Available online at www.sciencedirect.com
SCIENCE DIRECT®
Fuzzy Sets and Systems 152 (2005) 587–601
FUZZY sets and systems
www.elsevier.com/locate/fss

Genetic algorithm based framework for mining fuzzy association rules

M. Kaya[a], R. Alhajj[b,*]

[a] *Department of Computer Engineering, Firat University, 23119, Elaziğ, Turkey*
[b] *Department of Computer Science, Advanced Database Systems; Appl., University of Calgary, 2500 University Drive, Calgary - Alta., Canada T2N 1N4*

Received 27 February 2003; received in revised form 22 September 2004; accepted 29 September 2004
Available online 26 October 2004

Abstract

It is not an easy task to know a priori the most appropriate fuzzy sets that cover the domains of quantitative attributes for fuzzy association rules mining, simply because characteristics of quantitative data are in general unknown. Besides, it is unrealistic that the most appropriate fuzzy sets can always be provided by domain experts. Motivated by this, in this paper we propose an automated method for mining fuzzy association rules. For this purpose, we first present a genetic algorithm (GA) based clustering method that adjusts centroids of the clusters, which are to be handled later as midpoints of triangular membership functions. Next, we give a different method for generating the membership functions by using Clustering Using Representatives (CURE) clustering algorithm, which is known as one of the most efficient clustering algorithms described in the literature. Finally, we compared the proposed GA-based approach with other approaches from the literature. Experiments conducted on 100K transactions from the US census in the year 2000 show that the proposed method exhibits a good performance in terms of execution time and interesting fuzzy association rules.

Keywords: CURE clustering algorithm; Fuzzy sets; Data mining; Genetic algorithms; Quantitative attributes; Association rules

1. Introduction

Data mining is the process of extracting previously unknown and potentially useful hidden predictive information from large amounts of data. Discovering association rules is one of the several data mining

* Corresponding author. Tel.: +1 403 2109453; fax: +1 403 2844707.
E-mail address: alhajj@cpsc.ucalgary.ca (R. Alhajj).

0165-0114/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2004.09.014

techniques described in the literature. Associations allow capturing all possible rules that explain the presence of some items according to the presence of other items in the same transaction.

An association rule is defined in the form: $X \Rightarrow Y$, where both X and Y are defined as sets of attributes or items; it is interpreted as follows: "for a specified fraction of the existing transactions, particular values of the attributes in set X determine the values of the attributes in set Y as other particular values under a certain confidence". For instance, a binary association rule in a supermarket basket data may be expressed as: $butter \Rightarrow eggs$ [20%, 75%]. This rule is interpreted as follows: "20% of the people buy butter and eggs together in the same transaction; and 75% of the people buying butter also buy eggs in the same transaction". The two percentage values are referred to as *support* and *confidence*, respectively; these are the basic measures for the significance of an association rule. Simply, support is the percentage of transactions that contain both X and Y, while confidence is the ratio of the support of $X \cup Y$ to the support of X. So, the problem can be stated as: *find all association rules that satisfy user-specified minimum support and confidence*.

Early research in the field concentrated on boolean association rules, which are concerned only with whether an item is present in a transaction or not, without considering its quantity [1,2]. However, quantity is a very useful piece of information. Realizing the importance of quantity, people started to concentrate on quantitative attributes. The main reason for quantitative association rules mining is that numerical attributes typically contain many distinct values. The support for any particular value is likely to be low, while the support for intervals is much higher.

Although current quantitative association rules mining algorithms solved some of the problems particular to quantitative attributes, they introduced some other problems [22]. For instance, the major problem of Srikant and Agrawal's work is the sharp boundary between intervals. In other words, existing quantitative mining algorithms either ignore or over-emphasize elements near the boundary of an interval. The use of sharp boundary intervals is also not intuitive with respect to human perception as illustrated next.

Consider three intervals for the income of a person, namely [0, 30K], [30K, 70K], and [70K, 120K] to represent the three categories "poor", "moderate" and "rich", respectively. It is not easy to distinguish the degree of membership for the interval method. For instance, the interval method may classify a person as poor if annual income is less than or equal to \$30K and moderate if annual income is greater than \$30K. Further, using the interval method, incomes of \$70K and \$120K may both be classified into the rich category. However, it has intuitively been known that the income of \$120K is much richer than the income of \$70K. This problem can be handled smoothly by introducing fuzziness into the model as described in this paper.

As a remedy to the sharp boundary problem mentioned above, the fuzzy set concept has recently been used more frequently in mining quantitative association rules. Unlike classical set theory where membership is binary, the fuzzy set theory introduced by Zadeh [25] provides an excellent means to model the "fuzzy" boundaries of linguistic terms by introducing gradual membership. Some example linguistic terms include "poor", "young", "rich", "excellent", etc. Based on this and instead of using sharp boundary intervals, some work has recently been done on the use of fuzzy sets in discovering association rules for quantitative attributes e.g., [16,24]. However, in existing approaches fuzzy sets are either supplied by an expert or determined by applying an existing known clustering algorithm. The former is not realistic, in general, because it is extremely hard for an expert to specify fuzzy sets in a dynamic environment. On the other hand, approaches that applied classical clustering algorithms to decide on fuzzy sets have not produced satisfactory results. This is demonstrated by our research results

reported in [15], where we also showed that the fuzzy sets obtained by CURE-based approach exhibit better results than CLARANS-based approach.

To handle this problem, in this paper we present a GA-based method to derive the fuzzy sets from a set of given transactions. The method finds the optimum centroid points for a given number of clusters such that the membership functions generated using these points will extract the maximum number of large itemsets. The obtained membership functions are triangular and have a uniform structure. We then compared our approach with a CURE-based approach, which is identified as one of the most efficient clustering algorithms described in the literature. Finally, we considered some data from the United States census in the year 2000 and conducted some experiments to test our approach; the obtained results support the efficiency and effectiveness of the proposed method.

The rest of this paper is organized as follows. Related work is discussed in Section 2. Fuzzy quantitative association rules mining are defined in Section 3. The GA process to find fuzzy sets and their membership functions is described in Section 4. A brief overview of CURE clustering algorithm and the process of generating membership functions by using CURE are included in Section 5. Experimental results on 100K transactions from the United States census in year 2000 are given in Section 6. Section 7 is the conclusions.

2. Related work

Intervals may not be concise and meaningful enough for human experts to obtain nontrivial knowledge from quantitative association rules mining. On the other hand, fuzzy sets provide a smooth transition between members and non members of a set. Fuzzy association rules are also easily understandable to humans because of the linguistic terms associated with fuzzy sets. In addition to fuzziness, researchers proposed different approaches to overcome the interval sharp boundary problem.

Srikant and Agrawal [22] used equi-depth partitioning to mine quantitative rules. They separate intervals by their relative ordering and quantities equally. Miller and Yang [19] applied Birch clustering to identify intervals and proposed a distance-based association rules mining process, which improves the semantics of the intervals. Lent et al. [17] presented a geometric-based algorithm, called *BitOP*, to perform clustering for numerical attributes. They showed that clustering is a possible solution to figure out meaningful regions and support the discovery of association rules. Finally, some other researchers investigated the mining of weighted association rules [4,9,24]

Another trend to deal with the problem is based on fuzzy theory. In contrast to quantitative clustering, fuzzy linguistic-based approaches focus on qualitative filtering. Yager [23] introduced fuzzy linguistic summaries on different attributes. Hirota and Pedrycz [10,21] proposed a context sensitive fuzzy clustering method based on fuzzy C-means to construct rule-based models. However, the context-sensitive fuzzy C-means method cannot deal with the data consisting of both numerical and categorical attributes. To solve the qualitative knowledge discovery problem, Au and Chan [3] applied fuzzy linguistic terms to relational databases with numerical and categorical attributes. Later, they proposed the F-APACS method (1998) to discover fuzzy association rules. They utilized adjacent difference analysis and fuzziness in finding the minimum support and confidence values instead of having them supplied by a user. They determine both positive and negative associations. Hong et al. [11] proposed an algorithm that integrates fuzzy set concepts and Apriori mining algorithm to find interesting fuzzy association rules from given transactional data. In another paper, Hong et al. [12] proposed definitions for the support and confidence of

fuzzy membership grades and designed a data mining approach based on fuzzy sets to find association rules with linguistic terms of human knowledge. Ishibuchi et al. [13] illustrated fuzzy versions of confidence and support that can be used to evaluate each association rule. The authors employed these measures of fuzzy rules for function approximation and pattern classification problems. Gyenesei [8,9] presented two different methods for mining fuzzy quantitative association rules, namely *without normalization* and *with normalization*. The experiments of Gyenesei showed that the numbers of large itemsets and interesting rules found by the fuzzy method are larger than the discrete method defined by Srikant and Agrawal [22]. The approach developed by Zhang [26] extends the equi-depth partitioning with fuzzy terms. However, it assumes fuzzy terms as predefined.

Fu et al. [6] proposed an automated method to find fuzzy sets for the mining of fuzzy association rules. Their method is based on CLARANS clustering algorithm [20]. After obtaining the k medoids for each quantitative attribute, these medoids are used to classify each quantitative attribute into k fuzzy sets. On the other hand, in a previous part of our research, we used less centroids than the other approaches described in the literature, and the membership functions of the determined sets are adjusted accordingly [15].

However, the specified fuzzy linguistic terms in fuzzy association rules can be given only when the properties of the attributes are estimated. In real life, contents of columns (i.e., values of attributes) may be unknown and meaningful intervals are usually not concise and crisp enough. For this reason, some work has been done automatically determining the number and intervals of the clusters. Chien et al. [5] proposed an automated clustering algorithm based on variation of density to solve the interval partitioning problem. Recently, we presented a novel automated clustering method based on multi-objective GA [14], where the values of a given quantitative attribute are automatically clustered. So, the most appropriate number of clusters is determined automatically.

3. Fuzzy association rules

Let $T = \{t_1, t_2, \ldots, t_n\}$ be a database of transactions; each transaction t_j represents the jth tuple in T. We use $I = \{i_1, i_2, \ldots, i_m\}$ to represent all attributes (items) that appear in T; each attribute i_k may have a binary, categorical or quantitative underlying domain D_{i_k}. Besides, each quantitative attribute i_k is associated with at least two fuzzy sets. Explicitly, it is possible to define some fuzzy sets for attribute i_k with a membership function per fuzzy set such that each value of attribute i_k qualifies to be in one or more of the fuzzy sets specified for i_k. The degree of membership of each value of attribute i_k in any of its fuzzy sets is directly based on the evaluation of the membership function of the particular fuzzy set with the value of i_k as input.

So, given a database of transactions T, its set of attributes I, and the fuzzy sets associated with quantitative attributes in I. Note that each transaction t_j contains values of some attributes from I and each quantitative attribute in I has two or more corresponding fuzzy sets. The target is to find out some interesting and potentially useful regularities, i.e., fuzzy association rules with enough support and high confidence. We use the following form for fuzzy association rules [20].

Definition 3.1. A fuzzy association rule is expressed as

$$\text{If } X = \{x_1, x_2, \ldots, x_p\} \text{ is } A = \{f_1, f_2, \ldots, f_p\} \quad \text{then}$$
$$Y = \{y_1, y_2, \ldots, y_q\} \text{ is } B = \{g_1, g_2, \ldots, g_q\},$$

Here, X and Y are disjoint sets of attributes called itemsets, i.e., $X \subset I$, $Y \subset I$ and $X \bigcap Y = \phi$; A and B contain the fuzzy sets associated with corresponding attributes in X and Y, respectively, i.e., f_i is the set of fuzzy sets related to attribute x_i and g_j is the set of fuzzy sets related to attribute y_j. Finally, "X is A" is called the antecedent of the rule while "Y is B" is called the consequent of the rule. For a rule to be interesting, it should have enough support and high confidence value, larger than user specified thresholds.

To generate fuzzy association rules, all sets of items that have a support above a user specified threshold should be determined first. Itemsets with at least a minimum support are called frequent or large itemsets. The process alternates between the generation of candidate and frequent itemsets until large itemsets are identified. The following formula is used to calculate the fuzzy support value of itemset Z and its corresponding set of fuzzy sets F, denoted $S_{\langle Z,F\rangle}$.

Definition 3.2. The fuzzy support of itemset Z in the pair $\langle Z, F\rangle$ is

$$S_{\langle Z,F\rangle} = \frac{\sum_{t_i \in T} \prod_{z_j \in Z} \mu_{z_j}(f_j \in F, t_i[z_j])}{|T|},$$

where $|T|$ is the number of transactions in database T.

So, the problem of mining all fuzzy association rules converts into generating each rule, confidence is larger than the user specified minimum confidence. Explicitly, each large itemset, denoted L, is used in deriving all association rules $(L - S) \Rightarrow S$, for each $S \subset L$. Strong association rules are discovered by choosing from among all the generated possible association rules only those with confidence over a pre-specified minimum confidence. However, not all strong rules are interesting enough to be reported to the user. Whether a rule interesting or not can be judged either subjectively or objectively. Ultimately, only users can judge if a given rule is interesting or not, and this judgment, being subjective, may differ from one user to another. However, objective interestingness criteria, based on the statistics behind the analyzed data, can be used as one step toward the goal of weeding out uninteresting rules from presentation to the user.

To illustrate this, consider a rule $X \Rightarrow Y$ with 50% support and 66.7% confidence. Further, assume that the support of Y is 70%. For such case, it can be said that the rule $X \Rightarrow Y$ is a strong association rule based on the support-confidence framework. However, this rule is incomplete and misleading since the overall support of X is 75%, even greater than 66.7%. In other words, this analysis leads to the following interpretation: a customer who buys X is less likely to buy Y than a customer about whom we have no information. The truth here is that there is a negative dependence between buying X and buying Y. This negative dependence leads to not considering $X \Rightarrow Y$ as strong rule. As a result, there should be some filtering criteria to eliminate such rules from consideration as interesting rules. Explicitly, to help filtering out such misleading strong association rules, the interestingness of a rule $X \Rightarrow Y$, denoted $I(X \Rightarrow Y)$, is defined as: $I(X \Rightarrow Y) = S(X, Y)/S(X)S(Y)$, in order to give a more precise characterization of the rule.

A rule is filtered out if its interestingness is less than 1, since the nominator is the actual likelihood of both X and Y being present together and the denominator is the likelihood of having the two attributes being independent. As the above example is concerned, we can calculate the interestingness of $X \Rightarrow Y$ as: $I(X \Rightarrow Y) = 0.5/0.75 \times 0.7 = 0.95 < 1.$, which means that this rule is not interesting enough to be

reported to the user. This process helps in returning only rules having positive interestingness, and hence the size of the reported result is reduced to include more precise rules.

4. The genetic algorithm-based method

One of the most important steps in mining fuzzy association rules is to decide on the fuzzy sets according to which the values of each quantitative attribute are to be classified. In other words, the quality of the results produced relies quite crucially on the appropriateness of the fuzzy sets to the given data. So, fuzzy sets must be consistent with the values of the corresponding attribute. Fuzzy sets can be either provided by an expert or automatically derived from the contents of the existing transactions. However, fuzzy sets provided by experts may not be suitable for mining fuzzy association rules from databases. Also, it is extremely difficult for experts to estimate the most appropriate fuzzy sets.

In order to cope with these problems, we first concentrate on how fuzzy sets are determined automatically from the values of the given attributes. For this purpose, we have used a GA-based clustering algorithm. GAs form a class of adaptive search techniques based on the principles of population genetics. The standard GA proceeds as follows. It starts with an initial population of randomly or heuristically generated individuals, and advances toward better individuals by applying genetic operators modeled on the genetic processes occurring in nature. The population undergoes evolution in a form of natural selection. During successive iterations, called *generations*, individuals in the population are rated for their adaptation as solutions, and on the basis of these evaluations. As a result, a new population of individuals is formed using a selection mechanism and specific genetic operators such as *crossover* and *mutation*. To form a new population, individuals are selected according to their fitness. Consequently, an *evaluation* or *fitness* function must be devised for each problem to be solved. Given a particular individual, a possible solution, the fitness function returns a single numerical fitness, which is supposed to be proportional to the utility or adaptation of the solution represented by that individual.

The rest of this section is organized as follows. The encoding of chromosomes is presented in Section 4.1. The fitness evaluation and selection process are described in Section 4.2.

4.1. Chromosome encoding

In this study, we cluster the values of quantitative attributes into fuzzy sets with respect to a given fitness evaluation criteria. For this purpose, the GA will be employed to adjust the appropriate centroid values of the clusters.

In our experiments, we used membership functions in triangular shape because it is in general the most appropriate shape and the most widely used in fuzzy systems. To illustrate this, consider a quantitative attribute i_k and assume it has five corresponding fuzzy sets. Membership functions for attribute i_k and its centroid variables are shown in Fig. 1. These centroid points also represent the midpoint values of membership functions.

So, based on the assumption of having five fuzzy sets per attribute, as it is the case with attribute i_k, a chromosome consisting of three midpoints, also called centroid points, per attribute is represented in the following form:

$$R^1_{i_1} R^2_{i_1} R^3_{i_1} \ldots R^1_{i_m} R^2_{i_m} R^3_{i_m}.$$

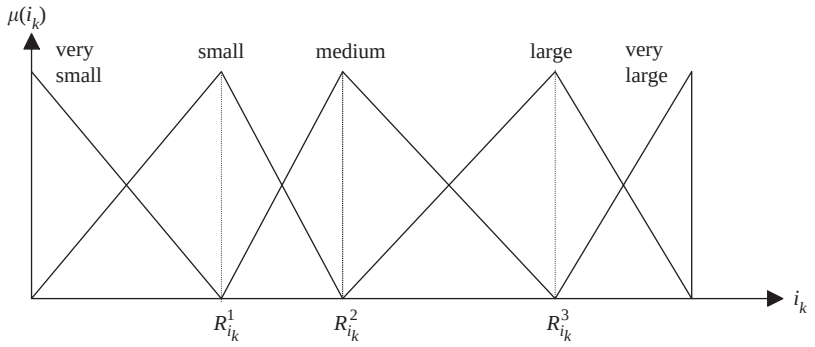

Fig. 1. Membership functions and centroid variables of attribute i_k.

We use real-valued coding, where chromosomes are represented as floating point numbers and their genes are the real parameters. These chromosomes form the input to the fitness function described in the next section.

4.2. Fitness evaluation and selection process

The fitness function measures the goodness of an individual in a given population. It is one of the key issues to a successful GA, simply because the main task in a GA is to optimize a fitness function.

The fitness function accepts a decoded chromosome and produces an objective value as a measure of the performance of the input chromosome. The aim of the GA employed in this study is to maximize the number of all the large itemsets extracted by the adjusted membership functions.

During each generation, individuals with higher fitness values survive while those with lower fitness values are destroyed. In other words, individuals who are strong according to parent selection policy are candidates to form a new population. Parent selection mimics the survival of the best individuals in the given population.

Many selection procedures are currently in use. However, we have adapted the *elitism* policy in our experiments. Elitism is to copy best solutions in present population to next generation. After selecting chromosomes with respect to the evaluation function, genetic operators such as, crossover and mutation, are applied to these individuals.

Crossover refers to information exchange between individuals in a population in order to produce new individuals. The idea behind a crossover operation is as follows. It takes as input two individuals, selects a random point, and exchanges the subindividuals behind the selected point. Since the length of the chromosomes in our study is very large, the multi-point crossover strategy was used with the crossover points determined randomly. As for the crossover, we used the arithmetical crossover method [18].

On the other hand, mutation is an operation that defines a local or global variation in an individual. Mutation is traditionally performed in order to increase the diversity of the genetic information. In our experiments, a probability test determines whether a mutation will be carried out or not.

Finally, the whole GAs process employed in this study can be summarized as follows. After generating each individual in the initial population, the executed GA includes the following steps:

(1) specify population size N and generate initial chromosomes,
(2) find the number of large itemsets according to current chromosomes,
(3) evaluate each chromosome with respect to the fitness function,
(4) perform selection, crossover and mutation,
(5) if not (end-test) go to step (2), otherwise stop and return the best chromosome.

The test in step 5 depends on the values returned by the fitness function. The fittest chromosomes are selected to form the next generation and the manipulation process is repeated until no significant improvement of the population can be observed.

5. Utilizing CURE clustering algorithm in generating membership functions

In order to evaluate our GA-based approach, we have used an existing clustering approach, namely CURE, throughout this study. CURE employs a novel hierarchical clustering algorithm that adopts a middle ground between the centroid-based and the all-point extremes [7]. In CURE, a constant number of well-scattered points in a cluster are first chosen. The scattered points capture the shape and extent of the cluster. The chosen scattered points are next shrunk toward the centroid of the cluster by a fraction α. After shrinking, these scattered points are used as representative of the cluster. The clusters with the closest pair of representative points are the clusters that are merged at each step of CURE's hierarchical clustering algorithm. The input parameters to CURE clustering algorithm are: (1) the input data set D containing $|D|$-values in n-dimensional space, where $|D|$ is the number of values in the database and n is the number of attributes; (2) the desired number of clusters k.

Starting with individual values as individual clusters, the closest pair of clusters are merged at each step to form a new cluster. The process is repeated until only k clusters are left. This way, the values of each attribute in the database are distributed into k clusters. The centroids of the k clusters are the set of midpoints of the fuzzy sets for the corresponding attribute.

To illustrate the process, suppose we want to find fuzzy sets for the ith attribute, which is quantitative with a range from $\min(i)$ to $\max(i)$. Let $\{f_{i1}, f_{i2}, \ldots, f_{ik}\}$ be the set of mid-points of the fuzzy sets for the ith attribute. Actually, the ith attribute has also two additional fuzzy sets, like in the GA-based approach, which cover the intervals $[f_{i0}, f_{i1}]$ and $[f_{ik}, f_{i(k+1)}]$, where $f_{i0} = \min(i)$. Then, the total $k + 2$ fuzzy sets will have the following ranges: $[f_{i0}, f_{i1}]$, $[f_{i0}, f_{i2}]$,…,$[f_{i(k-1)}, f_{i(k+1)}]$, and $[f_{ik}, f_{i(k+1)}]$, where $f_{i(k+1)} = \max(i)$.

After the fuzzy sets of each quantitative attribute are obtained, a corresponding membership function can be generated for each fuzzy set. In general, a membership function over a numerical universe is convex and normal. The process is illustrated next.

Suppose $\{f_{i1}, f_{i2}, \ldots f_{ik}\}$ is the set of mid-points of the fuzzy sets for attribute i; we use the following method to find the required membership functions.

For the fuzzy set with a range from f_{i0} to f_{i1}, the membership function is given by

$$F_0(x) = \begin{cases} \dfrac{x - f_{i1}}{f_{i0} - f_{i1}} & \text{if } x \leqslant f_{i1}, \\ 0 & \text{if } x > f_{i1}. \end{cases}$$

Table 1
The ranges of fuzzy sets found according to CURE clustering algorithm

Income	Fuzzy sets	Centroid
Quite poor	(10,10,30)	—
Poor	(10,30,70)	30
Moderate	(30,70,120)	70
Rich	(70,70,120)	—

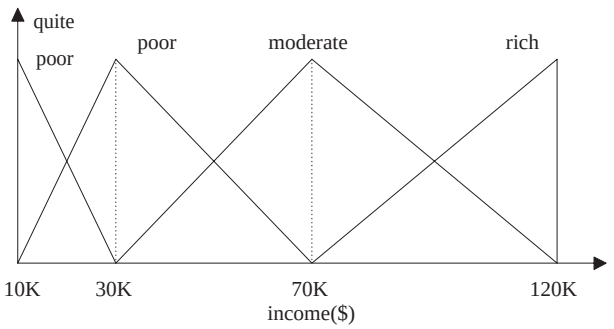

Fig. 2. The membership functions found according to the centroids.

For each fuzzy set with midpoint f_{ij}, where $1 \leqslant j \leqslant k$, the membership function is given by

$$F_{ij}(x) = \begin{cases} \dfrac{x - f_{i(j-1)}}{f_{ij} - f_{i(j-1)}} & \text{if } f_{i(j-1)} \leqslant x \leqslant f_{ij}, \\ \dfrac{x - f_{i(j+1)}}{f_{ij} - f_{i(j+1)}} & \text{if } f_{ij} \leqslant x \leqslant f_{i(j+1)}. \end{cases}$$

For the fuzzy set with a range from f_{ik} to $f_{i(k+1)}$, the membership function is given by

$$F_{k+1}(x) = \begin{cases} 0 & \text{if } x \leqslant f_{ik}, \\ \dfrac{x - f_{ik}}{f_{i(k+1)} - f_{ik}} & \text{if } f_{ik} < x \leqslant f_{i(k+1)}. \end{cases}$$

The following example illustrates the whole process. Given a quantitative attribute extracted from a synthetic database, say *income* with four different ranges as shown in Table 1. The values of *income* range from \$10K to \$120K, and can be classified into four fuzzy sets, i.e., two clusters, as shown in Fig. 2.

6. Experimental results

In this section, we present some experiments that have been carried out to test the efficiency and effectiveness of the proposed approach. All of the experiments were conducted on a Pentium III, 1.4 GHz CPU with 512 MB of memory and running Windows 2000. As experimental data, we used 100K transactional records from the United States census in the year 2000. In the experiments, we employed seven quantitative attributes and four fuzzy sets have been defined for each such attribute. Also, in all the experiments

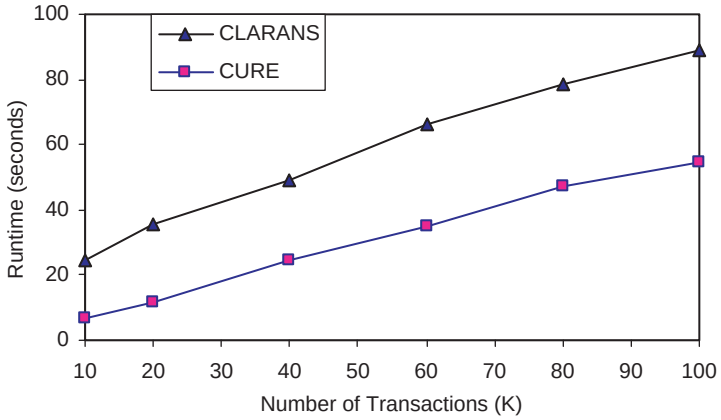

Fig. 3. The runtime required for CURE and CLARANS clustering algorithms to find 4 fuzzy sets.

conducted in this study, the GA process started with a population of 50 individuals and the maximum number of generations has been fixed at 500.

The first experiment is dedicated to support our position in using CURE clustering algorithm, we compare two of the well-known clustering algorithms described in the literature, namely CLARANS and CURE. We have run both clustering algorithms with database size ranging from 10K to 100K transactions. The number of fuzzy sets is set to four for both algorithms, i.e., the number of medoids to be found for CLARANS clustering algorithm is four while it is two for CURE. Fig. 3 shows the runtime required by each of the two algorithms to find the required four fuzzy sets per quantitative attribute.

As can be seen easily from Fig. 3, CURE outperforms CLARANS, as runtime is concerned. We run another experiment to compare discrete and fuzzy methods in the association rules construction process. We have used three attributes in the experiment and shown in Table 2 are the ranges of the attributes found with respect to four different methods: discrete intervals, Fu's work, CURE and GA-based clustering. These three attributes were chosen from the seven quantitative attributes used in the whole testing process. According to Table 2, the determined fuzzy sets and centroid points may be different from each other. It can be seen easily from Table 2 that the fuzzy sets generated by our GA-based approach are not too different than those of CURE. For example, for the linguistic term *medium below* for the *Age of Person*, the range of our approach is (16, 31, 40) while the range of CURE is (16, 25, 45). In general, the methods that use clustering algorithms to generate the proper membership functions, find a centroid point in a way similar to our approach. The only important difference is that other clustering methods take into account the distribution of the values of the attributes, while the proposed method optimizes the membership functions in favor of a given minimum support value. Finally, the fuzzy sets of the GA-based approach and discrete intervals were found by GA with respect to the given fitness criteria.

Plotted in Fig. 4 are the curves that show the number of interesting rules found according to four different methods for different values of minimum support. As expected, the number of interesting rules decreases as the minimum support value increases from 5% to 10%. In this experiment, the minimum confidence was set to 40% and the number of fuzzy sets is four for both GA and CURE-based approaches as well as for Fu's work, and the number of discrete intervals is also four. In order, Notice that two

Table 2
The fuzzy sets of attributes according to four different methods

Age of person					
Label	Young	Medium below	Medium above	Old	Centroids
Fuzzy sets for GA-based approach	(16,16,31)	(16,31,40)	(31, 40, 90)	(40, 90, 90)	31, 40
Fuzzy sets for CURE	(16,16,25)	(16,25,45)	(25, 45, 90)	(45, 90, 90)	25, 45
Fuzzy sets for Fu's work	(16,16,21,35)	(21,35,50)	(35,50,65)	(50, 65, 90, 90)	21, 35, 50, 65
Discrete intervals	[15,21]	[21,35]	[35, 55]	[55, 90]	21, 35, 55
Annual income					
Label	Low	Middle	High	Very high	
Fuzzy sets for GA-based approach	(10,10,62)K	(10,62,114)K	(62,114,200)K	(114,200,200)K	62K, 114K
Fuzzy sets for CURE	(10,10,50)K	(10,50,100)K	(50,100,200)K	(100K,200,200)K	50K, 100K
Fuzzy sets for Fu's work	(10,10,30,60)K	(30,60,90)K	(60,90,130)K	(90,130,200,200)K	30K, 60K, 90K, 130K
Discrete intervals	[10,30]K	[30,70]K	[70,120]K	[120,200]K	30K, 70K, 120K
Number of person					
Label	Small	Medium	Large	Very large	
Fuzzy sets for GA-based approach	(1,1,4)	(1,4,6)	(4,6,10)	(6,10,10)	4, 6
Fuzzy sets for CURE	(1,1,3)	(1,3,6)	(3,6,10)	(6,10,10)	3, 6
Fuzzy sets for Fu's work	(1,1,2,3)	(2,3,5)	(3,5,7)	(5,7,10,10)	2, 3, 5, 7
Discrete intervals	[1,3]	[3,5]	[5,7]	[7,10]	3, 5, 7

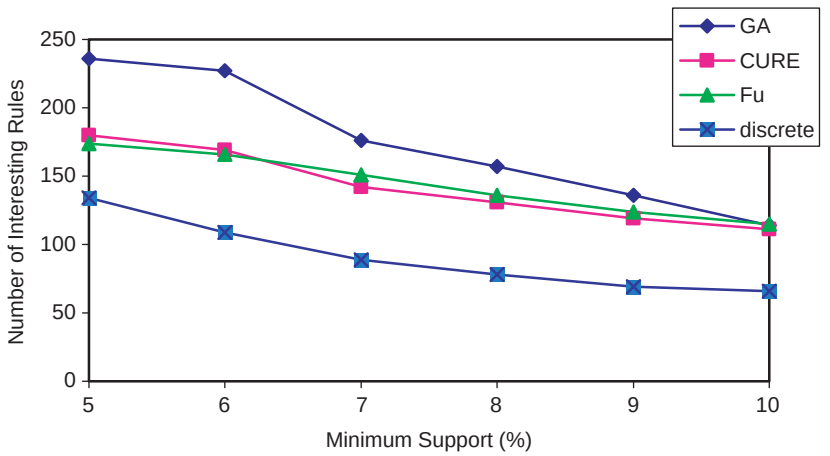

Fig. 4. Number of interesting rules for different values of minimum support.

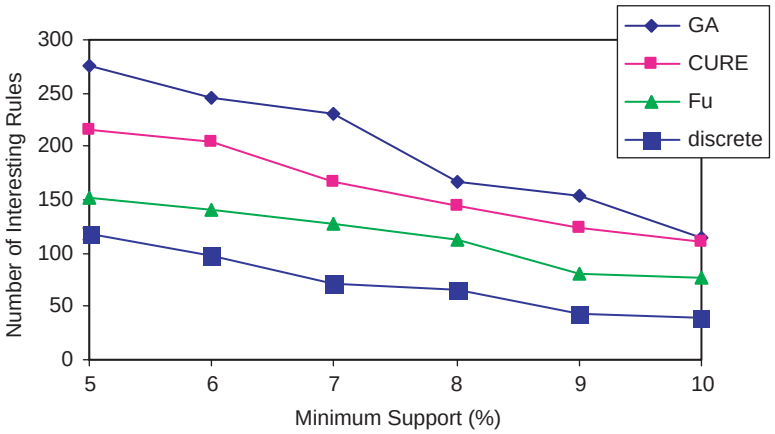

Fig. 5. Number of interesting rules obtained when the number of fuzzy sets and discrete intervals is increased to five.

centroid points are enough for GA and CURE-based approaches to find the value 4; however, the number of centroid points is four and three for Fu's approach and discrete method, respectively, to find the value four. The number of interesting rules found by CURE-based method is almost the same as that of Fu et al. approach.

The second experiment is dedicated to show the results obtained in case the number of fuzzy sets and the number of discrete intervals both increase to five. Fig. 5 shows the decrease in the number of interesting association rules as a function of the minimum support value. Here, another important point is that the difference between the curves of CURE and Fu increases in favor of CURE. This is quite consistent with our intuition since a large number of fuzzy sets or discrete intervals will make quantities of an item in different transactions easily scattered in difference sets.

The third experiment investigates the number of interesting association rules for different values of minimum confidence. The result is given in the curves plotted in Fig. 6. For this experiment, the minimum support value is set to 7%.

The last experiment compares CURE-based approach and the GA-based approach by considering the runtime required to find large itemsets for different numbers of transactions, varying from 10K to 100K. The runtime values reported in Fig. 7 represent optimal fitness at the end of 500 generations. It can be easily seen from Fig. 7 that the runtime of the GA-based approach is smaller than the CURE-based approach up to 20K transactions. Actually, this is completely compatible with our intuition because the extra time consumed by the GA-based approach is spent on optimizing membership functions to produce better result. Also, the number of membership functions is set to four for both approaches. It could be easily observed that the GA-based approach outperforms the CURE-based approach in case it is required to decide on the number of fuzzy sets as well as to optimize the ranges of the membership functions. The correctness of this idea is demonstrated and supported by the experiments conducted for our study described in [14].

Apart from the other studies described in the literature, membership functions are adjusted in this paper depending on a minimum support value given beforehand. In other words, membership functions are optimized such that the number of large itemsets is maximized with respect to a prespecified minimum support value. However, traditional clustering methods do this task by taking into account the distribution

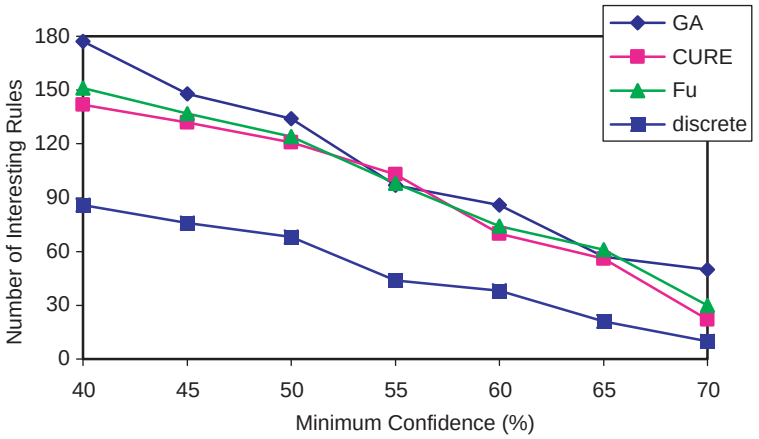

Fig. 6. Number of interesting rules for different minimum confidence values.

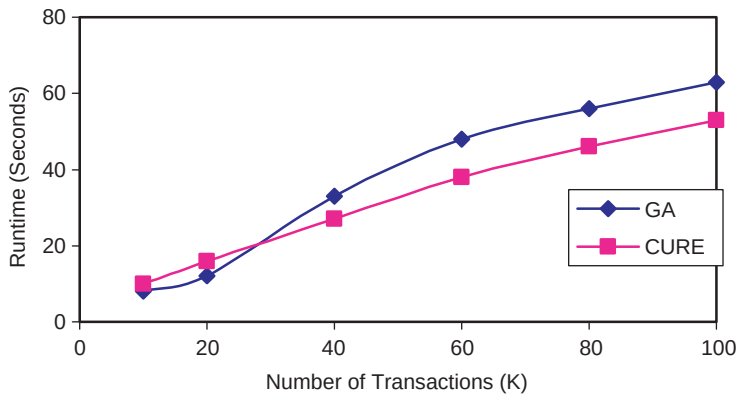

Fig. 7. The runtime required to find all large itemsets for four fuzzy sets.

of the values of attributes. In this regard, the method proposed in this paper can be autonomously applied in practice.

As our approach deals with an optimization problem, GA is used as a search procedure. In some other experiments carried out for another part of our work, we demonstrated that the runtime of the method will decrease exponentially if the number of only large 1-itemsets instead of the number of all large itemsets is taken into consideration. However, this may decrease the rule interestingness as well. In general, a linguistic term of an item with a larger support in 1-itemsets usually appears in itemsets with more items with a higher probability; this may imply more interesting association rules. But, items with support not exceeding minimum threshold value might be pruned if 2-itemsets are generated. This takes us away from the goal of developing efficient data mining framework. The paper presents an efficiency solution by achieving a trade-off between runtime and rule interestingness. Also, we concentrated on only the number of large itemsets since we wanted to optimize membership functions with respect to minimum support value. If we wanted to generate membership functions with respect to minimum confidence value

instead, then the minimum confidence value would be necessary. We demonstrated in our previous work [14] that it is enough to concentrate on only the number of large itemsets instead of the number of rules. This is because extracting association rules is a straight forward process.

In the present experiments, we used 100K transactional records. In case it is intended to apply this method on very larger data sets, say over million transactions, then it is recommended to employ the idea of random sampling in order to reduce the computation time. So, the underlying method selects and runs on a small manageable sample from the large collection of transactions.

As both Figs. 4 and 6 demonstrate, using fuzzy sets results in more rules than the discrete method. On the other hand, as using GA-based approach outperforms all the other three methods used in the comparison, using CURE clustering algorithm does improve the runtime. Some of the determined interesting fuzzy association rules are enumerated next:

IF age of person is medium AND educational level is high THEN annual income is medium.

IF age of person is medium AND income is very high AND education level is medium THEN marital status is divorced.

IF annual income is high AND number of persons in family is medium THEN number of kids in family is high.

At the end, it should be noted that in CURE-based approach, the function F_0 decreases between f_{i0} and f_{i1}, while this same interval is equal to one in CLARANS based Fu et al. work. In other words, Fu et al. employ trapezoidal membership function for the interval between f_{i0} and f_{i1}. Similarly, the function F_{k+1} increases between f_{ik} and $f_{i(k+1)}$ in CURE-based approach, while it is again fixed at 1 in Fu et al. work. Thus, the degree of membership of each attribute value differs in the two approaches.

7. Conclusions

In this paper, we have presented our approach of using GA-based clustering algorithm to implement the process of finding fuzzy sets to be used in mining interesting association rules. By using fuzzy sets, we describe association rules in concise manner. It is true that expert users can participate in the process by changing the number and position of the fuzzy sets, which lead to the discovered rules. However, it is unrealistic that users always have the capacity to intervene in the mining process. The approach proposed in this paper is a solution to these deficiencies. The position of the fuzzy sets has been adjusted by using GA. Another problem in mining association rules is to repeatedly tune the support value according to user requests. While the user may not find interesting rules for larger values of minimum support, smaller values of minimum support may lead to too many rules delivered to the user. One of the most important advantages of GA-based method is that, as apart from other methods, our approach depends on a minimum support value given beforehand. So it is possible to obtain more appropriate solutions by changing the minimum support value. Also, the number of interesting rules obtained with the GA-based approach is larger than those obtained by applying other methods. The experiments done on a real life data set showed that the proposed approach produces meaningful results and has reasonable efficiency and effectiveness providing a trade-off between runtime and rule interestingness.

References

[1] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proc. ACM SIGMOD Internat. Conf. Management of Data, May 1993, pp. 207–216.

[2] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proc. Internat. Conf. Very Large Databases, September 1994, pp. 487–499.
[3] W.H. Au, K.C.C. Chan, An effective algorithm for discovering fuzzy rules in relational databases, in: Proc. IEEE Internat. Conf. Fuzzy Systems, 1998, pp. 1314–1319.
[4] C.H. Cai, W.C. Fu, C.H. Cheng, W.W. Kwong, Mining association rules with weighted items, in: Proc. IDEAS, 1998, pp. 68–77.
[5] B.C. Chien, Z.L. Lin, T.P. Hong, An efficient clustering algorithm for mining fuzzy quantitative association rules, Proc. IFSA World Congr. and NAFIPS Internat. Conf., vol. 3, 2001, pp. 1306–1311.
[6] A.W.C. Fu, M.H. Wong, S.C. Sze, W.C. Wong, W.L. Wong, W.K. Yu, Finding fuzzy sets for the mining of association rules for numerical attributes, in: Proc. IDEL, October 1998, pp. 263–268.
[7] S. Guha, R. Rastogi, K. Shim, CURE: an efficient clustering algorithm for large databases, Inform. Syst. 26 (1) (2001) 35–58.
[8] A. Gyenesei, A fuzzy approach for mining quantitative association rules, TUCS Technical Report No: 336, March 2000.
[9] A. Gyenesei, Mining weighted association rules for fuzzy quantitative items, TUCS Technical Report No: 346, May 2000.
[10] K. Hirota, W. Pedrycz, Linguistic data mining and fuzzy modelling, Proc. IEEE Internat. Conf. Fuzzy Systems, vol. 2, 1996, pp. 1448–1496.
[11] T.P. Hong, C.S. Kuo, S.C. Chi, A fuzzy data mining algorithm for quantitative values, in: Proc. Internat. Conf. Knowledge-Based Intelligent Information Engineering Systems, 1999, pp. 480–483.
[12] T.P. Hong, C.S. Kuo, S.C. Chi, Mining association rules from quantitative data, Intell. Data Anal. 3 (1999) 363–376.
[13] H. Ishibuchi, T. Nakashima, T. Yamamoto, Fuzzy association rules for handling continuous attributes, in: Proc. IEEE ISIE, 2001, pp. 118–121.
[14] M. Kaya, R. Alhajj, Facilitating fuzzy association rules mining by using multi-objective genetic algorithms for automated clustering, in: Proc. IEEE Internat. Conf. Data Mining, November 2003.
[15] M. Kaya, R. Alhajj, F. Polat, A. Arslan, Efficient automated mining of fuzzy association rules, in: Proc. Internat. Conf. Database and Expert Systems with Applications, 2002.
[16] C.M. Kuok, A.W. Fu, M.H. Wong, Mining fuzzy association rules in databases, SIGMOD Rec. 17 (1) (1998) 41–46.
[17] B. Lent, A. Swami, J. Widom, Clustering association rules, in: Proc. IEEE Internat. Conf. Data Eng., 1997, pp. 220–231.
[18] Z. Michalewicz, Genetic Algorithms+Data Structures = Evolution Programs, Springer, Berlin, 1992.
[19] R.J. Miller, Y. Yang, Association rules over interval data, in: Proc. ACM SIGMOD Internat. Conf. Management of Data, 1997, pp. 452–461.
[20] R. Ng, J. Han, Efficient and effective clustering methods for spatial data mining, in: Proc. Internat. Conf. Very Large Databases, 1994.
[21] W. Pedrycz, Fuzzy sets technology in knowledge discovery, Fuzzy Sets and Systems (1998) 279–290.
[22] R. Srikant, R. Agrawal, Mining quantitative association rules in large relational tables, in: Proc. ACM SIGMOD Internat. Conf. Management of Data, 1996, pp. 1–12.
[23] R.R. Yager, Fuzzy summaries in database mining, Proc. Conf. Artif. Intell. Appl. (1995) 265–269.
[24] S. Yue, E. Tsang, D. Yeung, D. Shi, Mining fuzzy association rules with weighted items, Proc. IEEE Internat. Conf. Systems Man Cybernet. (2000) 1906–1911.
[25] L.A. Zadeh, Fuzzy sets, Inform. and Control 8 (1965) 338–353.
[26] W. Zhang, Mining fuzzy quantitative association rules, Proc. IEEE Internat. Conf. Tools Artif. Intell. (1999) 99–102.