DISCOVERING FUZZY CLASSIFICATION RULES FROM DATABASE BASED ON THE GENETIC ALGORITHM

Do Phuc, Hoang Kiem
University of Natural Sciences – VNU-HCM
Department of Information Technology
(Received 12 November 2001, Revised 19 December 2001)

ABSTRACT: In this paper, we propose a method for discovering the fuzzy classification rules from a database based on the genetic algorithm. We would like to propose a method for choosing a set of threshold values which can help us to convert a fuzzy information table to a binary information table by using the genetic algorithm. We develop an algorithm for mining the fuzzy classification rules from binary information table and a fitness function which can measure the goodness of the discovered classification rules. A chromosome of threshold values has been used in genetic algorithm for selecting the appropriate set of thresholds for converting the fuzzy information table to binary information table with high goodness measure. We also propose an application to a supermarket database. In this application, we study the quantity of purchased items instead of 'to be purchased or-not to be purchased' as the traditional approach did and mining the fuzzy classification rules from this database.

Keywords: association rules, binary information table, descriptor vector, fuzzy information table, fuzzy classification rules, frequent set, genetic algorithm

1. Introduction

Problem of classification rule discovery is one of the key problems in knowledge discovery from large database. We have developed an algorithm for discovering the classification rules from a binary information table [5]. In this paper, we would like to extend our proposed algorithm to discover the fuzzy classification rules from a fuzzy information table. Let d_i be the item and μ_i be the fuzzy member function for item d_i, we would like to discover the fuzzy classification rules which may have the following format:

If $\mu_2(d_2) \in [0.23,0.76]$ **and** $\mu_5(d_5) \in [0.45,0.87]$ **then** Class1.

The key issue of our research is how to select a set of intervals which can help us to convert a fuzzy information table to a binary information table and we use our proposed algorithm for discovering the fuzzy classification rules. We use genetic algorithm for choosing the appropriate set of intervals. A chromosome is a candidate of intervals. We also discuss a real life application in discovering the fuzzy classification rules from a supermarket database. This paper is organized as follows: 1) Introduction; 2) Binary Information table 3) An algorithm for discovering the descriptor sets and association rules from binary information table 4) Classification problem 5) Fuzzy information table 6) Using the genetic algorithm for discovering the thresholds 7) An application to a supermarket database 8) Conclusions

2. Binary information table

We would like to define some definitions to be used in our proposed algorithms.

Definition 1: Binary information table

A binary information table is a triple $S_B=(O,D,R_B)$ where O is a finite set of objects, D is a finite set of descriptors, R_B is a binary relation between O and D. R_B is represented by a function $\chi:OxD\rightarrow\{0,1\}$ where $\chi(o,d)=1\Leftrightarrow(o,d)\in R_B$. Let P(X) be the power set of X. Given $S\in P(D)$ and $X\in P(O)$, we define two binary information mappings $\rho_B:P(D)\rightarrow P(O)$, $\lambda_B: P(O) \rightarrow P(D)$.

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

We hold that if $S_1\subset S_2 \Rightarrow \rho_B(S_2)\subset\rho_B(S_1)$.

Definition 2: Frequent sets

Given $S_B=(O,D,R_B)$ and a threshold $\alpha\in[0,1]$. Let $S\in P(D)$, the support of S is denoted as SP(S) and calculated by:

$$SP(S) = |\rho_B(S))| / |O| \quad (1)$$

Where |X| is the cardinality of X. Given $S\in P(D)$ and a threshold $\alpha\in[0,1]$, S is a frequent set with threshold α iff $SP(S)\geq \alpha$. Given $S_B=(O,D,R_B)$, we denote $L_B(S_B,\alpha)$ as a set of all frequent sets with threshold υ of S_B. We have the property $\forall B\in L_B(S_B,\alpha)$, $A\subset B \Rightarrow A\in L_B(S_B,\alpha)$.

Let h be an integer, we denote $L_{B,h}=\{S\in L_B(S_B,\alpha) \mid |S|=h\}$.

Definition 3: Association rules

Given $S_B=(O,D,R_B)$ and a threshold $\alpha\in[0,1]$. Let $A\in L_B(S_B,\alpha)$, X and Y be the subsets of A where $A=X\cup Y$ and $X\cap Y=\varnothing$. An association rule between frequent set X and Y is a mapping $r:X\rightarrow Y$. The support of $r:X\rightarrow Y$ is $SP(X\cup Y)$ and the confidence of this rule is denoted as $CF(X\rightarrow Y)$ and calculated by:

$$CF(X\rightarrow Y) = |\rho_B(X)\cap\rho_B(Y)| / |\rho_B(X)| = SP(X\cup Y) / SP(X) \quad (2)$$

Given $S_B=(O,D,R_B)$, we denote $R_B(S_B,\alpha,\beta)$ as the set of all association rules of S_B where $SP(r)\geq\alpha$ and $CF(r)\geq\beta$. We also define two functions $Left(X\rightarrow Y)=X$ and $Right(X\rightarrow Y)=Y$.

Definition 4: Descriptor vector

Given $S_B=(O,D,R_B)$ where $O=\{o_1,o_2\ldots o_m\}$ and $D=\{d_1,d_2\ldots d_n\}$. Let $X\in P(D)$, a descriptor vector $v_B(X)=(X_1,\ldots,X_n)$ is a vector with n components where component X_i corresponds to descriptor d_j and takes a value 1 if descriptor $d_j\in X$, otherwise $X_i=0$. Let $V(S_B)$ be a set of all descriptor vectors of $S_B=(O,D,R_B)$. If card(X)=1, X is a descriptor of S_B and $X_j=\chi(o_j,X)$. Given $X, Y\in P(D)$, let $v_B(X)=(X_1,\ldots,X_n)$ and $v_B(Y)=(Y_1,\ldots,Y_n)$ be the elements of $V(S_B)$. We define the descriptor vector product $\otimes$ of $v_B(X)$ and $v_B(Y)$ as:

$$v_B(Z)=v_B(X)\otimes v_B(Y) \quad (3)$$

Where $Z_j = \min(X_j, Y_j)$, j=1,...,n and $Z=X\cup Y$. From vector $v_B(X)$, we know all objects having descriptor set X_1 and X_2. Therefore, we use $v_B(X)$ for calculating $\rho_B(X)$. If we have $v_B(X)$, $v_B(Y)$ and $Z=X\cup Y$, we can calculate $v_B(Z)$ and $\rho_B(Z)$ by using (3). We used this point in our algorithm for increasing the efficiency of calculating the frequent set. We denote $VS_{B,h}(S_B)$ as a subset of $V(S_B)$ containing only vector $v_B(X)$ where $X\in P(D)$ and card(X)= h.

3. An algorithm for discovering the frequent sets and association rules

3.1. Problem statement

Given $S_B=(O,D,R_B)$ and threshold $\alpha, \beta \in [0,1]$, find:

a) $L_B(S_B, \alpha)$ b) $R_B(S_B,\alpha,\beta)$

3.2. Discovering $L_B(S_B, \alpha)$

Step 1: Discover the frequent sets

Answer =∅
Generate $L_{B,1}(S_B, \alpha)$ from S_B;
For (k=2; $L_{B,k}(S_B, \alpha) \neq \emptyset$; k++) do
Begin
Generate $L_{B,k}(S_B, \alpha)$ from $L_{B,k-1}(S_B, \alpha)$
Answer = $\cup_k L_{B,k}(S_B, \alpha)$
End
Return Answer;

Step 2: Generating $L_{B,1}(S_B, \alpha)$

We test each descriptor set d_j of D if $SP(\{d_j\}) \geq \alpha$ then d_j will be saved to $L_{B,1}(S_B, \alpha)$.

Step 3: Generating $L_{B,k}(S_B, \alpha)$ from $L_{B,k-1}(S_B, \alpha)$

Based on the property : $T \in L_B(S_B, \alpha)$, $S \subset T \Rightarrow S \in L_B(S_B, \alpha)$, we generate $L_{B,k}(S_B, \alpha)$ from $L_{B,k-1}(S_B, \alpha)$.

The procedure is as follows:

Create a matrix which rows and columns are the elements of $L_{B,k-1}$.
$L_{B,k} = \emptyset$;
For(each $X \in L_{B,k-1}(S_B, \alpha)$) do
For (each $Y \in L_{B,k-1}(S_B, \alpha)$ and X <> Y) do begin
$Z = X \cup Y$;
Restore vector $v_B(X)$ and $v_B(Y)$ from $VS_{B,k-1}$
$v_B(Z) = v_B(X) \otimes v_B(Y)$
From $v_B(Z)$, calcualate $\rho_B(Z)$
If ($\rho_B(Z) \geq \alpha$*card(O)) and Card(Z) = k and $Z \notin L_{B,k}$) then
begin
SaveFreqSet(T, $L_{B,k}(S_B, \alpha)$);
SaveDescVector($v_B(Z)$, $VS_{B,k}$);
End;
End;
Answer = $L_{B,k}(S_B, \alpha)$;
Return Answer;

Where:

SaveFreqSet(T, $L_{B,k}(S_B, \alpha)$) is a function for saving frequent set T to $L_{B,k}(S_B, \alpha)$

SaveDescVector($v_B(Z)$,$VS_{B,k}(S_B)$) is a function for saving descriptor vector $v_B(Z)$to $VS_{B,k}(S_B)$.

Step 4: Discovering $R(S_B, \alpha, \beta)$

$R(S_B, \alpha, \beta) = \emptyset$;
For each $L \in L_B(S_B, \alpha)$ do begin
For (each $X,Y \in L$ and $X \cup Y = L$ and $X \cap Y = \emptyset$) do begin

If ($CF_B (X \rightarrow Y) \geq \beta$) and $X \rightarrow Y \notin R(S_B, \alpha, \beta)$ then SaveRule($X \rightarrow Y$, $R(S_B, \alpha, \beta)$).
If ($CF_B (Y \rightarrow X) \geq \beta$) and $Y \rightarrow X \notin R(S_B, \alpha, \beta)$ then SaveRule($X \rightarrow Y$, $R (S_B, \alpha, \beta)$).
End;
End;
Return ($R_B (S_B, \alpha, \beta)$) ;

where: SaveRule($X \rightarrow Y$, $R_B(S_B,\alpha,\beta)$ is a function for saving the association rule to $R_B(S_B,\alpha,\beta)$.

3.3. Increasing the efficiency of mining the frequent sets

To increase the efficiency and to reduce the time of scanning the database, we use a sparse matrix to organize the binary information table in the memory. The non zero element of the binary information table is a triple (row, column, value). The rows and columns of sparse matrix are stored in the double linked lists. We develop operations for processing the descriptor vector operations on the double linked lists of binary information matrix. Besides, we use the descriptor vector product to calculate the SP(Z) where $Z = X \cup Y$ as mentioned above.

4. Classification problem

4.1. Discovering the classification rules based on the association rules

Given a special binary information table $(O,D=H \cup C,R_B)$ where H is a set of condition descriptors and C is a set of decision descriptors. This special kind of binary information table is called binary decision table [5].

Given threshold $\alpha \in [0,1]$, $\beta=1.0$ we would like to discover classification rule $X \rightarrow Y$ where $X \in P(H), Y \in P(C)$, $card(Y)=1$, $SP(X \rightarrow Y) \geq \alpha$, $CF(X \rightarrow Y)=1.0$. We employ the above algorithm for discovering the classification rules.

4.2. The goodness of a set of classification rules.

Let $C=\{c_1,c_2,...,c_K\}$ be the set of K classes, we build $RC(c_i) = \{r: X \rightarrow \{c_i\} | CF(r)=1, SP(r) \geq \alpha\}$ and define the goodness of a set of classification rules for class i as follows:

$$GoodMeasure(S_B, \upsilon, 1.0, i) = (1/N_i) \left(\cup_{r \in RC(i)} card (\rho_B(Left(r)) \right) \quad (4)$$

Where Ni is the cardinality of the set of all objects of class i. The classification goodness for all classes of a binary decision table is calculated by:

$$GoodAllClass(S_B, \upsilon, 1.0) = (1/K) \left(\sum_{i=1}^{K} card \left(\cup_{r \in RC(i)} \rho_B(Left(r)) \right) \right) \quad (5)$$

We can adjust two thresholds α, β to get a high goodness of classification rule set.

5. Fuzzy information table

A fuzzy information table is a triple $S_F=(O,D,R_F)$ where O is a finite set of objects, D is a finite set of descriptors, R_F is a fuzzy relation between O and D. R_F is represented by a fuzzy

function μ: OxD$\rightarrow$[0,1], where $\mu(o,d)$ expresses the degree of object o having descriptor d. and [0,1] is a subset of real number set.

5.2. Converting a fuzzy information table to a binary information table

We can convert a fuzzy information table (O,D,R_F) to binary information table (O,D,R_B) by using the rule: $\chi(o,d)= 1$ iff $\mu(o,d)\in[L(d),U(d)]$ and otherwise $\chi(o,d)= 0$ where $[L(d),U(d)]$ is an interval of dom(d) and $L(d)\leq U(d)$. Depending on the $[L(d),U(d)]$, we can build R_B from R_F or $S_B(O,D,R_B)$ from $S_F=(O,D,R_F)$.

5.3.Problem statement

Given a fuzzy information table $S_F=(O,D,R_F)$ and threshold υ,β, $\alpha \in[0,1]$, our problem is to find a set of n interval $[L(d),U(d)]$ where $d\in D$ which maximize the function $GoodAllClass(S_B,\upsilon,1.0)$. R_B is created from R_F by using the n intervals $[L(d),U(d)]$ and $d\in D$.

6. Using the genetic algorithm for discovering the thresholds

Genetic algorithm (GA) is a form of adaptive search techniques initially introduced by Holland. GA is an iterative search procedures that maintain a population of candidate solution to the object function, each member of population is called chromosome. During each iteration step, called a generation, the current population is evaluated and on the basis of that evaluation, a new population of candidate solution is formed. A general sketch of genetic algorithm is in figure 1.

```
Making the initial population of chromosomes
Evaluating the initial population
While terminate condition not satisfied do
        Select the current population from the previous one
        If crossover condition satisfied
                perform crossover
        If mutation condition satisfied
                perform mutation
        Evaluate the fitness of the offsprints
Endwhile
```

Figure 1: A sketch of genetic algorithm

There are four basic steps in genetic algorithm:

Step 1: Making the initial population

The initial population of chromosomes is chosen randomly. Let n=card(D) then the chromosome i denoted CHR_i will have the format $<v_{i1},v_{i2}...,v_{in}>$. An element v_{ij} of chromosome i is a couple $<L(d_j),U(d_j)>$ where $L(d_j),U(d_j)\in[0,1]$ and $L(d_j) \leq U(d_j)$.

Step 2: Crossover operator

Given two parental chromosomes $<v_{i1},.v_{i2},..,v_{ik},...v_{in}>$ and $<v_{j1},v_{j2},...,v_{jk},...,v_{jn}>$

Select a random position $k \in[1,2,...,n]$ and swap two portions of two parental chromosomes.

The result after crossover operation $<v_{i1},.v_{i2},..,v_{jk},...,v_{jn}>$ and $<v_{j1},v_{j2},...,v_{ik},...v_{in}>$

Step 3: Mutation operator

Given a chromosome $<v_{i1},.v_{i2},..,v_{ik},...v_{in}>$, select a random position $k \in[1,2,...,n]$ and select randomly $L(d_k)$ or $U(d_k)$, replace $L(d_k)$ or $U(d_k)$ with a random value in [0,1].

Step 4: Evaluation

The fitness function is defined as follows:

Given a chromosome $CHR_i = <v_{i1},.v_{i2},..,v_{ik},...v_{in}>$, From this chromosome , we can convert R_F to R_B and change (O,D,R_F) to (O,D,R_B). The fitness function $f(CHR_i)$ is defined as:

$$f(CHR_i) = GoodAllClass(S_B, \upsilon, 1.0) = (1/K)(\sum_{i=1}^{K} card\ (\cup_{r \in RC(i)} \rho_B(Left(r)))) \quad (6)$$

7. An application to a supermarket database

Suppose that, we have a table where rows are transactions and columns are items and classes as in table 1.

Table 1: *Table of transaction and items*

	d_0	d_1	d_2	d_3	d_4	d_5	c_1	c_2
o_0	42	65	0	0	12	12	1	0
o_1	12	12	65	12	13	13	1	0
o_2	43	60	0	0	0	0	1	0
o_3	12	26	65	12	0	0	1	0
o_4	12	0	14	89	0	0	0	1
o_5	14	0	14	0	88	99	0	1
o_6	15	0	0	90	12	0	0	1
o_7	12	0	11	0	80	87	0	1

The element $b_{i,j}$ of this table is the quantity of the purchased item d_j in transaction o_i and c_1, c_2 are two transaction categories Let O be the set of transactions, H be the set of items, C={ c_1, c_2} be a set of transaction categories. f is fuzzy membership functions μ_i: $dom(d_i) \rightarrow [0,1]$, i=1,...,card(H).

Each descriptor has a particular membership function. This function expresses the meaning of quantity. We define fuzzy functions MANY, FEW, AVERAGE for each descriptor, because when we buy three cars we think that we buy many cars, but when we buy three pens, we think that we buy few pens. Based on a group of fuzzy membership functions, we can convert the above table to a fuzzy information table as shown in table 2.

Table 2: *Fuzzy information table*

	D_0	d_1	d_2	d_3	d_4	d_5	c_1	c_2
o_0	0.6	0.6	0.0	0.0	0.0	0.1	1	0
o_1	0.1	0.1	0.8	0.1	0.0	0.2	1	0
o_2	0.6	0.5	0.0	0.0	0.0	0.0	1	0
o_3	0.1	0.3	0.8	0.1	0.0	0.0	1	0
o_4	0.1	0.2	0.2	0.7	0.1	0.0	0	1
o_5	0.2	0.0	0.2	0.0	0.9	0.9	0	1
o_6	0.3	0.0	0.0	0.9	0.2	0.0	0	1
o_7	0.1	0.0	0.1	0.0	0.8	0.7	0	1

We use genetic algorithm with the following parameters :

Population size: 50; Selection Probability: 0.1; Crossover Probability: 0.1; Mutation Probability: 0.1

We have the following solution:

(0.75.0.95)(0.31.0.95)(0.36.0.76)(0.61.0.92)(0.81.0.91)(0.66.0.74)

The binary information table appropriated with the above set of threshold is shown in table 3.

Table 3: A binary information table

	d_0	D_1	d_2	d_3	d_4	d_5	c_1	c_2
o_0	0	1	0	0	0	0	1	0
o_1	0	0	0	0	0	0	1	0
o_2	0	1	0	0	0	0	1	0
o_3	0	0	0	0	0	0	1	**0**
o_4	0	0	0	1	0	0	0	1
o_5	0	0	0	0	1	0	0	1
o_6	0	0	0	1	0	0	0	1
o_7	0	0	0	0	0	1	0	1

The association rule based classification rules are as follows:

R:{d1}->{c0} ; SP(r)= 0.25; CF(r)= 1.00; R:{d3}->{c1} ; SP(r)= 0.25; CF(r)= 1.00

R:{d4}->{c1} ; SP(r)= 0.13; CF(r)= 1.00; R:{d5}->{c1} ; SP(r)= 0.13; CF(r)= 1.00

With rule R:{d1}->{c0} ; SP(r)= 0.25; CF(r)= 1.00

A couple of threshold for d1 is (0.31.0.95). It means that, we have the following fuzzy classification rules:

If $\mu_1(d_1) \in [0.31,0.95]$ Then Class C0.

7. Conclusions

We develop algorithms for discovering the frequent sets from binary information table. Based on the frequent sets, we build association rules, classification rules. From binary information table, we extend our application to fuzzy information table. We use the genetic algorithm for selecting a set of appropriate thresholds for converting the fuzzy information table to binary information table and use the existing algorithm for discovering the binary classification rules. The fitness function, the genetic operations are built and an application in supermarket is proposed to discover the fuzzy classification rules for categorizing the transactions based on the purchased quantity of items.

KHÁM PHÁ CÁC LUẬT MỜ TỪ CƠ SỞ DỮ LIỆU DỰA TRÊN THUẬT TOÁN DI TRUYỀN

Đỗ Phúc, Hoàng Kiếm
Khoa Công nghệ Thông tin
Đại học Khoa học tự nhiên Tp.HCM
(Bài nhận ngày 12 tháng 11 năm 2001, hoàn chỉnh sửa chữa ngày 19 tháng 12 năm 2001)

TÓM TẮT: Bài báo đề xuất một phương pháp để khám phá các luật phân lớp mờ từ cơ sở dữ liệu dựa trên thuật toán di truyền. Mục tiêu là tìm kiếm một tập ngưỡng để chuyển bảng thông tin mờ sang bảng thông tin nhị phân. Sau đó sử dụng các thuật toán rút luật từ bảng nhị phân. Các cá thể của thuật toán di truyền sẽ lưu tập các ngưỡng tiềm năng. Các thao tác lai ghép, đột biến và hàm phù hợp được định nghĩa theo yêu cầu của bài toán để tìm tập ngưỡng thích hợp. Chúng tôi đã sử dụng mô hình đề xuất vào bài toán bán hàng siêu thị, trong đó có chú ý đến số lượng mặt hàng bán ra thay vì chỉ lưu ý đến việc có bán hay không bán hàng như các tiếp cận truyền thống.

REFERENCES

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

[2] C-w,K-Chen,D.Y.Y. Yun: Integrating classification and association rules mining: a concept lattice framework, Proceedings of PAKDD'99 conference, Beijing, China, (1999)

[3] Dao I Lin Pincer-Search: A new algorithm for discovering the maximal frequent set, Proceedings of VLDB, (1995)

[4] 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)

[5] Hoang Kiem, Do Phuc: On the extension of lower approximation of the rough set theory for the classification in data mining, WCC2000 international conference , Beijing, China (2000).

[6] Ho Tu Bao: A measure of attribute selection in inductive decision tree, Proceedings of IOIT conference, Hanoi, Vietnam, (1996).

[7] K Rajamani, Sam Sung: Extending the applicability of Association rule In Proc of PAKDD'99, Beijing, China, (1999)

[8] Pieter Adrians, Dolf Zantige: Data Mining, Addison Wesley, Longman, (1996)

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

[10] R. Agrawal, R. Srikant: Mining generalized association rules. Proceedings of 21st VLDB, (1995)

[11] Z.Pawlak: Rough Set Perspective, Proceedings of PAKDD99, Beijing, China, (1999)

[12] Z.Pawlak: Decision Rules, Bayes's rules and Rough Set, Proceedings of RSFDGRC'99 conference, University of Yamaguchi, UBE, Japan, (1999)