Application of Data Mining Based on Artificial Immunity in Marketing

Jun Ju and Hong Zhang

School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221008, Jiangsu, P.R. China jujuncumt@163.comb
hongzh@cumt.edu.cn

Abstract. Facing fierce market competition, how to make the effective marketing strategy rapidly according to the market demand is of vital importance to the survival and development of enterprises. Data mining can discover and extract the latent predictable information from the large database and data warehouse. Using this technology, marketers may obtain potential associations among sales data, so that they can make a market analysis, adopt pertinent marketing strategy, reduce costs and raise profits. This paper proposes an algorithm of association rule mining based on artificial immunity. Practice proves that this algorithm is robustness, hidden parallelism and commonality. It can discover useful association rule from sales data rapidly and effectively, and provide forceful support for enterprises to make accurate marketing strategy.

Keywords: *Data mining, Artificial immunity, Association rule, Marketing strategy*

1. INTRODUCTION

Along with the increasingly fierce competition of enterprises, how to make the effective marketing strategies rapidly according to the market demand is of vital importance to the survival and development of enterprises. The existing marketing decision support system lacks ability of data analysis and mining function. It is difficult for enterprises to find out business operating laws and patterns from a mass of data and make timely and effective response to the customers' demand, which affects the improvement of the competitiveness of enterprises. Data mining is a new technology, which can discover and extract the latent predictive information from the large database or data warehouse. Using this technology, marketers may obtain potential relations among sales data, so that they can make a market analysis, adopt pertinent marketing strategies, reduce costs and raise profits. This paper proposes an algorithm of association rule mining based on artificial immunity. It can discover potentially useful association rules and mining sales data more quickly and more effectively.

2. OVERVIEW OF ASSOCIATION RULES MINING

Association rule mining is one of the important methods of data mining. It discovers interesting associations and relationships between data with different attributes by analyzing them. In the daily marketing process, enterprises store a mass of sales data. Analyzing these data shows that there are large numbers of association rules among them. These association rules may help the policy-maker to analyze the characteristics and the laws of sales data and facilitate them to make pointed marketing strategies.

2.1 Description of Association Rule

Given a set of items $I=\{I_1, I_2, \ldots, I_3\}$ and a transaction database $D=\{t_1, t_2, \ldots, t_n\}$, where $t_i=\{I_{i1}, I_{i2}, \ldots, I_{ik}\}$, and $I_{ij} \in I$. An association rule is an implicit expression like $A \rightarrow B$, where A and B are two item sets, and $A \subset I$, $B \subset I$, $A \cap B = \emptyset$ [1].

The support of the rule is defined as the percentage of transactions containing $A \cup B$ in all transactions [1], which is assumed as S (A, B). The support is used to measure the importance of the rule.

The confidence of the rule is defined as the ratio of the number of transactions containing $A \cup B$ to the number of transactions containing A [1], that is S(A,B)/S(A). The confidence is a measurement of the accuracy of the rule.

In order to obtain meaningful association rules, two threshold values are needed to given: min_support and min_confidence. Association rule mining is aimed at discovering all association rules satisfying min_support and min_confidence.

2.2 The Algorithm of Association Rule Mining

Many algorithms for association rule mining have been studied since Agrawal proposed mining association rules between item-sets in transaction database in 1993. Apriori [2] is the most popular algorithm in many algorithms listing frequent item sets. This algorithm discovers association rules in two steps. The fist step finds out all frequent item sets according to the min_support. The second step generates association rules according to the min_confidence and frequent item sets. But the algorithm may produce massive candidate item sets and scan the database repeatedly. The complexity of calculation is very high. In the application of marketing, this algorithm will find out a large number of rules. But, it is actually difficult to discover useful rules and is unable to satisfy the enterprises' needs for policy-making.

In addition, Park proposed the algorithm of DHP. But its trimming and pruning characteristics are unrealistic in many practical applications. Although partitioning algorithm reduces the consumption of I/O, it has many problems in dealing with higher dimensional item sets transactions. FP-growth algorithm creates a kind of affinitive tree structure, which contributes to generating candidate item sets. But it is inefficient to the production sale database, because it has to traverse the whole database, causing large amount of computation [3].

In order to solve problems in above algorithms in application of enterprise's sales data mining, this paper introduces a concept of artificial immunity and proposes an algorithm of association rule mining based on it by using its multiple spot random intelligent search technique and its unique mechanism of immune memory.

3. BASIC CONCEPTS OF ARTIFICIAL IMMUNITY

Inspired by the biologic immune system, artificial immune system can imitate the function of the natural immune system. It provides evolutionary learning mechanism, such as noise tolerance, self-learning, self-organization and memory, by studying the learning technologies of creatures' natural defense mechanism. It has the potential ability of providing an original approach to the problems [4].

The response process of immune system's resisting external invasion and protecting the body from pathogen is called immunity. External harmful pathogen invades the organism, activates the immune cell and induces it to have the response. This process is called immune response. Immune response is divided into two kinds: inherent immunity and acquired immunity. The former is congenital for the organism, which can get rid of viruses quickly. The latter specifically identifies and removes pathogens, which has many good characteristics such as specificity, memory, differentiating between self and non-self, diversity and self-regulation. The material, which induces the immune system to produce immune responses, is called antigen. The immune cell, which specifically combines with antigen, is called antibody. Affinity shows the bonding strength of an antigen-antibody combination. The higher affinity indicates the higher bonding strength. The antibody with high-affinity is promoted, otherwise inhibited.

4. ASSOCIATION RULE MINING BASED ON ARTIFICIAL IMMUNITY

We can see from above concepts that artificial immune system has powerful abilities of identification, learning and memory and characteristics of distribution, self-organization and diversity. Its capability of differentiating between self and non-self is a good solution for association rule mining. The algorithms of mining association rule based on artificial immunity proposed in this paper aims at improving overall performance in mining association rules, helping policy-maker discovery useful rules in sales data rapidly and make targeted marketing strategy.

4.1 Problem Description

First, we need describe association rule in the artificial immune system. In the ordinary way, problems being solved are expressed as antigens, each feasible result is expressed as antibody, and the object function of the solution is expressed as the affinity between the antigen and the antibody. In this paper, sales attribution which

users are interested in is considered as antigen (assuming as A), the antibody (assuming as A→B) is produced by the artificial immune system. If the affinity (assuming as C (A, B)) between the antigen and the antibody is larger then the threshold of minimal affinity, the antigen is recognized as the non-self, and the antibody is just the association rule to be mined. Otherwise, the antigen is recognized as the self, and the antibody is abandoned.

4.2 Code Scheme

Secondly, we must encode the initial data. Now, there are three kinds of encoded modes for antigen and antibody in immune algorithm: binary coding, real number coding and character coding, and a few is grayscale coding. This paper is directed at transaction database relating to sales data, so the antigen and the antibody in the algorithm we proposed are encoded in character coding.4.3 An Enterprise Architecture Provision.

4.3 Affinity Calculation

In order to make the affinity function reflect the matching degree between rules and data sets well, some definitions are given as follows:

Definition 1: the expected confidence of the rule A→B is the percentage of transactions containing B in all transactions, assuming as S(B).

Definition 2: the lift of the rule A→B is the ratio of the confidence to the expected confidence, that is lift(A→B)=S(A,B)/S(A)S(B).

The expected confidence describes the support of B itself in the absence of A. The lift shows how great influence A has on B. The bigger the lift is, the greater influence A has on B.

Definition 3: the affinity function of the rule is defined as follows:

C(A,B)= S(A,B)+ S(A,B)/S(A)+ S(A,B)/S(A)S(B)

In this paper, the affinity function of the rule is the sum of the support, the confidence and the lift. The bigger support of the rule indicates that the proportion of the rule in the data sets space is larger and the universal significance of the rule is better. The confidence shows the accuracy of conclusions derived from conditions. The rule is constant true when the confidence is equal to 1. The bigger lift of the rule indicates that the conclusion is greater influenced by conditions. Generally, the lift of useful association rule should be larger than 1. Only when the confidence is larger than the expect confidence can it indicates that conditions contribute to conclusions and there are relevancies between conclusions and conditions to a certain extent.

4.4 Generation of Association Rule

In this paper, we set a relation-rule table and an interest-rule table in the algorithm memory bank, which record such items as structure, confidence, support and lift of rules. Relation rules and interest rules are stored in the relation-rule table and the

interest-rule table respectively. In order to guarantee that the antibody has a higher affinity, it is needed to inhibit or promote the antibody. Therefore, we size the interest-rule table according to the actual needs. To the antibody newly generated, that is newly possible rule, calculate its support, confidence and lift, and add the result satisfying the affinity to the interest-rule table. If the interest-rule table is full, the antibody, which is newly generated and has a higher affinity with the antigen, will replace the lower one. If the support and the confidence of the rule in the interest-rule table are larger than min_support and min_confidence respectively, add it to the relation-rule table. Output association rules from the relation-rule table in the end of the algorithm.

4.5 Algorithm Flow

Suppose that there are m records in the database, and each record has n attributes.

Step1 Antigen recognition. Choose interested attributes to be antigens. The purpose is to discover antibodies (association rules), which can be combined with the antigens. Assuming that a user chooses the attribute value A to be an antigen.

Step2 Produce initial antibody groups.

Take k records at random

For the record i

If the antigen is contained in the record

{

Calculate the support of the antigen A (S(A))

Combine the antigen A with other attribute supposed to be attribute j (j=1,…,n, and $j \neq i$) to form the antibody (association rule), and calculate the support of attribute j and the antibody $A \rightarrow j$ (S(j), S(A,j))

}

Else deal with the next record

Step 3 Calculate the affinity: $C(A,j)= S(A,j)+ S(A,j)/S(A)+ S(A,j)/S(A)S(j)$

Step 4 If no new antibody is found, then turn to step 5

Else {

Add the antibody satisfied the affinity to the interest-rule table. If the interest-rule table is full, the antibody that is newly generated and has a higher affinity with the antigen will replace the lower one.

}

Step 5 Generate antibodies

Scan the database once, calculate the support and the confidence of all antibodies in the interest-rule table. If the support and the confidence of the rule in the interest-rule table are larger than min_support and min_confidence respectively, add it to the relation-rule table. Output association rules from the relation-rule table in the end.

5. THE APPLICATION OF ASSOCIATION RULE MINING BASED ON ARTIFICIAL IMMUNITY IN MARKETING

The following is an example of customers' purchase. Next, we will mine the sales data with the algorithm proposed in this paper, and analyze the association rules acquired.

An electronic products shop has retained transaction records during a certain time. Each record shows the details of one customer purchasing once, shown in table1. We list six customers and five products for example for the lack of space.

Table 1. Customers Purchase Table

Customers	Goods purchased
A	computer, scanner, printer
B	scanner, duplicator, ink box
C	computer, scanner
D	computer, printer, duplicator, ink box
E	ink box
F	duplicator, ink box

Analyze the table 1 with the algorithm proposed. Suppose the threshold value of the affinity is 2, the min_support is 0.3 and the min_confidence is 0.6. Then obtain association rules as follows:

Rule1：printer→computer，support=0.33，confidence=1，affinity =3.33

Rule2：computer→printer，support=0.33,confidence=0.67，affinity =3

Rule3：computer→scanner，support=0.33，confidence=0.67，affinity =2.33

Rule4：scanner→computer，support=0.33，confidence=0.67，affinity =2.33

Rule5：duplicator→ink box，support=0.5，confidence=1，affinity =3

Rule6：ink box→duplicator，support=0.5，confidence=0.75，affinity =2.75

From above association rules we can draw initial conclusions:

1. Customers who buy a printer almost certainly buy a computer. The proportion of customers buying a computer and customers buying a scanner or a duplicator are equal, according to the actual needs of individuals.
2. Customers who buy an ink box tend to purchase a duplicator, and customers who buy a duplicator almost certainly buy an ink box. It indicates that once customers buy a duplicator, it needs regular replace the ink box.

According to above rules, the shop can take measures as follows in marketing:

1. Put printers with computers and duplicators with ink boxes to make purchasing convenient for the customers.
2. Determine which commodities can be bundled together for bargain sale. For example, if a customer buys a computer, it is more than likely that he will buy a printer bundled together with the computer for bargain sale.

3. After a customer buys one kind of commodity, the shop assistant can recommend another kind of commodity to him.
4. Produce and transport commodities associated with each other together.

After taking these measures, the cross-consumption of customers is increased significantly, and the satisfaction degree of shops and customers is also raised.

6. CONCLUSIONS

We can see from above analysis that association rule mining can discovery useful association knowledge from a mass of business transaction records and help enterprises to design targeted marketing strategies. An algorithm of association rule mining based on artificial immunity is proposed in this paper. It adopts the tactic of "random parallel search", and identifies association rules from sales data rapidly by using the mechanism of identification, learning and memory of artificial immunity. It scans the database only once during the whole process of mining, i.e. identify real association rules from interest rules. There is no need to generate large number of candidate item sets during the process of mining. So it improves the performance of association rule mining dramatically. Practice proves that the algorithm is robustness, hidden parallelism and commonality. It can discover useful association rules from sales data rapidly and effectively, and provide forceful support for enterprises to make accurate marketing strategy. This project is supported by the Natural Science Foundation of Jiangsu Province (serial number: BK2005021).

REFERENCES

1. M.H. Dunham, *Data Mining Courses* (Tsinghua University Press: Beijing, 2005).
2. R. Agrawal and R. Srikant, *Fast Algorithms for Mining Association Rules*, IBM Almaden Research Center (1994). http://rakesh.agrawal-family.com/papers/vldb94apriori.pdf (Accessed February 6, 2007).
3. T. Liu, *Research on Artificial Immune System and Application in Data Mining*. Ph.D Thesis, China University of Mining and Technology (2005).
4. L. Jiao and H. Du, Development and Prospect of the Artificial Immune System, *Acta Electronica Sinica.* Volume 31, Number 10, pp.1540-1548, (2003).