Fuzzy Association Rules Extraction Based on FCV Algorithm

Toshihiko Watanabe
Osaka Electro-Communication University
Neyagawa, Osaka, Japan
t-wata@isc.osakac.ac.jp

Hirokazu Takahashi
NEC Soft, Ltd.
Koto-ku, Tokyo, Japan
m04322@mail.osakac.ac.jp

Abstract— In order to develop a data mining system for huge database mainly composed of numerical attributes, there exists necessary process to decide valid quantization of the numerical attributes. Though the clustering algorithm can provide useful information for the quantization problem, it is difficult to formulate appropriate clusters for rule extraction in terms of cluster size and shape. In this paper, we study fuzzy association rules extraction method that can quantize the attributes by applying FCV clustering algorithm and extract rules simultaneously. From the results of numerical experiments using benchmark data, the method is found to be promising for actual applications.

I. INTRODUCTION

Recently, Data Mining to extract knowledge or rules from massive data sets stored in the database or dataware house, is studied for utilizing in various business scenes. As the promising applications of Data Mining, Association Rules [1-7] have been applied to various marketing problems. As for the other needs like in manufacturing area, there still exist many problems to cope with the stored valuable data, such as business decision, process improvement, and so on. In such domain, it seems that conventional approach such as computer assisted data analysis of stored data has been applied depending upon the human skill. Several reasons can be stated why the approach based on Data Mining methods is not applicable for such problems, one of the main reasons is that the mining system should deal with quantitative attributes appropriately.

In order to deal with the quantitative attributes in mining association rules, algorithms based on the generalized association rules that handle the continuous attributes as the Boolean vector by partitioning into several intervals are proposed [1,2]. Though several methods were also proposed to improve the computational time, the results applying the algorithms are still time consuming and are complicated to the user. Fuzzy association rules approaches [3-7] are proposed to overcome such disadvantages based on the fuzzy set concept. These approaches are based on the fuzzy extensions to the classical association rules mining by defining support and confidence of the fuzzy rule. Though the mining results are easy to understand by human operator, two drawbacks of applying such fuzzy methods to the actual problems still remain. One is the computational time for mining from database, and the other is accuracy of extracted rules.

The accuracy of extracted rules depends upon the fineness of the defined partition. In existing approaches, it is assumed that the input space is divided by grid-type fuzzy partitions in advance[4,5,7]. However, this assumption leads to deterioration of accuracy as well as to waste of computation time when the true fuzzy set exists just between the grids.

From these points of view, we proposed a quantitative association rule extraction algorithm based on clustering algorithm[12]. The algorithm extracts quantitative association rules along with simultaneous quantization based on clustering algorithm such as Fuzzy c-means(FCM). However some problems still remained such as improvement of computational efficiency compared with conventional association rules mining algorithm.

In this paper, we study fuzzy association rules extraction based on Fuzzy c-Varieties(FCV) algorithm that can quantize the numerical attributes by applying clustering and extract rules simultaneously. Results of numerical experiments using benchmark data are also shown.

II. EXTRACTION OF ASSOCIATION RULES FROM DATABASE

A. Extraction of Quantitative Association Rules

Assume that the database consists of numerical attributes. Let $X$ denote the set of numerical attributes(items) as:

$$X = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$$

(1)

$k$-th transaction data is defined as:

$$X^k = [x_1^k, x_2^k, \ldots, x_i^k, \ldots, x_n^k]^T, k = 1, 2, \ldots, m$$

(2)

where $m$ is the number of transactions in the database. As it is assumed that the database is composed of numerical attributes for simplicity, the transaction data also corresponds to $n$-dimensional input vector. In order to express discovering knowledge as the rule form, the item in (1) is transformed to quantized items in each attribute as:

$$Q = \left\{ C_{i,1}, C_{i,2}, \ldots, C_{i,f(i)}, C_{2,1}, C_{2,2}, \ldots, C_{2,f(2)}, \ldots \right\}$$

(3)

where $C_{i,j}$ denotes quantized item of $x_i$ and $f(i)$ denotes the number of fuzzy sets(fineness) in the partition of $x_i$. This transformation is performed by interval division or definition of fuzzy sets(fuzzy partition). Let $F$ denote the itemset which consists of items in (3). Support of the itemset $F$ is defined as:

$$s(F) = \frac{\sum_{i \in F} \mu_F(X_i)}{m}$$

(4)

where $\mu_F(X_i)$ denotes the membership value to the
quantized set $F$, i.e. multidimensional interval set or fuzzy set, calculated by the product operation of each membership value of each item in $F$. From the support value, confidence of the association rule $G \Rightarrow H$ is calculated by:

$$c(G \Rightarrow H) = s(G \cup H)/s(G)$$

(5)

where $G$ and $H$ are quantized itemsets. Association rules are extracted when the value of the rule is more than pre-defined minimal support and pre-defined minimal confidence. When we apply the mining algorithm to the actual huge problems, the support calculation is critical calculation concerning the number of query to the database. The itemset which has the value greater than predefined threshold is called “frequent itemset.” One of the main problems of mining quantitative association rules is how efficiently find the “frequent itemsets” from the database.

B. The Apriori Algorithm

The Apriori algorithm is essential and effective method for finding the frequent itemsets. The basic idea is that the frequent itemset should contain the subsets of frequent itemsets. Owing to this characteristic, frequent itemsets can be compounded from the smaller frequent itemsets one after another. Let $k$-itemset denote an itemset having $k$ items. Let $L_k$ represent the set of frequent $k$-itemsets, and $D_k$ the set of candidate $k$-itemsets. The algorithm to generate the frequent itemsets is as follows:

A1) $D_k$ is generated by joining the itemsets in $L_{k-1}$.

A2) The itemsets in $D_k$ which have some ($k$-I)-subset that is not in $L_{k-1}$ are deleted.

A3) The support of itemsets in $D_k$ is calculated through database scan to decide $L_k$.

After $L_1$ is decided first through database scan, the above A1-A3 procedures are iterated until $L_k$ becomes empty set. The association rules are decided by calculating the confidence of the rule combining the extracted frequent itemsets.

III. QUANTIZATION BASED ON CLUSTERING ALGORITHM

A. Quantization for Mining

In the above described extraction of quantitative association rule, the precision of extracted rules highly depends upon validness of quantization of numerical attributes. In traditional approach, the numerical attribute is usually quantized by means of definition of interval sets or fuzzy sets in advance. The necessary itemsets are generated by combining the quantized item in (3) together through the mining algorithm. In other words, multidimensional area for rule expression is formulated as the Cartesian product set of the 1-dimensional quantized sets. The clustering algorithm[8,13] is effective approach for such quantization process. In this paper, Fuzzy $c$-Varieties(FCV), Fuzzy $c$-Means(FCM), and hard clustering(HCM) are applied for the quantization.

B. Fuzzy $c$-Means Clustering

When the numerical attribute should be quantized by fuzzy partition, fuzzy clustering algorithm, i.e. fuzzy $c$-means, is effective. The regulation of the algorithm is as follows:

$$M_f = \{(U_a) | u_a \in [0,1], \sum_{i=1}^{m} u_{ak} = 1 \}$$

(6)

$$J_{sw}(U,V) = \sum_{i=1}^{m} \sum_{j=1}^{n} (u_{aj})^p \| X_i - v_j \|^2$$

(7)

where $v_i$ denotes the center of i-th cluster and $u_{ak}$ denotes the membership degree to i-th cluster. $p$ denotes the parameter for the degree of fuzziness.

C. Hard Clustering

In (7), “crisp clustering” is realized when the parameter $p$ is set as 1.0. The process is equivalent to $k$-means clustering algorithm. The quantization is expressed as the interval value.

D. Fuzzy $c$-Varieties Clustering

In addition to centroid for representative of each cluster, linear varieties are taken into consideration in FCV algorithm. Owing to this formulation, the shape of clusters becomes flexible as shown in Fig.1. The regulation of the algorithm is as follows:

$$J_{pl}(U,V,S) = \sum_{i=1}^{m} \sum_{j=1}^{n} (u_{aj})^p \| X_i - v_j \|^2$$

$$D_a(X_i,v_j,s_i) = \| X_i - v_j \|^2 - \sum_{i=1}^{m} \langle X_i - v_j, s_i \rangle^2$$

(8)

where $D$ denotes distance for clustering, $s_i$ is principal component vector of i-th cluster, and $\langle \rangle$ denotes inner product of vector. By using FCV algorithm for quantization of numerical attributes, more appropriate quantization is expected to be achieved for extraction of association rules.

IV. SYNCHRONOUS ALGORITHM OF ASSOCIATION RULE EXTRACTION AND CLUSTERING BASED QUANTIZATION

A. Basic Concept

Though the clustering algorithm is effective for quantization of numerical attribute, appropriate items should be selected for proper representation of area in $n$-dimensional space for the association rule. Instead of using the Cartesian product for multidimensional rule expression as described above, our idea is to use the clusters of multi-dimension for...
Let $T$ denote whole set of itemsets for clustering as:

$$T = 2^X - \emptyset$$

$$= \{\{x_1\}, \{x_2\}, \ldots, \{x_1, x_2\}, \{x_1, x_3\}, \ldots, \{x_1, x_2, x_3, \ldots, x_n\}\} \quad (9)$$

where $2^X$ denotes the power set of $X$, $\emptyset$ is empty set, and $q_i$ denotes the itemset in $T$. From (9), the quantization is performed using clustering algorithm with $q_i$ as the input variables. The quantized cluster is as follows:

$$Q = \{C_{q_1}, C_{q_2}, \ldots, C_{q_j}, \ldots\} \quad (10)$$

where $f(q_i)$ denotes the number of clusters calculated with input data represented as itemset $q_i$. For example, cluster set $\{C_{q_1}, C_{q_2}, C_{q_3}\}$ is generated by 2-dimensional clustering with inputs $\{x_1, x_3\}$ corresponding to itemset $q_i = \{x_1, x_3\}$ when $f(q_i)$ is 3. It is obvious that all clusters based on the itemsets(combination of input variables) in (10) cannot be calculated in actual huge database because of its combinatorial explosion.

The idea is to generate the clusters of appropriate dimension in turn through mining process. It is realized as a synchronous algorithm of association rule extraction and fuzzy clustering based quantization. Figure 2 shows the conceptual diagram of the algorithm. For reducing the database scan, $q_i$ is generated in turn like the apriori algorithm.

![Fig. 2. Conceptual Diagram of Proposed Algorithm](image)

### B. Algorithm

We define that the cluster is the “frequent cluster” when $C$ in (10) satisfies support restriction. It is also defined that the itemset $q_i$ is “frequent itemset” when frequent $w$-th cluster $C_{q_{i,w}}$ exists. The flowchart of the algorithm is shown in Fig.3. First, “Candidate 1-Itemsets” is set as $\{\{x_1\}, \{x_2\}, \ldots, \{x_n\}\}$. Then “1-dimensional clustering” procedures are employed based on the input variable corresponding to the “Candidate 1-Itemset”. The frequent 1-dimensional clusters are selected based on the support restriction. “Frequent 1-Itemsets” are selected whether there exists frequent cluster in “Frequent 1-dimensional clusters”. For example, 1-Itemset $q_i = \{x_1\}$ is frequent if there exist frequent cluster $C_{q_{i,1}}$ in “Frequent 1-dimensional clusters”. “Candidate 2-Itemsets” are generated by joining the itemsets in “Frequent 1-Itemsets”. 2-dimensional process is employed in the same manner. These procedures are iterated until “Frequent $n$-Itemsets” becomes empty set. It should be noted that a anti-monotone behavior of the support values of the clusters is not guaranteed in this algorithm. However, it is empirically expected that proper itemsets and clusters can be selected through the algorithm.

In the clustering procedure in the algorithm, the number of clusters should be decided appropriately. In order to save the computation, the numbers of clusters might be fixed reasonably in advance. In this study, threshold of the proportion of variance explained by the clusters is adopted as the necessary parameter in FCM and HCM. The number of clusters is decided as minimum number of clusters over the threshold value. As for FCV, variance of principal component scores is used in the same manner. For applying the number decision, several iterations are needed in each clustering process.

As shown in the figures, the quantization and the rule extraction are employed simultaneously. It is expected that the quantization and the extraction can be fulfilled in reduced computation time avoiding combinatorial explosion neatly.

![Fig. 3. Flowchart of Proposed Algorithm](image)
V. NUMERICAL EXPERIMENTS

A. Classification Problem

We develop a mining system based on the proposed algorithm and evaluate the performance through numerical experiments. In order to evaluate the precision of extracted rules, we apply the system to the classification problems such as “Iris”[9], “Wine”[9], “Liver”[10], and “Diabetes”[11]. Table 1 shows the parameter settings for numerical experiments, where “Decision coefficient” denotes the threshold parameter for decision of the number of clusters.

Let \( W = \{ e_1, e_2, \ldots, e_p \} \) denote class(output) set where \( p \) is the number of classes. The extracted \( i \)-th rule is expressed as:

\[
R^i : \text{IF } X^i \text{ belongs to } C_i \text{ THEN class is } E_i \text{ with } CV_i, \quad (11)
\]

where \( X^i \) is the input vector, \( C_i \) denotes the cluster of \( i \)-th rule, and \( E_i \) denotes the class of \( i \)-th rule. From the rules extracted as (11), the performance of classification is investigated. In order to evaluate the extracted rules, proper reasoning method should be defined. In this study, the reasoning output is calculated by the following two manners as:

\[
E^* = E_{\text{cl}}, \quad cl = \arg \max_k (u_k(X) \cdot CV_k) \quad (12)
\]

\[
E^* = e_{\text{cl}}, \quad cl = \arg \max \left( \sum_{j \in N(e_k)} u_j(X) \cdot CV_j \right) \quad (13)
\]

where \( u_k() \) denotes the membership value of the \( k \)-th rule, \( CV_k \) denotes the confidence value of the \( k \)-th rule, \( E^* \) is the output class, and \( N(e_k) \) denotes the set of rule numbers that consequent part is class \( e_k \). Eq.(12) corresponds to the output decision using the rule of maximum membership value. Eq.(13) corresponds to the output decision applying aggregation of the rules. These two different reasoning manners are used for the numerical experiments.

B. Evaluation of Basic Performance of Algorithm

We evaluate basic performance of the proposed algorithm compared with traditional “rule extraction after quantization” method. In the “rule extraction after quantization” method, every quantization is generated by clustering algorithm in advance. In other words, all of the quantized cluster in Eq.(10) are generated before starting mining. From the results of quantization, the association rules are extracted by mining algorithm described in Section.2. In the proposed algorithm, FCM is used for clustering. Results of evaluation are shown in Fig. 4. Where “MAX” stands for using (12) as the output reasoning, “MOD” stands for using (13) as the output reasoning, “(Revised)” denotes using \( CV \) as it is(extracted rule confidence value), and no entry of “(Revised)” denotes using \( CV \) as 1.0 for all extracted rules. From the results, the proposed algorithm decrease the number of extracted rules along with keeping good quality of classification compared with “rule extraction after quantization”.

Table 1. Parameter Settings for Mining Experiments

<table>
<thead>
<tr>
<th></th>
<th>Iris</th>
<th>Wine</th>
<th>Liver</th>
<th>Diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Minimal Support[%]</td>
<td>15</td>
<td>15</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>The Minimal Confidence[%]</td>
<td>75</td>
<td>75</td>
<td>70</td>
<td>75</td>
</tr>
<tr>
<td>Decision Coefficient</td>
<td>0.85</td>
<td>0.85</td>
<td>0.70</td>
<td>0.85</td>
</tr>
</tbody>
</table>

![Fig. 4. Basic Performance of Proposed Algorithm Based on FCM](image)
C. Results of Experiments by FCV Based Algorithm

The performance of the proposed FCV based algorithm is compared with the other clustering algorithms, i.e. FCM and HCM. Figures 5-8 show the number of itemsets/clusters through mining process and recognition rate in (d). Except for “Liver”, recognition rate of FCV can be considered to be approximately the same as the one of FCM. However, the computation time of FCV can be reduced compared with FCM, because the number of the candidate itemsets and the frequent itemsets is decreased especially at the second clustering steps. This characteristic can be expected to lead to a reduction of database scan in huge problem. From the results, the proposed FCV based algorithm has potential of efficient mining computation depending upon data distribution.
VI. CONCLUSION

In this paper, we proposed an algorithm for extraction of fuzzy association rules based on FCV algorithm that can quantize the numerical attribute by applying clustering algorithm and extract rules simultaneously. From the results of numerical experiments using benchmark data, the method is found to be promising compared with FCM based algorithm for data mining applications in the viewpoint of computational efficiency. In order to verify the evaluation definitely, we plan to apply the method to the other various datasets.

REFERENCES