Archives of Control Sciences
Volume 12(XLVIII), 2002
No. 4, *pages 379–395*

An Algorithm of granulation on numeric attributes for association rules mining

BEEN-CHIAN CHIEN, ZIN-LONG LIN, YI-XUE CHEN and TZUNG-PEI HONG

Mining association rules from numeric data is relatively more difficult than categorical data. The main reason is that the domain of real number lacks of the user's abstraction on reality. In this paper, we propose an algorithm to granulate numeric intervals automatically. The proposed method defines two threshold factors, information density-similarity and information closeness, to measure the condition if two granules should be merged and construct an abstraction hierarchy of intervals. For abstracting the best level of interval from the interval hierarchy automatically, we develop a determination function based on the threshold factors. After the intervals are determined, the fuzzy membership functions for each interval can be generated. Then an algorithm for mining fuzzy association rules can be used to mine qualified association rules from the fuzzy intervals.

Key words: data analysis, information granulation, data mining, fuzzy association rule, clustering

1. Introduction

Data mining is a key step in the processing of knowledge discovery. Many approaches like Apriori-based algorithms for mining association rules [1] and ID3 [18] for pattern classification have been proposed to handle knowledge discovery in symbolic attributes. For numeric attributes, it is relatively difficult for users to find a general solution to analyze numeric data effectively and efficiently. The simplest granulation for a numeric attribute is to partition the domain of attribute into several intervals. Existing interval partition techniques include Equal Width Interval (EWI) and Equal Frequency Interval (EFI). The EWI technique divides the range of observed values for the variable into n equal sized intervals. The EFI technique divides a continuous variable into n intervals and each interval contains the same number of instances. For the above two

B.-C. Chien, Z.-L. Lin and Y.-X. Chen are with Institute of Information Engineering I-Shou University 1, Section 1, Hsueh-Cheng Rd., Ta-Hsu Hsiang, Kaohsiung County, Taiwan, 840, R.O.C. T.-P. Hong is with Department of Electrical Engineering National University of Kaohsiung No. 251, Lane 280, Der-Chung Road, Nan-Tzu District Kaohsiung 811, Taiwan, tel: +886-7-6577711 ext. 6517 fax: +886-7-6578944 e-mail: cbc@isu.edu.tw

Received 2.12.2001, revised 25.04.2002.

methods, users must require the number of intervals and the number of instances in an interval usually.

Mining association rules from numeric attributes or quantitative data using interval partition has been discussed widely in previous researches. Srikant *et al*. [19] use equi-depth partitioning (EWI) to mine quantitative rules. However, they do not consider the relative distance between values and density of an interval. Miller and Yang [14] applied Birch clustering [23] to identify intervals and proposed distance-based association rules to improve the semantics of intervals. At the same time, ARCS [13] proposed by Lent *et al*. present a geometric-based algorithm, BitOP, for performing the clustering in numeric attributes. They show that clustering is a possible solution to granulate the meaningful regions and support the discovery of association rules. Although clustering provides a useful technique to discriminate granulation, it is not feasible for all clustering algorithms. Cheng *et al*. [4] suppose that most of the clustering algorithms, such as DBSCAN [6], BIRCH [23], CURE [7], and STING [20], do not satisfy the requirement of identifying clusters embedded in subspace of multi-attributes. Thus, they develop an entropy-based subspace clustering method called ENCLUS to handle the mining of subspace in numeric attributes.

Another possible granular approach for discovering association rules is fuzzy set theory. In contrast with quantitative clustering, fuzzy linguistic-based approaches focus on the qualitative filtering. Yager [22] introduced fuzzy linguistic summaries to provide summarization on different attributes. Hirota and Pedrycz [8][15] proposed a context-sensitive fuzzy clustering method based on Fuzzy C-means to construct rule-based models. However, the context-sensitive FCM method cannot deal with the data consisting of both numerical and categorical attributes. For solving the qualitative knowledge discovery from relational databases with numerical and categorical attributes, Au and Chan employed fuzzy linguistic terms to propose the F-APACS method [3]. Hong *et al*. [10] gave definitions of the support and confidence for fuzzy membership grade and designed a data mining approach based on fuzzy sets to find association rules with linguistic terms of human knowledge. However, the specified fuzzy linguistic terms in the above methods could be given only if we knew the properties of attributes. In real life, unfortunately, contents of attributes may be unknown and meaningful intervals are often not concise and crisp enough.

In this paper, we propose an algorithm to cluster numeric data into meaningful granulation and discover fuzzy association rules. The proposed method is based on the two threshold factors: information density-similarity and information closeness. For two adjacent granules C_1 and C_2, we merge them into a single granule C' if the threshold value of C' is close to the values in original granules of C_1 and C_2. We develop a measurement function and a agglomerative algorithm to produce the interval hierarchy recursively. We also give a determination function to decide the best granulation from the interval hierarchy of an attribute. The membership grade of fuzzy interval then can be derived to mine fuzzy association rules can be generated.

This paper is organized as follows. In Section 2, we give the definitions of fuzzy association rules. Then, we introduce the idea and have a formal description for the

proposed method in Section 3. The experimental results of the proposed method with different parameters are shown in Section 4. Finally, Section 5 concludes a summary and gives some directions of future research.

2. Fuzzy association rules

We first give formal definitions for fuzzy association rules. Let $R = \{A_1, A_2, \ldots, A_m\}$ be a set of attributes in a relational database, where $|R| = m$ is the number of attributes in R. Let $t[A]$ stand for the tuple values with the set of attributes A in R. Assume that A_i is an attribute with numeric domain D_{A_i}, $1 \leqslant i \leqslant m$. Let $I^k_{A_i} = [l_{ik}, u_{ik}]$ be the kth interval of the attribute A_i, where $l_{ik} \in D_{A_i}$, $u_{ik} \in D_{A_i}$, and $l_{ik} \leqslant u_{ik}$. An item value of attribute A_i is denoted as $t[A_i]$. Let C_{A_i} denote the condition $t[A_i] \in I^k_{A_i}$, C_X and C_Y be the conjunctions of conditions C_{A_i}, $1 \leqslant i \leqslant m$. The sets of X and Y represent the sets of attributes in the conjunctions of conditions C_X and C_Y, respectively. The number of tuples in the database satisfying the conditions C_X and C_Y are respectively denoted as $|C_X|$ and $|C_Y|$. A quantitative association rule over a set of crisp intervals is defined as follows:

Definition 1: [14] *A quantitative association rule is an implication of the form $C_X \Rightarrow C_Y$, where, $X \subset R$, $Y \subset R$, and $X \cap Y = \emptyset$. A rule $C_X \Rightarrow C_Y$ holds with confidence c and support s if $|C_X \wedge C_Y|/|CX| \geqslant c$, and $|C_X \wedge C_Y|/n \geqslant s$.*

In Definition 1, the conditions in association rules are based on crisp interval data. For unknown numeric values, intervals may not be clear enough in general. Therefore, we define the concept of fuzzy intervals as follows. Assume that a set of fuzzy intervals sets denoted as $\{A_{i1}, A_{i2}, \ldots A_{iK}\}$ is given for the attribute A_i. The membership function, $\mu_{A_{ik}}$, for the kth fuzzy interval in attribute A_i is defined as

$$\mu_{A_{ik}} \ : \ D_{A_i} \rightarrow [0, 1].$$

The fuzzy interval A_{ik}, $k = 1, 2, \ldots, K$ are defined as

$$A_{ik} = \sum_{D_{A_i}} \mu_{A_{ik}} \left(t[A_i]\right) / t[A_i], \text{ for all } t[Ai] \in D_{A_i}.$$

If $\mu_{A_{ik}}(t[Ai]) = 1$, the item $t[A_i]$ belongs to the fuzzy interval A_{ik} certainly. If $\mu_{A_{ik}}(t[Ai]) = 0$, it is no doubt that the item $t[A_i]$ are not in the fuzzy interval A_{ik}. Let I_X and I_Y be the conjunctions of fuzzy intervals in the attributes sets of X and Y. Based on fuzzy interval, we define fuzzy association rules over a set of fuzzy intervals as follows:

Definition 2: *A fuzzy association rule is an implication of the form* $I_X \Rightarrow I_Y$*, where,* $X \subset R$*,* $Y \subset R$*, and* $X \cap Y = \emptyset$*. A rule* $I_X \Rightarrow I_Y$ *holds with confidence* c *and support* s *if*

$$\frac{\sum_{t \in t[R]} \min_{A_i \in X, B_i \in Y} \left\{ \max_{A_{ik} \subseteq A_i} \{\mu_{A_{ik}}(t[A_i])\}, \max_{B_{ik} \subseteq B_i} \{\mu_{B_{ik}}(t[B_i])\} \right\}}{\sum_{t \in t[R]} \min_{A_i \in X} \left\{ \max_{A_{ik} \subseteq A_i} \{\mu_{A_i}(t[A_i])\} \right\}} \geqslant c,$$

$$\frac{\sum_{t \in t[R]} \min_{A_i \in X, B_i \in Y} \left\{ \max_{A_{ik} \subseteq A_i} \{\mu_{A_{ik}}(t[A_i])\}, \max_{B_{ik} \subseteq B_i} \{\mu_{B_{ik}}(t[B_i])\} \right\}}{n} \geqslant s.$$

For mining fuzzy association rules, as the flow shows in Fig. 1, the approach of fuzzy interval analysis for numeric attributes is the first significant step. A successful partition not only present the implicit characteristics in an attribute but also improve the linguistic interpretation in association rules. The interval granulation algorithm proposed in Section 3 will give an efficient approach for automatic generation of fuzzy intervals. After meaningful fuzzy intervals are found, the corresponding membership functions will be assigned to the intervals. The fuzzy mining algorithms then can be applied to discover the fuzzy association rules.

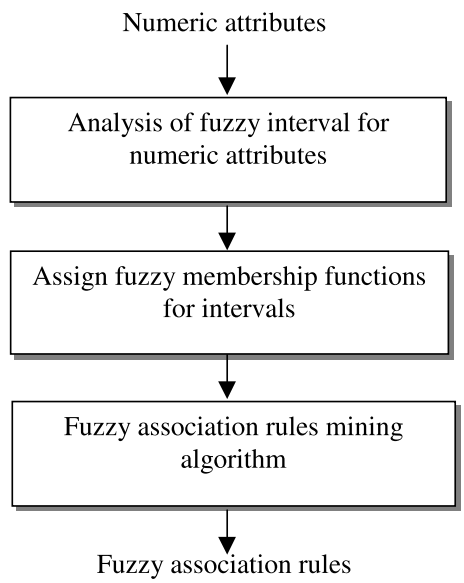

Figure 1. The flow for mining fuzzy association rules

3. The proposed algorithm

In the section, we present the proposed algorithm for interval granulation formally. The symbols used in the algorithm are defined in Section 3.1. The proposed algorithm is described in Section 3.2. Section 3.3 depicts the assignment of fuzzy membership functions for the granular results of fuzzy intervals.

3.1. Notations and evaluation of discrimination

Since we only consider the algorithm on a single attribute here, each interval on an attribute can be treated as a granule. The symbols used in our proposed method are defined in the following.

n : The total number of tuples in the original dataset.
r : The initial number of intervals on an attribute.
k : The current level of a granular hierarchy, $k = 0$ for initial level.
N^k : The number of intervals at level k, $N^0 = r$.
C_i^k : The ith interval at level k.
(i.e. The subset of dataset which are contained in ith interval at level k.)
h_i^k : The number of data contained in the ith intervals at level k.
$x_{i,j}^k$: The data contained in the ith intervals at level k.
(i.e. $C_i^k = \{x_{i,0}^k, x_{i,1}^k, \cdots, x_{i,h_i^k-1}^k\}$)
l_i^k : The length of the ith intervals at level k.
Ψ_i^k : The centroid of ith interval at level k,

$$\Psi_i^k = \frac{\sum_{j=0}^{h_i^k-1} x_{i,j}^k}{h_i^k} \tag{1}$$

$\delta_{i,i+1}^k$: The difference between the density of interval C_i^k and the density of the interval merged by neighbor intervals of C_i^k and C_{i+1}^k. We have

$$\delta_{i,i+1}^k = \left| \frac{h_i^k + h_{i+1}^k}{l_i^k + k_{i+1}^k} - \frac{h_i^k}{l_i^k} \right| \quad \text{and} \quad \delta_{i+1,i}^k = \left| \frac{h_i^k + h_{i+1}^k}{l_i^k + k_{i+1}^k} - \frac{h_{i+1}^k}{l_{i+1}^k} \right| \tag{2}$$

$\sigma_{i,i+1}^k$: The distance of the centroid of interval C_i^k moved after the merging of neighbor intervals C_i^k and C_{i+1}^k.

$$\sigma_{i,i+1}^k = \left| \Psi'^k_{i,i+1} - \Psi_i^k \right|, \quad \sigma_{i+1,i}^k = \left| \Psi'^k_{i,i+1} - \Psi_{i+1}^k \right|, \quad \text{where } \Psi'^k_{i,i+1} = \left| \frac{h_i^k \Psi_i^k + h_{i+1}^k \Psi_{i+1}^k}{h_i^k + h_{i+1}^k} \right|. \tag{3}$$

$DS_{i,i+1}^k$: Information density-similarity, the degree of density-similarity between two neighbor intervals C_i^k and C_{i+1}^k,

$$DS^k_{i,i+1} = \frac{h^k_i \delta^k_{i,i+1} + h^k_{i+1}\delta^k_{i+1,i}}{h^k_i + h^k_{i+1}} \tag{4}$$

$IC^k_{i,i+1}$: Information closeness, the degree of information closeness between two neighbor intervals C^k_i and C^k_{i+1},

$$IC^k_{i,i+1} = \frac{h^k_i \sigma^k_{i,i+1} + h^k_{i+1}\sigma^k_{i+1,i}}{h^k_i + h^k_{i+1}} \tag{5}$$

$E^k_{i,i+1}$: The threshold value is used to evaluate if C^k_i and C^k_{i+1} should be merged.

$$E^k_{i,i+1} = IC^k_{i,i+1} \times \left(DS^k_{i,i+1}\right)^{\alpha} \tag{6}$$

where α is the parameter used for adjusting the importance between $DS^k_{i,i+1}$ and $IC^k_{i,i+1}$, $\alpha > 0$.

U^k : The set of all N^k intervals at level k after k merging processes. Let $U^k = \{C^k_0, C^k_1, \cdots, C^k_{N^k-1}\}$.

ξ^k : The information density-similarity variance of U^k, the h^k_i is the number of data in interval C^k_i at level k,

$$\xi^k = \sum_{i=0}^{N^k} \left| \frac{n}{r} - \frac{h^k_i}{l^k_i} \right|. \tag{7}$$

η^k : The information closeness variance of U^k, the h^0_j is the number of data in initial interval C^k_j,

$$\eta^k = \sum_{i=0}^{N^k} \sum_{C^0_j \subseteq C^0_i} \left| \frac{h^k_i}{k^k_i} - \frac{h^0_j}{l^0_j} \right|. \tag{8}$$

Example 1. Assume that there are three intervals initially, $U^0 = \{C^0_0, C^0_1, C^0_2\}$ and $C^0_0 = \{1.8, 2.2\}$, $C^0_1 = \{3.2, 3.1, 3.0, 3.3, 3.4\}$, $C^0_2 = \{3.9, 3.8, 4.0, 4.2, 4.1, 4.0\}$. Suppose that $l^0_0 = l^0_1 = l^0_2 = 1$, we have $h^0_0 = 2$, $h^0_1 = 5$, $h^0_2 = 6$, and $\Psi^0_0 = 2$, $\Psi^0_1 = 3.2$, $\Psi^0_2 = 4$ as shown in Fig. 2.

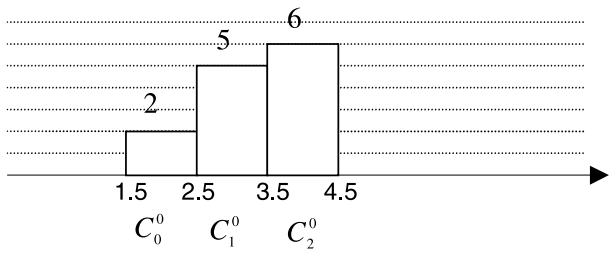

Figure 2. An example of three intervals

For granulating the intervals, we calculate the threshold value between C_0^0 and C_1^0 as follows:

$$\delta_{0,1}^0 = \left|\frac{h_0^0 + h_1^0}{l_0^0 + l_1^0} - \frac{h_0^0}{l_0^0}\right| = \left|\frac{2+5}{1+1} - \frac{2}{1}\right| = 1.5 \text{ and } \delta_{0,1}^0 = \left|\frac{2+5}{1+1} - \frac{5}{1}\right| = 1.5,$$

$$\Psi'^0_{0,1} = \left|\frac{h_0^0\Psi_0^0 + h_1^0\Psi_1^0}{h_0^0 + h_1^0}\right| = \frac{2\cdot 2 + 5\cdot 3.2}{2+5} \cong 2.857,$$

$$\sigma_{0,1}^0 = \left|\Psi'^0_{0,1} - \Psi_0^0\right| = |2.857 - 2| = 0.857 \text{ and } \sigma_{1,0}^0 = |2.857 - 3.2| = 0.343.$$

Thus, the information density-similarity and information closeness are

$$DS_{0,1}^0 = \frac{h_0^0\delta_{0,1}^0 + h_1^0\delta_{1,0}^0}{h_0^0 + h_1^0} = \frac{2\cdot 1.5 + 5\cdot 1.5}{2+5} = 1.5,$$

$$IC_{0,1}^0 = \frac{h_0^0\sigma_{0,1}^0 + h_1^0\sigma_{1,0}^0}{h_0^0 + h_1^0} = \frac{2\cdot 0.857 + 5\cdot 0.343}{2+5} = 0.490.$$

Let $\alpha = 2$, the threshold value for C_1^0 and C_2^0 is

$$E_{0,1}^0 = IC_{0,1}^0 \cdot (DS_{0,1}^0)^2 = 0.490 \cdot 1.5^2 = 1.1025.$$

Similarly, the threshold value for C_1^0 and C_2^0 can be calculated by

$$\delta_{1,2}^0 = \left|\frac{5+6}{1+1} - \frac{5}{1}\right| = 0.5 \text{ and } \delta_{2,1}^0 = \left|\frac{5+6}{1+1} - \frac{6}{1}\right| = 0.5,$$

$$\Psi'^0_{1,2} = \frac{5\cdot 3.2 + 6\cdot 4}{5+6} \cong 3.636,$$

$$\sigma_{1,2}^0 = |3.636 - 3.2| = 0.436 \text{ and } \sigma_{2,1}^0 = |3.636 - 4| = 0.364.$$

$$DS_{1,2}^0 = \frac{h_1^0\delta_{1,2}^0 + h_2^0\delta_{2,1}^0}{h_1^0 + h_2^0} = \frac{5\cdot 0.5 + 6\cdot 0.5}{5+6} = 0.5, \text{ and}$$

$$IC_{1,2}^0 = \frac{h_1^0\sigma_{1,2}^0 + h_2^0\sigma_{2,1}^0}{h_1^0 + h_2^0} = \frac{5\cdot 0.436 + 6\cdot 0.364}{5+6} = 0.397.$$

The threshold value for C_1^0 , C_2^0 and $\alpha = 2$ is

$$E_{1,2}^0 = IC_{1,2}^0 \cdot (DS_{1,2}^0)^2 = 0.397 \cdot 0.5^2 = 0.099.$$

Obviously, the threshold value $E_{0,1}^0$ for C_0^0 and C_1^0 is larger than the threshold value $E_{0,1}^0$ for C_1^0 and C_2^0. It means that the adjacent intervals with minimum $E_{i,i+1}^k$ should be granulated into an interval if the user needs a higher abstraction for their purpose. After merging C_1^0 and C_2^0 at level 0, the granulation of level 1 is generated. We obtain the new set of intervals $U^1 = \{C_0^1, C_1^1\}$, where $C_0^1 = C_0^0$ and $C_1^1 = C_1^0 \cap C_2^0$.

3.2. The algorithm for granulating intervals

For starting the algorithm of intervals granulation, the original numeric attribute must be divided into equal minimum intervals firstly. The minimum interval is the lowest abstraction level for the user's concept. It could be set to be the minimum distance among the data or be specified by the user for a specific application. Once upon the minimum interval was determined, the domain of the attribute has been granulated into r intervals initially. Then, we compute the threshold values for all neighbor intervals and merge the intervals with minimum $E^k_{i,i+1}$. Of course, the number of intervals will be decreasing and the size of merged intervals will be increasing after each merging process. At this time, we only re-compute the threshold values of intervals that are merged at last level and their neighbors again at the new level. Such merging processes are repeated until remaining only one interval. When the process of merging is stop, we find the granular level k with $\min_k\{\beta \cdot \xi^k + (1-\beta) \cdot \eta^k\}$, where β is the weight for the two evaluators ξ^k and η^k, $0 \leqslant \beta \leqslant 1$. The set of U^k is the best abstraction of interval partitions. The detailed algorithm is shown in the following.

Algorithm: Interval granulation

Input: A data set $t[A]$ of attribute A, and α, β.

Output: The interval set U^k.

Step 1: Initialize the original data set $t[A]$ into r intervals with the same length of interval l^0_i and compute h^0_i. Initially $k = 0$, $N^0 = r$, and $U^k = \{C^k_0, C^k_1, \cdots, C^k_{N^k-1}\}$.

Step 2: Compute ξ^k and η^k.

Step 3: Compute $E^k_{i,i+1}$, for $0 \leqslant i \leqslant (N^0 - 2)$.

Step 4: Find C^k_i with $\min\limits_{0 \leqslant i \leqslant N^k} E^k_{i,i+1}$, then merge C^k_i and $C^k_{i,i+1}$, $N^{k+1} = N^k - 1$.

Step 5: If $N^{k+1} > i$, we generate U^{k+1} from U^k, $k = k + 1$, and go to Step 2; otherwise, go to Step 6.

Step 6: Output U^k with $\min\limits_k\{\beta \cdot \xi^k + (1-\beta) \cdot \eta^k\}$.

The parameters, ξ^k and η^k, are used to determine the best abstraction of interval partitions. The parameter ξ^k is used to estimate the density-similarity among intervals. The parameter η^k is used to estimate the information closeness within an interval. While the summation of the two parameters has the smallest value, we claim that there is

the best abstraction at this time. In the traditional quantitative method, users need to give the number of partitions. However, it is difficult for users to assign the number of partitions when the characteristics of data are unknown. The proposed method estimates the variance between inter-cluster and intra-cluster and determines the best number of abstraction automatically.

Example 2. We give an example to explain the work of the proposed algorithm in the following. Assume that there are 127 numeric data in an attribute A as given in Table 1. The domain of the attribute A is in $[1.0, 11.0]$. The initial length of an interval is set to be 1 and thus there are 10 intervals initially. Further, let the parameters $\alpha = 2$ and $\beta = 0.5$. We start the step of granulation in the following

Table 1: Data set for Example 2.

$T\#$	A	$T\#$	A	$T\#$	A	$T\#$	A	$T\#$	A	$T\#$	A	$T\#$	A	$T\#$	A
1	6.2	17	9.4	33	10.2	49	4.3	65	4.3	81	4.5	97	10.0	113	1.6
2	3.1	18	1.3	34	7.4	50	10.5	66	8.2	82	10.7	98	4.3	114	9.5
3	5.4	19	6.6	35	8.3	51	8.3	67	5.7	83	5.7	99	6.4	115	9.1
4	1.1	20	2.4	36	2.4	52	2.6	68	3.5	84	6.3	100	8.3	116	10.1
5	5.1	21	4.3	37	8.2	53	2.3	69	3.8	85	8.5	101	2.7	117	4.7
6	4.6	22	2.3	38	1.6	54	2.0	70	10.7	86	2.4	102	8.8	118	8.2
7	5.2	23	2.5	39	1.2	55	8.8	71	2.8	87	7.2	103	9.3	119	8.5
8	7.4	24	5.8	40	5.6	56	5.6	72	10.2	88	1.2	104	3.2	120	10.4
9	10.3	25	3.3	41	4.3	57	7.3	73	1.0	89	9.3	105	3.3	121	10.7
10	1.5	26	8.4	42	3.6	58	4.5	74	9.8	90	7.7	106	7.3	122	1.2
11	8.1	27	1.7	43	3.7	59	6.8	75	3.6	91	6.3	107	3.5	123	7.5
12	10.0	28	10.1	44	3.1	60	9.2	76	8.8	92	9.2	108	8.3	124	3.4
13	4.7	29	3.3	45	9.0	61	3.5	77	2.1	93	7.4	109	2.6	125	3.5
14	10.3	30	10.9	46	10.2	62	10.3	78	3.2	94	10.2	110	9.8	126	3.3
15	4.4	31	4.5	47	4.2	63	9.2	79	9.1	95	8.3	111	8.2	127	4.2
16	4.2	32	10.1	48	9.9	64	9.1	80	9.1	96	7.4	112	2.0		

Input: The dataset Table 1, we have $r = 10$, $\alpha = 2$, $\beta = 0.5$.

Output: The interval set U^k.

Step 1: Quantize data set $t[A]$ into 10 intervals with initial length of interval $l_i^0 = 1$, and compute h_i^0 where i$= 0, \ldots, 9$, $N^0 = 10$, $k = 0$, $U^0 = \{C_0^0, C_1^0, \cdots C_9^0\}$ as shown in Fig. 3 and Table 2.

Table 2: Height of each initial interval.

i	0	1	2	3	4	5	6	7	8	9
h_i^0	10	13	17	15	8	6	9	16	15	18

Step 2: Compute ξ^k and η^k.

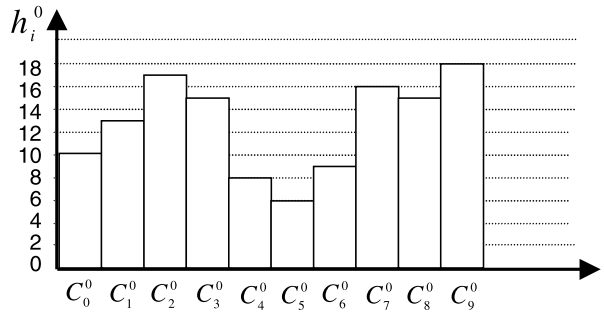

Figure 3. Initial intervals for Example 2

$$\begin{aligned}
\xi^k &= \sum_{i=0}^{N^k-1}\left|\frac{n}{r}-\frac{h_i^k}{l_i^k}\right| = \sum_{i=0}^{9}\left|\frac{127}{10}-\frac{h_i^0}{l_i^0}\right| \\
\cdots &= \left|\frac{127}{10}-\frac{10}{1}\right|+\left|\frac{127}{10}-\frac{13}{1}\right|+\left|\frac{127}{10}-\frac{17}{1}\right|+\left|\frac{127}{10}-\frac{15}{1}\right|+\left|\frac{127}{10}-\frac{8}{1}\right| \\
&+ \left|\frac{127}{10}-\frac{6}{1}\right|+\left|\frac{127}{10}-\frac{9}{1}\right|+\left|\frac{127}{10}-\frac{16}{1}\right|+\left|\frac{127}{10}-\frac{15}{1}\right|+\left|\frac{127}{10}-\frac{18}{1}\right| = 35.6 \\
\eta^k &= \sum_{i=0}^{N^k-1}\sum_{C_j^0\in C_i^0}\left|\frac{h_i^k}{l_i^k}-\frac{h_j^0}{l_j^0}\right| = \sum_{i=0}^{9}\sum_{C_j^0\in C_i^0}\left|\frac{h_i^k}{l_i^k}-\frac{h_j^0}{l_j^0}\right| \\
\cdots &= \left|\frac{10}{1}-\frac{10}{1}\right|+\left|\frac{13}{1}-\frac{13}{1}\right|+\left|\frac{17}{1}-\frac{17}{1}\right|+\left|\frac{15}{1}-\frac{15}{1}\right|+\left|\frac{8}{1}-\frac{8}{1}\right| \\
&+ \left|\frac{6}{1}-\frac{6}{1}\right|+\left|\frac{9}{1}-\frac{9}{1}\right|+\left|\frac{16}{1}-\frac{16}{1}\right|+\left|\frac{15}{1}-\frac{15}{1}\right|+\left|\frac{18}{1}-\frac{18}{1}\right| = 0
\end{aligned}$$

Step 3: Compute $E_{i,i+1}^k$, for $0 \leqslant i \leqslant (N^0-2)$. We give the results of $k=0$ in Table 3.

Table 3: The threshold values of $E_{i,i+1}^k$ for $k=0$.

i	0	1	2	3	4	5	6	7	8
$IC_{i,i+1}^0$	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
$DS_{i,i+1}^0$	1.5	2.0	1.0	3.5	1.0	1.5	3.5	0.5	1.5
$E_{i,i+1}^0$	1.125	2.000	0.500	6.125	0.500	1.125	6.125	0.125	1.125

Step 4: Find C_i^0 with $\min\limits_{0 \leqslant i \leqslant N^0}\{E_{i,i+1}^0\}$, then merge C_i^0 and $C_{i,i+1}^0$, $N^1 = N^0 - 1$. As Table 3 shows, $E_{7,8}^0$ is minimum. Hence, the intervals C_7^0 and C_8^0 should be merged at level 0.

Step 5: If $N^1 > 1$, we generate U^1 from U^0: $U^{k+1} = U^1 = \{C_0^1, C_1^1, C_2^1, C_3^1, C_4^1, C_5^1, C_6^1, C_7^1, C_8^1\}$, where $C_0^1 = C_0^0$, $C_1^1 = C_1^0$, $C_2^1 = C_2^0$, $C_3^1 = C_3^0$, $C_4^1 = C_4^0$, $C_5^1 = C_5^0$, $C_6^1 = C_6^0$, $C_7^1 = C_7^0 \cap C_8^0$ $C_8^1 = C_9^0$; the interval set of level 1 is generated. Then, $k = k + 1 = 0 + 1 = 1$ and go to Step 2. After the procedures from Step 2 to Step 5 are repeated until $N^k = 1$, the hierarchy of merging process is shown as Fig. 4.

Step 6: We list the values of ξ^k and η^k for all level k as shown in Table 4 and Fig. 5. In Table 4, we find that when $k = 5$, we have the minimum of $(\beta \cdot \xi^k + (1 - \beta) \cdot \cdot \eta^k)$. The final result of granulation has five intervals, that is, $U^5 = \{C_0^5, C_1^5, C_2^5, C_3^5, C_4^5\}$, as shown in Fig. 6.

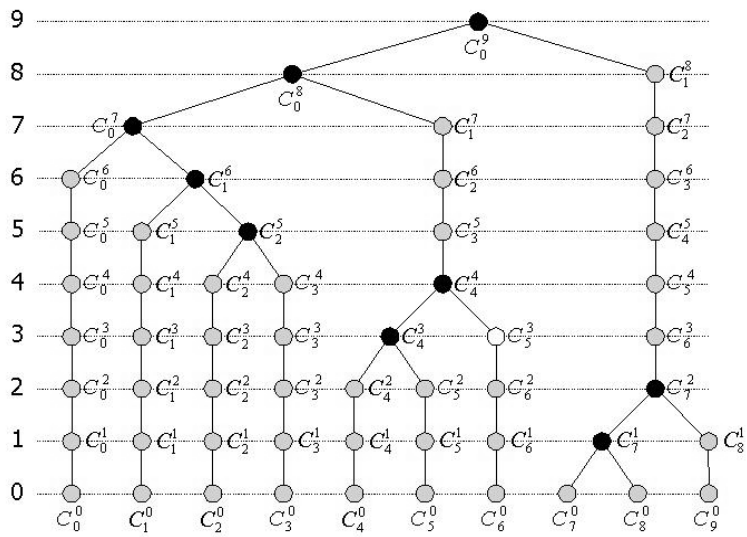

Figure 4. The abstraction hierarchy for merging process

Table 4: ξ^k, η^k and $\min\limits_k\{\beta \cdot \xi^k + (1 - \beta) \cdot \eta^k\}$ at all levels $(\beta = 0.5)$.

k	0	1	2	3	4	5	6	7	8	9
ξ^k	35.6	32.80	28.33	22.63	18.27	14.97	13.67	9.72	5.19	0.0
η^k	0.0	1.00	3.33	5.33	6.67	8.67	10.67	15.67	26.48	35.6
$\beta\xi^k + (1-\beta)\eta^k$	17.8	16.90	15.83	13.98	12.47	11.82	12.17	12.69	15.83	17.8

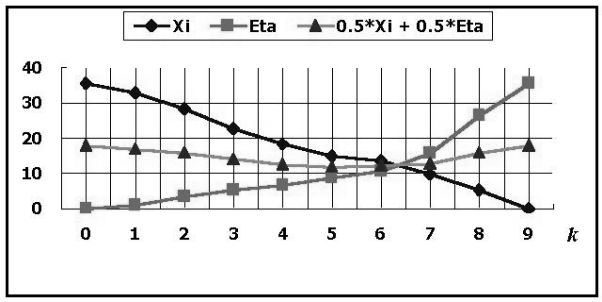

Figure 5. ξ^k, η^k and $\beta \cdot \xi^k + (1-\beta) \cdot \eta^k$, where $\beta = 0.5$

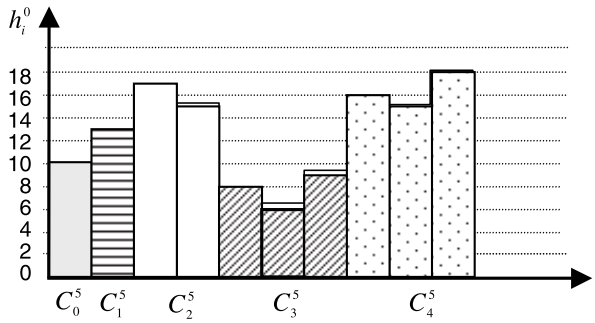

Figure 6. The interval abstraction for the proposed method

3.3. Assignment of membership functions for data mining

Since the ranges of intervals in a numeric attribute are continuous, as discussed in Section association rules from large numeric data. Here, we propose two efficient assignments of fuzzy membership functions based on the centroids of the partitioned intervals by the proposed algorithm of interval granulation.

The first type of assignment uses triangular membership function. Assume that $U^k = \{C_0^k, C_1^k, \cdots, C_{N^k}^k\}$ are the set of intervals abstracted from an attribute A_i by the proposed method, we map the fuzzy interval A_{ij} to C_{j-1}^k. A triangular membership function is assigned to each fuzzy interval by using the centroid of C_{j-1}^k to be the center of fuzzy interval. We have the following assignments:

$$\mu_{A_{i1}}(x) = \begin{cases} \frac{\Psi_0^k + x}{2\Psi_0^k}, & \text{for } x \leqslant \Psi_0^k \\ \max\left\{0, 1 - \frac{|x - \Psi_0^k|}{|\Psi_1^k - \Psi_0^k|}\right\}, & \text{for } x > \Psi_0^k \end{cases}$$

$$\mu_{A_{i(j+1)}}(x) = \begin{cases} \max\left\{0, 1 - \frac{|\Psi_j^k - x|}{|\Psi_j^k - \Psi_{j-1}^k|}\right\}, & \text{for } x \leqslant \Psi_j^k \\ \max\left\{0, 1 - \frac{|x - \Psi_j^k|}{|\Psi_{j+1}^k - \Psi_j^k|}\right\}, & \text{for } x > \Psi_j^k \end{cases}$$

$$\mu_{A_{i(N^k+1)}}(x) = \begin{cases} \max\left\{0, 1 - \frac{|\Psi^k_{N^k} - x|}{|\Psi^k_{N^k} - \Psi^k_{N^k-1}|}\right\}, & \text{for } x \leqslant \Psi^k_{N^k} \\ \frac{(1-\Psi^k_{N^k})+(1-x)}{2(1-\Psi^k_{N^k})}, & \text{for } x > \Psi^k_{N^k} \end{cases}$$

where $x \in D_{A_i}$, A_{i1} is the first fuzzy interval, $A_{i(N^k+1}$ is the last interval and Ψ^k_j is the centroid of C^k_j as defined in Section 3.1. For example, in Example 2, we have the membership functions of U^k as shown in Fig. 7.

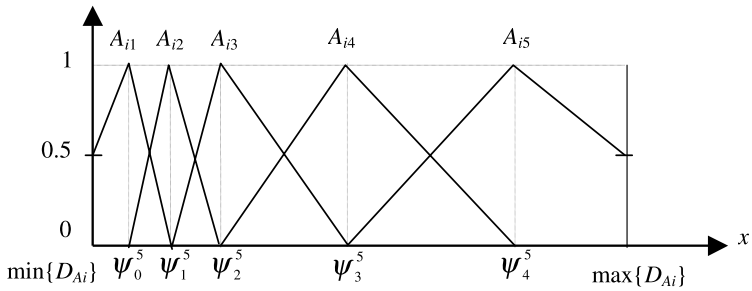

Figure 7. The membership functions for U^5 in Example 2

Another type of membership functions can be assigned by applying the Fuzzy C-mean model directly. That is,

$$\mu_{A_{i(j+1)}}(x) = \frac{1}{\sum_{i=0}^{N^k-1} \left(\frac{|x-\Psi^k_j|}{|x-\Psi^k_i|}\right)^2},$$

where $x \in D_{A_i}$ and Ψ^k_j is the centroid of C^k_j. In fact, the value of Ψ^k_j is very close to the center generated by FCM from empirical results. Since the two threshold factors, information density-similarity and information closeness, provide a way of calculation for interval center similar the C-Mean approach except that we always consider the neighboring interval with similar granular information.

After the membership functions are assigned to all numeric attributes, the mining algorithms for fuzzy association rules, like [3] and [10], can be used to generate the fuzzy association rules as defined in Section 2.

4. Experimental results

We made experiments in several types of numeric data with different distributions. The four data sets shown in Fig. 8, Fig. 9, Fig. 10 and Fig. 11 are quantized by $r = 256$. The numbers of data in the four data sets are all about 8,000 and the data are real numbers in the range of $[0, 640]$. We demonstrate the best partition results for the data set 1 in

Fig. 12, Fig. 13, and Fig. 14 with parameters $\alpha = 0$, $\alpha = 1.5$ and $\alpha = 2.5$, respectively, and the same $\beta = 0.5$.

Table 5 lists the number of granules generated automatically under different values of α in the four data sets. As Table 5 shows, the number of granules usually decreases while the value of α is getting large. However, the number of granules will be increasing again when the value of α is larger than a certain value. We may choose the value of α with the smallest number of granules as the possible abstraction. Let's look at Table 5, for example, the better choice of α in the data set 1 is $\alpha = 2.5$. The data set 2 is $\alpha = 2.5$, too. The data set 3 and data set 4 then had better pick the values of $\alpha = 0.5$ and $\alpha = 1.5$, respectively. The granular result of the data set 1 in Fig.14 is reasonable for our observation visually. For the data sets with different distributions, the value of α will not have the same choice for obtaining the best result. Generally speaking, the value of α is approximately equal to 1 for data governed by a uniform distribution.

Table 5: The number of granules for the test data sets.

α	0	0.5	1	1.5	2	2.5	3
Data set 1	49	25	25	22	20	18	18
Data set 2	90	58	49	37	29	17	44
Data set 3	67	48	49	49	56		
Data set 4	69	48	43	31	41		

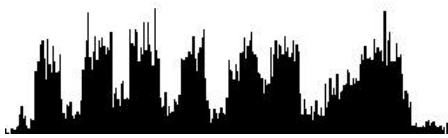

Figure 8. The data set 1

Figure 9. The data set 2

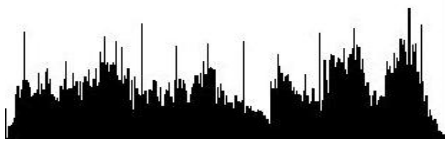

Figure 10. The data set 3

Figure 11. The data set 4

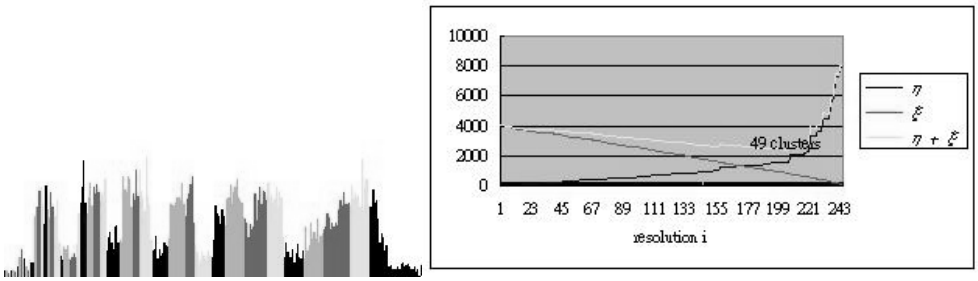

Figure 12. The best partition for data set 1, $r = 256$, $\alpha = 0$, $N^k = 48$

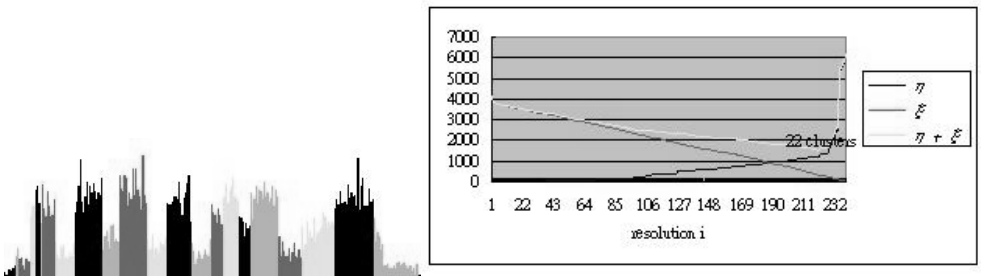

Figure 13. The best partition for data set 1, $r = 256$, $\alpha = 1.5$, $N^k = 21$

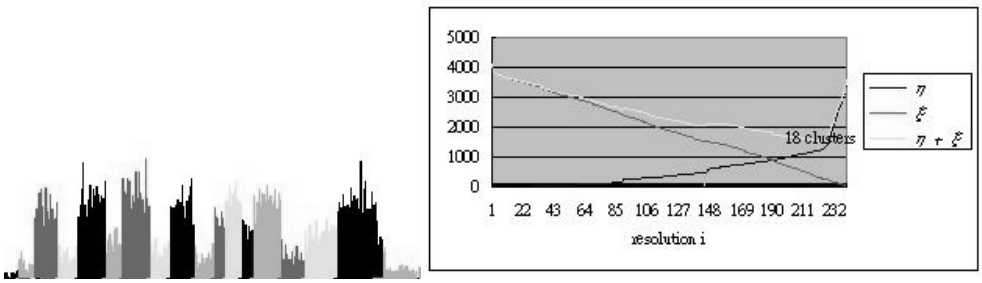

Figure 14. The best partition for data set 1, $r = 256$, $\alpha = 2.5$, $N^k = 17$

5. Conclusions

The data mining process from numerical attributes is more difficult for analysts than categorical data in general. The main obstacle is the lack of meaningful property in data and the vagueness of intervals. The contribution of this paper is to propose an automatic interval abstraction method and an efficient fuzzy membership function generator based on the concept of granular computing. By the proposed method, a set of qualified fuzzy association rules can be generated effectively by the fuzzy association rules mining algorithms [3][10]. According to the empirical results, the proposed approach of granulation may have a good approximation for the C-Mean approach. However, it needs more the-

oretical analysis and further discussion. Another future work is to extend the proposed method to the granulation of multi-dimensional numeric data features.

References

[1] R. AGRAWAL, T. IMIELINSKI and A. SWAMI: Mining Association Rules between Sets of Items in Large Databases. *Proc. ACM SIGMOD Int. Conf. on Management of Data*, Washington, (1993), 207-216.

[2] W.H. AU and K.C.C. CHAN: An Effective Algorithm for Discovering Fuzzy Rules in Relational Databases. *Proc. IEEE Int. Conf. on Fuzzy Systems*, (1998), 1314-1319.

[3] K.C.C. CHAN and W.H. AU: Mining Fuzzy Association Rules. *Proc. 6th ACM Int. Conf. on Information and Knowledge Management*, Las Vegas, (1997), 209-215.

[4] C.H. CHENG, A.W. FU and Y. ZHANG: Entropy-based Subspace Clustering for Mining Numerical Data. *Proc. Int. Conf. on Knowledge Discovery and Data Mining*, San Diego, USA, (1999), 84-93.

[5] B.C. CHIEN, Z.L. LIN and T.P. HONG: An Efficient Clustering Algorithm for Mining Fuzzy Quantitative Association Rules. *Proc. Int. Conf. IFSA*, (2001), 1306-1311.

[6] M. ESTER, H. KRIEGEL, J. SANDER and X. XU: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proc. 2nd Int. Conf. on Knowledge Discovery in Databases*, Menlo Park, USA, (1996), 226-231.

[7] S. GUHA, R. RASTOGI and K. SHIM: CURE: An Efficient Clustering Algorithm for Large Databases. *Proc. ACM SIGMOD Int. Conf. on Management of Data*, Seattle, USA, (1998), 73-84.

[8] K. HIROTA and W. PEDRYCZ: Linguistic Data Mining and Fuzzy Modeling. *IEEE Int. Conf. on Fuzzy Systems*, **2** (1996), 1488-1496.

[9] K. HIROTA and W. PEDRYCZ: Fuzzy Computing for Data Mining. *Proc. IEEE*, **87**(9), (1999), 1575-1600.

[10] T.P. HONG, C.S. KUO and S.C. CHI: A Data Mining Algorithm for Transaction Data with Quantitative Values. *Intelligent Data Analysis*, **3**(5), (1999), 363-376.

[11] G. KARYPIS, E.H. HAN and V.KUMAR, CHAMELEON: Hierarchical Clustering Using Dynamic Modeling. *IEEE Computer*, **32**(8), (1999), 68-75.

[12] L.H. LEE and L.K. HYUNG: An Extension of Association Rules Using Fuzzy Sets. *Proc. Int. Conf. IFSA*, (1997), 399-402.

[13] B. LENT, A. SWAMI and J. WIDOM: Clustering Association Rules. *Proc. IEEE Int. Conf. on Data Engineering*, (1997), 220-231.

[14] R.J. MILLER and Y. YANG: Association Rules over Interval Data. *Proc. ACM SIGMOD Int. Conf. on Management of Data*, AZ, USA, (1997), 452-461.

[15] W. PEDRYCZ: Fuzzy Set Technology in Knowledge Discovery. *Fuzzy Sets and Systems*, (1998), 279-290.

[16] W. PEDRYCZ: Granular Computing: An Introduction. *Proc. Int. Conf. IFSA*, (2001), 1349-1354.

[17] A. BARGIELA and W. PEDRYCZ: Classification and Clustering of Granular Data. *Proc. Int. Conf. IFSA*, (2001), 1696-1701.

[18] J.R. QUINLAN: Introduction of decision trees. *Machine Learning*, **1** (1986), 81-106.

[19] R. SRIKANT and R. AGRAWAL: Mining Quantitative Association Rules in Large Relational Tables. *Proc. ACM SIGMOD Int. Conf. on Management of Data*, Montreal, Canada, (1996), 1-12.

[20] W. WANG, J. YANG and R. MUNTS: STING: A Statistical Information Grid Approach to Spatial Data Mining. *Proc. 23rd Conf. on Very Large Data Bases*, Athens, Greece, (1997), 186-195.

[21] X. XU, M. ESTER, H.P. KRIEGEL and J. SANDER: A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases. *Proc. IEEE Int. Conf. on Data Engineering*, (1998), 324-331.

[22] R.R. YAGER: Fuzzy Summaries in Database Mining. *Proc. 11th Conf. on Artificial Intelligence for Application*, Los Angeles, USA, (1995), 265-269.

[23] T. ZHANG, R. RAMAKRISHNAN and M. LIVNY: BIRCH: An Efficient Data Clustering Method for Very Large Databases. *Proc. ACM SIGMOD Int. Conf. on Management of Data*, Montreal, Canada, (1996), 103-114.