1999 Third International Conference on Knowledge-Based Intelligent Information Engineeing Systems, 31st Aug-1st Sept 1999, Adelaide, Australia

A Fuzzy Data Mining Algorithm for Quantitative Values

Tzung-Pei Hong†, Chan-Sheng Kuo‡ and Sheng-Chai Chi‡
†Department of Information Management
‡Graduate School of Management Science
I-Shou University
Kaohsiung, 84008, Taiwan, R.O.C.
e-mail: tphong@csa500.isu.edu.tw

ABSTRACT

This paper attempts to propose a new data-mining algorithm to enhance the capability of exploring interesting knowledge from transactions with quantitative values. The proposed algorithm integrates the fuzzy set concepts and the apriori mining algorithm to find interesting fuzzy association rules from given transaction data. Experiments on student grades in I-Shou University are also made to verify the performance of the proposed algorithm.

1. Introduction

In data mining researches, inducing association rules from transaction data is the most commonly seen. Most of the previous research works can, however, only handle transaction data with attributes of binary values. In real-world applications, transaction data are usually composed of quantitative values. Designing a sophisticated data-mining algorithm to deal with different types of data turns a challenge in this research topic.

Recently, fuzzy set theory is more and more frequently used in intelligent systems, because of its simplicity and similarity to human reasoning [8]. The theory has been successfully applied to many fields such as manufacturing, engineering, diagnosis, economics, and others [5, 8, 9, 11]. This paper integrates the fuzzy-set concepts and the apriori mining algorithm to find interesting itemsets and fuzzy association rules from transaction data with quantitative values. A fuzzy data mining algorithm is proposed, which is specially capable of transforming quantitative values in transactions into linguistic terms, then filtering them, and finding association rules by modifying the apriori mining algorithm [4] .

2. Review of Agrawal et al's data mining algorithms

As mentioned above, the goal of data mining is to discover the important associations among items such that the presence of some items in a transaction will imply the presence of some other items. For achieving this purpose, Agrawal and his co-workers proposed several mining algorithms based on the concept of large itemsets to find association rules from transactions [1, 2, 3, 4]. They decomposed the mining process into two phases. In the first phase, candidate itemsets are generated and counted by scanning the transactions. If the number of an itemset appearing in the transactions is larger than a pre-defined threshold value (called minimum support), the itemset is thought of as a large itemset. Itemsets with only one item are first processed. The large itemsets with one item are then combined to form candidate itemsets of two items. This process is repeated until all large itemsets are found. In the second phase, the desired association rules are induced from the large itemsets found in the first phase. All the possible combination ways of association rules for each large itemset are formed, and the ones with their calculated confidence values larger than a predefined threshold (called minimum confidence) are output as desired association rules.

In addition to proposing methods for mining association rules from transactions of binary values, Agrawal et al also proposed a method [10] to mine association rules from those with quantitative and categorical attributes. Their proposed method first determines the number of partitions for each quantitative attribute, and then maps all possible values of each attribute into a set of consecutive integers. It then finds the large itemsets whose support values are greater than the user-specified minimum support. These large itemsets are then processed to generate association rules, and the interesting rules are output from the viewpoint of users.

In this paper, we adopt the fuzzy set concepts to mine associate rules from transactions with quantitative attributes. The rules mined out are expressed in linguistic terms, which are more natural and understandable for human beings.

3. The fuzzy data-mining algorithm for quantitative values

In this section, the fuzzy concepts are used in the apriori data-mining algorithm to discover useful association rules from quantitative values. Notation used in this paper is first stated as follows.

n: the total number of transaction data;
m: the total number of attributes;
$D^{(i)}$ *: the i-th transaction data,* $1 \le i \le n$;
A_j *: the j-th attribute,* $1 \le j \le m$;
$|A_j|$ *: the number of fuzzy regions for* A_j;

0-7803-5578-4/99/$10.00©1999 IEEE

R_{jk} : *the k-th fuzzy region of* A_j, $1 \le k \le |A_j|$;

$v_j^{(i)}$: *the quantitative value of* A_j *for* $D^{(i)}$;

$f_j^{(i)}$: *the fuzzy set converted from* $v_j^{(i)}$;

$f_{jk}^{(i)}$: *the membership value of* $v_j^{(i)}$ *in Region* R_{jk};

$count_{jk}$: *the summation of* $f_{jk}^{(i)}$ *for i=1 to n;*

α : *the predefined minimum support;*

λ : *the predefined minimum confidence value;*

C_r: *the set of candidate itemsets with r attributes (items);*

L_r: *the set of large itemsets with r attributes (items).*

The proposed fuzzy mining algorithm first transforms each quantitative value into a fuzzy set with linguistic terms using membership functions. The algorithm then calculates the scalar cardinality of each linguistic term on all the transaction data. The mining process based on fuzzy counts is then performed to find fuzzy association rules. The detail of the proposed mining algorithm is described as follows.

The Fuzzy Data Mining Algorithm:

INPUT: A set of n transaction data, each with m attribute values, a set of membership functions, a predefine minimum support value α, and a predefined confidence value λ.

OUTPUT: A set of fuzzy association rules.

STEP 1: For each transaction data $D^{(i)}$, i=1 to n, and for each attribute A_j, j=1 to m, transfer the quantitative value $v_j^{(i)}$ into a fuzzy set $f_j^{(i)}$ represented as $\left(\frac{f_{j_1}^{(i)}}{R_{j_1}} + \frac{f_{j_2}^{(i)}}{R_{j_2}} + ... + \frac{f_{j_l}^{(i)}}{R_{j_l}}\right)$ using the given membership functions, where R_{jk} is the k-th fuzzy region of attribute A_j, $f_{jk}^{(i)}$ is $v_j^{(i)}$'s fuzzy membership value in region R_{jk}, and l ($=|A_j|$) is the number of fuzzy regions for A_j.

STEP 2: For each attribute region R_{jk}, calculate its scalar cardinality on the transactions:

$$count_{jk} = \sum_{i=1}^{n} f_{jk}^{(i)} .$$

STEP 3: For each R_{jk}, $1 \le j \le m$ and $1 \le k \le |A_j|$, check whether its $count_{jk}$ is larger than or equal to the predefined minimum support value α. If R_{jk} satisfies the above condition, put it in the set of large 1-itemsets (L_1). That is:

$$L_1 = \{ R_{jk} \mid count_{jk} \ge \alpha,\ 1 \le j \le m \text{ and } 1 \le k \le |A_j| \}.$$

STEP 4: Set r=1, where r is used to represent the number of items kept in the current large itemsets.

STEP 5: Generate the candidate set C_{r+1} from L_r in a way similar to that in the apriori algorithm [4] except that two regions belonging to the same attribute cannot simultaneously exist in an itemset in C_{r+1}. Restated, the algorithm first joins L_r and L_r under the condition that r-1 items in the two itemsets are the same and the other one is different. It then keeps in C_{r+1} the itemsets which have all their sub-itemsets of r items existing in L_r and do not have two items R_{jp} and R_{jq} ($p \ne q$).

STEP 6: For each newly formed (r+1)-itemset s with items $(s_1, s_2, ..., s_{r+1})$ in C_{r+1}, do the following substeps:

(a) For each transaction data $D^{(i)}$, calculate its fuzzy value on s as $f_s^{(i)} = f_{s_1}^{(i)} \wedge f_{s_2}^{(i)} \wedge ... \wedge f_{s_{r+1}}^{(i)}$, where $f_{s_j}^{(i)}$ is the membership value of $D^{(i)}$ in region s_j. If the minimum operator is used for the intersection, then $f_s^{(i)} = \underset{j=1}{\overset{r+1}{Min}} f_{s_j}^{(i)}$.

(b) Calculate the scalar cardinality of s on the transactions as:

$$count_s = \sum_{i=1}^{n} f_s^{(i)} .$$

(c) If $count_s$ is larger than or equal to the predefined minimum support value α, put s in L_{r+1}.

STEP 7: IF L_{r+1} is null, then do the next step; otherwise, set r=r+1 and repeat STEPs 5 to 7.

STEP 8: For each large q-itemset s with items $(s_1, s_2, ..., s_q)$, $q \ge 2$, construct the association rules by the following substeps:

(a) Form each possible association rule as follows:

$$s_1 \wedge ... \wedge s_{k-1} \wedge s_{k+1} \wedge ... \wedge s_q \rightarrow s_k ,$$

k=1 to q.

(b) Calculate the confidence value of the above association rule as:

$$\frac{\sum_{i=1}^{n} f_s^{(i)}}{\sum_{i=1}^{n} (f_{s_1}^{(i)} \wedge ... \wedge f_{s_{k-1}}^{(i)}, f_{s_{k+1}}^{(i)} \wedge ... \wedge f_{s_q}^{(i)})}$$

STEP 9: Output the rules with their confidence values larger than or equal to the predefined confidence threshold λ.

After STEP 9, the rules output can act as the meta-knowledge for the given transactions.

4. Experimental Results

Student score data from the Department of Information Management at I-Shou University, Taiwan, were used to show the feasibility of the proposed mining algorithm. A total of 260 transactions were included in the data set. Each transaction consisted of scores that a student had gotten. Execution of the mining algorithm was performed on a Pentium-PC. The minimum support value α was set at 40 and the minimum confidence value λ was set at 0.8. Totally, 42 rules were mined out. Two rules mined out are shown below as examples.

1. If the Data Structure score is Low, then the System Analysis and Design score is middle, with a confidence factor of 0.81.

2. If the Management Information Systems score is middle and the C Programming Language score is middle, then the Business Data Communication score is middle, with a confidence factor of 0.85.

Experiments were also made to show the relationships between numbers of large itemsets and minimum support values. Results are shown in Figure 1.

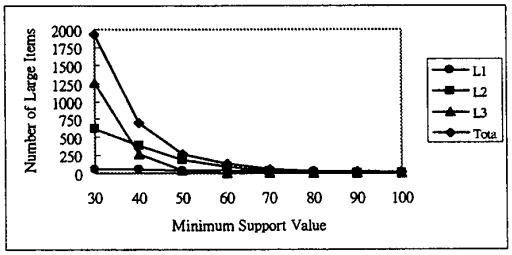

Figure 1. The relationship between numbers of large itemsets and minimum support values.

From Figure 1, it is easily seen that the numbers of large itemsets decreased along with an increase in minimum support values. This is quite consistent with our intuition. The curve of the numbers of large 1-itemsets was also smoother than that of the numbers of large 2-itemsets, meaning that the minimum support value had a larger influence on itemsets with more items. Also, appropriate minimum support values can avoid too many large itemsets and uninteresting patterns.

Experiments were then made to show the relationship of the numbers of association rules and the minimum support values along with different minimum confidence values. Results are shown in Figure 2.

From Figure 2, it is easily seen that the numbers of association rules decreased along with the increase in minimum support values. This is also quite consistent with our intuition. Also, the curves for larger minimum confidence values were smoother than those for smaller minimum confidence values, meaning that the minimum support value had a large effect on the numbers of association rules derived from small minimum confidence values.

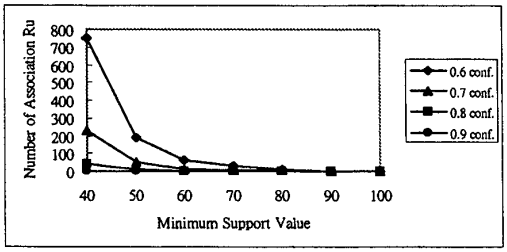

Figure 2. The relationship between numbers of association rules and minimum support values.

The relationship of the numbers of association rules and the minimum confidence values along with different minimum support values is shown in Figure 3.

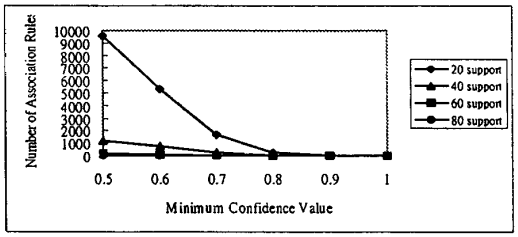

Figure 3. The relationship between numbers of association rules and minimum confidence values.

From Figure 3, it is easily seen that the numbers of association rules decreased along with an increase in minimum confidence values. This is quite consistent with our intuition. The curves for larger minimum support values were smoother than those for smaller minimum support values, meaning that the minimum confidence value had a larger effect on the number of association rules when smaller minimum support values were used. All of the various curves however converged to 0 as the minimum confidence value approached to 1.

In these experiments, the association rules mined out can actually be used to help the faculty in the Department of Information Management check the course programs and understand the students' learning interest and capability on the courses.

5. Conclusion and future work

In this paper, we have proposed a generalized data-mining algorithm, which can process transaction data with quantitative values and discover interesting patterns among them. The rules thus mined exhibit quantitative regularity in databases and can be used to provide some suggestions to appropriate supervisors. The proposed algorithm can also solve conventional transaction-data problems by using degraded membership functions. Experimental results with the students' scores in the Department of Information Management at I-Shou University, Taiwan, show the feasibility of the proposed mining

algorithm.

Although the proposed method works well in data mining for quantitative values, it is just a beginning. There is still much work to be done in this field. Our method assumes that the membership functions are known in advance. In [6, 7], we also proposed some fuzzy learning methods to automatically derive the membership functions. In the future, we will attempt to dynamically adjust the membership functions in the proposed mining algorithm to avoid the bottleneck of membership function acquisition. We will also attempt to design specific data-mining models for various problem domains.

References:

[1] R. Agrawal, T. Imielinksi and A. Swami, "Mining association rules between sets of items in large database," *The 1993 ACM SIGMOD Conference,* Washington DC, USA, 1993.

[2] R. Agrawal, T. Imielinksi and A. Swami, "Database mining: a performance perspective," *IEEE Transactions on Knowledge and Data Engineering,* Vol. 5, No. 6, 1993, pp. 914-925.

[3] R. Agrawal, R. Srikant and Q. Vu, "Mining association rules with item constraints," *The Third International Conference on Knowledge Discovery in Databases and Data Mining,* Newport Beach, California, August 1997.

[4] R. Agrawal and R. Srikant, "Fast algorithm for mining association rules," *The International Conference on Very Large Data Bases*, 1994, pp. 487-499.

[5] I. Graham and P. L. Jones, *Expert Systems – Knowledge, Uncertainty and Decision,* Chapman and Computing, Boston, 1988, pp.117-158.

[6] T. P. Hong and J. B. Chen, "Finding relevant attributes and membership functions," *Fuzzy Sets and Systems,* Vol.103, No. 3, 1999, pp. 389-404.

[7] T. P. Hong and C. Y. Lee, "Induction of fuzzy rules and membership functions from training examples," *Fuzzy Sets and Systems,* Vol. 84, 1996, pp. 33-47.

[8] A. Kandel, *Fuzzy Expert Systems*, CRC Press, Boca Raton, 1992, pp. 8-19.

[9] E. H. Mamdani, "Applications of fuzzy algorithms for control of simple dynamic plants, " *IEEE Proceedings*, 1974, pp. 1585-1588.

[10] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," *The 1996 ACM SIGMOD International Conference on Management of Data*, Monreal, Canada, June 1996, pp. 1-12.

[11] L. A. Zadeh, "Fuzzy logic," *IEEE Computer*, 1988, pp.83-93.