. Let L_F is a set of all large fuzzy descriptor sets of S_F, let $L_{F,k}$ be a subset of L_F, if $X \in L_{F,k}$, card(X) = k (k is an positive integer) and $R_{F,\beta}$ be a set of all fuzzy association rules which have confidence greater than a given threshold β. The algorithm for discovering the fuzzy association rules is the same as the procedure mentioned in 3 for discovering binary association rules. In this algorithm, S_B, L_B $L_{B,k}$ $R_{B,\beta}$ are replaced by S_F, L_F $L_{F,k}$ $R_{F,\beta}$ respectively.

7. USING A CUBE DATABASE FOR DISCOVERING THE RELATION AMONG MANY FUZZY INFORMATION MATRICES

With (O,D) and three sets of $\{\mu^d_{MANY}\}$,$\{\mu^d_{AVER}\}$,$\{\mu^d_{FEW}\}$ we have three fuzzy information matrices which are built up from the same (O,D). We build a cube database where dimension #1 is the transaction; dimension #2 is the items and dimension #3 is the membership functions μ^d_{MANY} , μ^d_{AVER} , μ^d_{FEW}. In each dimension, we discover the fuzzy large sets. Let $F1= (O,D, \mu^d_{MANY}, \upsilon , \tau)$ is the first layer of dimension #3. $F2= (O,D, \mu^d_{AVER}, \upsilon , \tau)$ is the second layer of dimension #3 and $F3=(O,D, \mu^d_{FEW}, \upsilon , \tau)$ is the third layer of dimension #3. With F1 we discover the set LF1 containing all fuzzy large descriptor set of F1.With F2 we discover the set LF2 containing all fuzzy large descriptor sets of F2. With F3 we discover the sets LF3 containing all fuzzy large descriptor sets of F3.Let SF1 be the subset of LF1, SF2 be the subset of LF2 and $SF12=SF1\cap SF2 <> \{\}$. We build the fuzzy association rules based on SF12. Let Li and Lj be the subset of S12 and $S12 = Li\cup Lj$ and $Li\cap Lj=\{\}$. we define a fuzzy association rule between Li and Lj and denoted as Li --- > Lj. The meanings of these rules are as follows:

There are n% of customer who buy MANY item X, also buy FEW item Y, or

There are m% of customer who buy AVERAGE item T, also buy MANY item U.

8. CONCLUSIONS AND FUTURE WORKS

The usage of fuzzy information system and some fuzzy concepts help us to upgrade the meaning of binary and fuzzy association rules discovered from a large database. This approach give us a chance to integrate the fuzzy set with rough set into data mining application and give us an idea to increase the speed of mining process. We also think about how to change the continuous data and discrete data to an information by using fuzzy membership functions. We continue to study the partitioning methods for converting the relational database to our information systems and improving the speed of our algorithm by using a parallel algorithm.

KHÁM PHÁ CÁC LUẬT KẾT HỢP NHỊ PHÂN VÀ KẾT HỢP MỜ TRONG CƠ SỞ DỮ LIỆU

Hoàng Kiếm - Đỗ Phúc

TÓM TẮT: Trong bài báo này, chúng tôi đề xuất một phương pháp để phát hiện các luật kết hợp nhị phân và kết hợp mờ trong cơ sở dữ liệu. Chúng tôi nghiên cứu một kiểu luật kết hợp mới: luật kết hợp mờ. Chúng tôi dùng lý thuyết tập thô của Z. Pawlak và lý thuyết tập mờ của L.A. Zadeh để định nghĩa một số khái niệm và phát triển phương pháp dựa trên các khái niệm mới này. Chúng tôi cũng đề xuất một ứng dụng nhằm khám phá các luật kết hợp mờ từ các cơ sở dữ liệu trong siêu thị. Trong ứng dụng này, chúng tôi chú ý đến số lượng mặt hàng được mua thay vì chỉ lưu ý " là có mua hay không mua một mặt hàng nào đó" như các tiếp cận truyền thống đã làm.

REFERENCES

[1]. A. Surasere, E. Omiecinsky. An efficient algorithm for mining association rules in large database In Porc 21st VLDB,(1995)

[2]. Hoang Kiem, Do Phuc, Using MDDM for mining association rules in a large database, proceedings of conference, VNU-HCMC, Vietnam, (1998).

[4]. Hoang Kiem, Do Phuc, A combined multi-dimensional data model, self-learning algorithm and genetic algorithm for cluster discovery in data mining-Proceedings of the third international conference in KDD, Bejing, China, (1999)

[5]. Hoang Kiem, Do Phuc, Using data mining in education and training, Magazine of science and technology development, Vol 1, No. 4, VNU-HCM, Vietnam, (1999).

[6]. Hoang Kiem, Do Phuc Using rough genetic and Kohonen 's neural network for conceptual discovery in data mining, Proceedings of RSFDGRC'99 conference, University of Yamaguchi, UBE, Japan.(1999).

[7]. Ho Tu Bao, Automatic unsupervised learning on Galois lattice, Proceedings of IOIT conference, Hanoi, Vietnam, (1996).

[8]. Jun Yan, Using Fuzzy Logic, Prentice Hall, (1994)

[9]. K Rajamani, Sam Sung Extending the applicability of Association rule In Proc of PAKDD 99, Bejing, China, (1999)

[10]. Pieter Adrians, Dolf Zantige, Data Mining, Addison Wesley, Longman, 1996.

[11]. Timothy J. Ross, Fuzzy logic with engineer application, Mac Graw-Hill, (1995).

[12]. R. Agrawal , R. Srikant Fast Algorithm for mining association rules in large database, Research report RJ, IBM Almaden Research Center, San Jose, CA (1994)

[13]. R. Agrawal , R. Srikant Mining generalized association rules. In Porc of 21st Int'l conference on VLDB, Zurich, Switchzerland (1995)

[14] Z.Pawlak, Rough Set Int. Journal of Information and computer. Sciences 11(1982)