Finding Fuzzy Sets for the Mining of Fuzzy Association Rules for Numerical Attributes

Ada Wai-chee Fu, Man Hon Wong, Siu Chun Sze, Wai Chiu Wong, Wai Lun Wong and Wing Kwan Yu

Department of Computer Science and Engineering
The Chinese University of Hong Kong,
Shatin, Hong Kong

Abstract. Fuzzy association rules is introduced in [3]. However, the algorithms proposed by [3] for mining fuzzy association rules assumes that fuzzy sets are given. Here we propose a method to find the fuzzy sets based on clustering techniques. We have implemented our proposed method and showed that it is feasible and produces desirable results.

1 Introduction

Mining association rules is one of the important research problems in data mining [5]. The problem of boolean mining association rules over basket data was introduced in [2]. An example of such an association rule is the statement that 90% of the transactions that purchase bread and butter also purchase milk. The antecedent of this rule consists of bread and butter, and the consequent consists of milk alone. The number 90% is the confidence factor of the rule. There are many known algorithms for mining boolean association rule. For example, [1] has proposed Apriori, AprioriTID and AprioriHybrid algorithms for mining association rule. Moreover, a hash-based algorithm for mining association rule has been introduced in [4].

The problem of mining quantitative association rules is introduced in [7]. For illustration, the following is a table with three non-key attributes. Age and NumCars are quantitative attributes, whereas Married is a categorical attributes. A quantitative association rule present in this table is: $\langle Age : 30..39\rangle$ and $\langle Married : Yes\rangle \rightarrow \langle NumCars : 2\rangle$.

RecordID	Age	Married	NumCars
100	23	No	1
200	25	Yes	1
300	29	No	0

The algorithm proposed in [7] is to first partition the attribute domains into small intervals and combines adjacent intervals into larger one such that the combined intervals will have enough supports. Replacing the original attribute by its attribute-interval pairs, the quantitative problem can be transformed to a boolean one.

Although the current quantitative association rule mining algorithms can solve some of the problems introduced by quantitative attributes, they introduce some other problems. The first problem is caused by the sharp boundary between intervals. The algorithms either ignore or over-emphasize the elements near the boundary of the intervals in the mining process. The use of shape boundary interval is also not intuitive with respect to human perception. For example, the interval method may classify a person as young if age is less than 40 and old if age is greater than 40. This obviously does not correspond to the human perception of young and old. Furthermore, it is not easy to distinguish the degree of membership for the interval method. For instance, using the interval method, age of 45 and age of 90 will both be classified into old. However, we intuitively know that that age of 90 is much older than age of 45.

In [3], an algorithm for mining fuzzy association rule was proposed. The fuzzy association rule is of the form, “If X is A then Y is B”. As in the binary association rule, “X is A” is called the antecedent of the rule while “Y is B” is called the consequent of the rule. X, Y are sets of attributes of database and A, B are sets containing fuzzy sets which characterize X and Y respectively. With fuzzy sets, a person may be both a member of ”old” with 80% membership, and also a member of ”middle-age” with 20% membership. The membership function determines how much a data object belong to each fuzzy set defined for the numerical attributes. Fuzzy sets provide a smooth transition between member and non-member of a set. The fuzzy association rule is also easily understandable to human because of the linguistic terms associated with the fuzzy sets.

2 Finding Fuzzy Sets

The algorithm proposed in [3] for fuzzy association rule mining suffers from the following problem. The user or an expert must provide this algorithm the required fuzzy sets of the quantitative attributes and their corresponding membership functions. Moreover, the fuzzy sets and their corresponding membership functions provided by the experts may not be suitable for mining fuzzy association rules in the database. The quality of the results produced by the algorithm relies quite crucially on the appropriateness of the fuzzy sets to the given data. The problem for most applications is that it is very difficult to know a priori which fuzzy sets will be the most appropriate. Besides, it is unrealistic that experts can always provide the fuzzy sets of the quantitative attributes in the database for fuzzy association rule mining. To deal with these problems, we explore whether fuzzy sets can be determined automatically from the data. We intend to find the fuzzy sets for each quantitative attribute in a database by using clustering techniques. In addition, we will propose an algorithm to find the corresponding membership function for each fuzzy set discovered.

In this section, we describe the basic steps for finding fuzzy sets from a given database. The first step is to transform the database into one with the format required for finding clusters. The required format is that the value of all attributes (including boolean, categorical and quantitative attributes) should be positive integer. After this transformation, we can use a known clustering method to find the medoids of k clusters. By regarding the attribute, we can generate the fuzzy sets and the corresponding membership functions. The steps for finding fuzzy sets can be summarized as: (1) Transforming the original database, (2) finding the k medoids of the clusters in the database, (3) constructing fuzzy sets with the k medoids for each quantitative attribute, and (4) deriving the corresponding membership functions.

After converting the format of the given database into the required one, we can start to find the fuzzy sets. In order to apply the clustering method for finding fuzzy sets in a database, we consider the search space of a database with n attributes (including boolean, categorical and quantitative attributes) as an n-dimensional space. Moreover, the records of the database are regarded as the n-dimensional points or vectors distributed in the space. Now, we can use the clustering method to find the k medoids from the space (or database). In this project, we have proposed to use the CLARANS clustering algorithm introduced in [6] to find the k medoids because it is efficient for large data sets.

After we have obtained the k medoids of the database, we can use these medoids to classify each quantitative attribute into k fuzzy sets. Let $\{m_1, m_2, ..., m_k\}$ be the k medoids found from the database with n attributes, where $m_i = \{a_{i1}, a_{i2}, ..., a_{in}\}$ is the i-th medoid. Suppose we want to find the fuzzy sets for the j-th attribute which is quantitative with a range from min_j to max_j, then $\{a_{1j}, a_{2j}, ..., a_{kj}\}$ is the set of mid-points of the fuzzy sets for the j-th attribute. The k fuzzy sets will have ranges of $\{min_j - a_{2j}\}$, $\{a_{1j} - a_{3j}\}$,..., $\{a_{(i-1)j} - a_{(i+1)j}\}$,..., and $\{a_{(k-1)j} - max_j\}$.

To quote an example, suppose a database contains a quantitative attributes (Salary ranged from 4000 to 32000). Suppose we have found 3 medoids = $\{m_1, m_2, m_3\}$ from the database, where $m_1 = \{Salary = 7000, IQ = 100\}$, $m_2 = \{Salary = 10000, IQ = 120\}$, and $m_3 = \{Salary = 20000, IQ = 150\}$. Then we classify each attribute into 3 fuzzy sets shown in the following.

Fuzzy Set Label	Range	Mid-point
Low	4000 - 10000	7000
Medium	7000 - 20000	10000
High	10000 - 32000	20000

3 Generating the Membership Function

Next we describe how to generate corresponding membership function for each fuzzy set of an quantitative attribute. Let $\{m_1, m_2, ..., m_k\}$ be the k medoids found from the database with n attributes, where $m_i = \{a_{i1}, a_{i2}, ..., a_{in}\}$ is the i-th medoid. Suppose we want to find the membership functions of fuzzy sets for the j-th attribute which is quantitative with a range from min_j to max_j, and $\{a_{1j}, a_{2j}, ..., a_{kj}\}$ is the set of mid-points of each fuzzy set for the jth attribute, then we use the following method to find the required membership functions. For the fuzzy set with the mid-point a_{1j}, the membership function is given by

$$f_{1j}(x) = \begin{cases} 1.0 & \text{if } x \leq a_{1j} \\ \frac{x-a_{2j}}{a_{1j}-a_{2j}} & \text{if } a_{1j} < x < a_{2j} \\ 0 & \text{if } x \geq a_{2j} \end{cases}$$

For the fuzzy set with the mid-point a_{kj}, the membership function is given by

$$f_{kj}(x) = \begin{cases} 0 & \text{if } x \leq a_{1j} \\ \frac{x-a_{(k-1)j}}{a_{kj}-a_{(k-1)j}} & \text{if } a_{(k-1)j} < x < a_{kj} \\ 1.0 & \text{if } x \geq a_{2j} \end{cases}$$

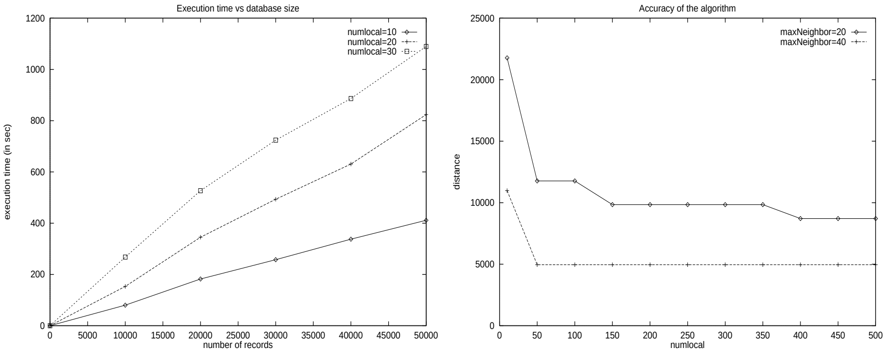
For the fuzzy set with the mid-point a_{ij}, where $2 \leq i \leq k-1$, the membership function is given by

$$f_{ij}(x) = \begin{cases} 0 & \text{if } x \leq a_{1j} \\ \frac{x-a_{(i-1)j}}{a_{ij}-a_{(i-1)j}} & \text{if } a_{(i-1)j} < x < a_{ij} \\ 1.0 & \text{if } x = a_{2j} \\ \frac{x-a_{(i+1)j}}{a_{ij}-a_{(i+1)j}} & \text{if } a_{ij} < x < a_{(i+1)j} \\ 0 & \text{if } x \geq a_{(i+1)j} \end{cases}$$

4 Experiments on the Clustering Method

First we want to determine if the clustering algorithm is efficient enough. We experimented with a database with three numerical attributes, and the values of the records are generated randomly. We have run the program with database size ranging from 10000 to 50000 records. The number of medoids is fixed at 3. We tried values of numLocal of 10, 20 ,30 (see [6]). We found that the execution time of all values of numLocal grows linearly as the number of records increases. The execution time required varies from 100 to about 1000 seconds Hence we find that the algorithm is sufficiently efficient for large databases.

Next we study the effects of different parameters on the accuracy. We have generated some clustered 2-Dimensional data with 3 known medoids, and apply the clustering method with varying parameters of numLocal and maxNeighbor (see [6]). We measured the sum of distances of the discovered medoids from the expected ones. Figure 1(b) shows that increase in the value of numLocal and maxNeighbor leads to decrease in the distance (i.e. increase in the accuracy). Also the result with numLocal = 50 is already very close to the minimum distance. We conclude that a reasonable value of numLocal and maxNeighbor is sufficient for good results.



(a) (b)

Figure 1: Experiment on the clustering algorithm

5 Experiments on Real Life Data

We apply our algorithm to a real life data set. After that, we use the identified fuzzy sets for mining fuzzy association rules and examine the interesting rules found. The data set comes from a research by the Hong Kong Institute of Asia-Pacific Studies. This research is proposed to investigate the point of view of the Hong Kong adults about the prospect of Hong Kong after 1997. This set of data contains 1000 records with 8 attributes (including boolean, categorical and quantitative). Boolean attributes include BORN IN HK and SEX, quantitative attributes include AGE, YEARS IN HK, and INCOME, categorical attributes incldue EDUCATION, WORK, and ASSESSMENT, where ASSESSMENT is the opinion of prospect of Hong Kong after 1997.

We want to find the fuzzy sets of the quantitative attributes. In order to find the fuzzy sets, we use the clustering program to find the k medoids of the database first. In this case, the k is 3 (i.e. there are three fuzzy sets for each quantitative attribute). The following table shows the 3 medoids found by the program (Remark : numLocal" = 30 and maxNeighbor" = 30).

	AGE	BIRTH	LIVE	EDU	WORK	INCOME	SEX	ASSESS
Medoid 1	28	1	20	3	0	9	1	0
Medoid 2	34	0	34	4	0	22	0	1
Medoid 3	56	1	45	1	0	4	0	1

After finding the medoids, we can create the fuzzy sets for each quantitative attribute by using these medoids. The ranges of fuzzy sets of AGE, LIVE and INCOME are shown in the following table.

AGE	Range	YEARS IN HK	Range	INCOME	Range
YOUNG	18 to 34	SHORT	0 to 34	LOW	0 to 9000
MEDIUM	28 to 56	MEDIUM	20 to 45	MEDIUM	4000 to 22000
OLD	34 to 98	LONG	34 to 98	HIGH	9000 to 43000

Next, we derive the membership functions for the fuzzy sets of each attribute by the algorithm mentioned before. We use the fuzzy set discovered above for mining fuzzy association rules in the database. We set the certainty threshold to be 0.5 and the significance threshold to be 0.3. Moreover, we restrict the consequent side of rules to contain only one attribute, ASSESSMENT, which is called the decision attribute. We use the algorithm given in [3]. We have chosen the definition of rules based on significance as described in [3]. The following shows some interesting rules. The certainty of the rules is equal to 1.

If {AGE *is* MEDIUM & EDUCATION *is*PRIMARY & WORK *is* UNEMPLOY & INCOME *is* LOW & SEX *is* MALE} then {ASSESSMENT *is* WORSE}

If {AGE *is* OLD & BORN IN HK & YEARS IN HK *is* MEDIUM & EDU *is* S_SEC & WORK *is* UNEMPLOY & INCOME *is* LOW & SEX *is* FEMALE} then {ASSESSMENT *is* BETTER}

6 Conclusion

For fuzzy association rule mining [3], previous work assume that the fuzzy sets are given. This assumption may not always be realistic. We propose a method to find the fuzzy sets from the data based on clustering techniques. From experiments, we show that the method produces meaningful results and has reasonable efficiency.

References

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, pages 487–499, 1994.
2. Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD, Washington, DC, USA*, pages 207–216, 1993.
3. C.M. Kuok, A. Fu, and M.H. Wong. Fuzzy association rules in large databases with quantitative attributes. In *ACM SIGMOD Records*, March, 1998.
4. J.S. Park, M-S. Chen, and P.S. Yu. An effective hash-based algorithm for mining association rules. In *Proceedings of ACM SIGMOD*, pages 175–186, 1995.
5. G. Piatesky-Sshapiro and W.J. Frawley. *Knowledge Discovery in Databases.* AAAI Press/The MIT Press, Menlo Park, California, 1991.
6. R.Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th VLDB Conference*, 1994.
7. Ramakrishnan Srikant and Rakesh Agrawal. Mining quantitative association rules in large relational tables. In *Proceedings of ACM SIGMOD*, pages 1–12, 1996.

This article was processed using the LaTeX macro package with LLNCS style