Applied Intelligence 24, 7–15, 2006

Utilizing Genetic Algorithms to Optimize Membership Functions for Fuzzy Weighted Association Rules Mining

MEHMET KAYA
Department of Computer Engineering, Firat University, 23119 Elazig, Turkey
kaya@firat.edu.tr

REDA ALHAJJ
Department of Computer Science, University of Calgary, Calgary, Alberta Canada;
Department of Computer Science, Global University, Beirut, Lebanon
alhajj@cpsc.ucalgary.ca

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. In general, it is unrealistic that experts can always provide such sets. And finding the most appropriate fuzzy sets becomes a more complex problem when items are not considered to have equal importance and the support and confidence parameters required for the association rules mining process are specified as linguistic terms. Existing clustering based automated methods are not satisfactory because they do not consider the optimization of the discovered membership functions. In order to tackle this problem, we propose Genetic Algorithms (GAs) based clustering method, which dynamically adjusts the fuzzy sets to provide maximum profit based on user specified linguistic minimum support and confidence terms. This is achieved by tuning the base values of the membership functions for each quantitative attribute with respect to two different evaluation functions maximizing the number of large itemsets and the average of the confidence intervals of the generated rules. To the best of our knowledge, this is the first effort in this direction. Experiments conducted on 100 K transactions from the adult database of United States census in year 2000 demonstrate that the proposed clustering method exhibits good performance in terms of the number of produced large itemsets and interesting association rules.

Keywords: data mining, clustering, fuzzy association rules, genetic algorithms, linguistic terms, weighted 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 techniques described in the literature. Associations allow capturing almost all possible rules that explain the presence of some attributes according to the presence of other attributes in the same transaction. For instance, an association rule in a supermarket basket data may be stated as, "in 20% of the transactions, 75% of the people buying butter also buy eggs in the same transaction"; 20% and 75%, respectively, represent rule's support and confidence, which are the major factors in measuring the significance of an association rule. Simply, support is the percentage of transactions that contain both butter and eggs, while confidence is the ratio of the support of butter and eggs together to the support of butter. So, the problem can be stated as: *find all association rules that satisfy user-specified minimum support and confidence values*.

Research in the field concentrated mainly on boolean and quantitative association rules. Also, some researchers investigated weighted mining based on the fact that the degree of interest in items of a database may differ from one user to another; even the same user may not show the same degree of interest in all items. For instance, facing the adult information selected from the census of United States in 2000, a pretty girl might be more interested in the attribute "occupation" than other attributes in selecting a boyfriend. On the other hand, "age" is expected to be one of the most important attributes for an insurance consultant. To satisfy such cases, Cai et al. [4] proposed weighted mining to reflect different importance in different attributes; a user specified numerical weight is assigned to each attribute. Weighted support and weighted confidence were then defined to determine interesting association rules. Later on, Yue et al. [28] extended these concepts to fuzzy

item vectors. In these studies minimum support and minimum confidence are specified as numerical values. However, linguistic minimum support and minimum confidence values are more natural and understandable for humans [15]; it is essential to incorporate people intuition in the process because data mining is mainly intended for decision making.

Although current quantitative association rules mining algorithms solved some of the problems introduced by quantitative attributes, they introduced some other problems. The major problem is caused by the sharp boundary between intervals. Using sharp boundary intervals is also not intuitive with respect to human perception. The problem can be handled smoothly by introducing fuzziness into the model as in the approach described in this paper.

Unlike classical set theory where membership is binary, the fuzzy set theory introduced by Zadeh [29] provides excellent means to model the "fuzzy" boundaries of linguistic terms by introducing gradual membership. Some example linguistic terms include "important", "young", "rich", "excellent", etc. Based on this and instead of using sharp boundary intervals, some work have recently been done on the use of fuzzy sets in discovering association rules for quantitative attributes. 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. In particular, they have not considered the optimization of membership functions. The number of fuzzy sets is given as a constant beforehand and membership functions are tuned in terms of this fixed value.

Having all of this in mind, this paper contributes to the ongoing research on data mining by combining advantages of several concepts, including fuzziness, association rules, weighted mining and specifying both minimum support and minimum confidence as linguistic terms. We also propose a clustering method that employs GA to optimize membership functions used in determining fuzzy quantitative association rules. The base values of membership functions for each quantitative attribute are tuned by GA in order to maximize the number of large itemsets in a certain continuous interval of minimum support values, or the average of confidence intervals of the rules exceeding the threshold interval of minimum confidence. To achieve this, we defined two fitness functions. Advantages and the effectiveness of the proposed method are demonstrated by testing it on 100 K transactions taken from the adult database of United States census in year 2000.

GAs demonstrated high success in solving some of the major research problems in computer science. They are general purpose search algorithms that use principles inspired by natural genetic populations to generate solutions to complicated search problems [13]. In other words, GAs are theoretically and empirically proven to provide robust search capabilities in complex spaces, offering a valid approach to problems requiring efficient and effective searching. The basic idea is to maintain a population of *individuals* called *chromosomes*,[1] which represent candidate solutions to the concrete problem that evolves over time through a process of competition and controlled variation. Each chromosome in the population has an associated *fitness*, which is utilized in determining chromosomes to be used in constructing new chromosomes in the competition process, called *selection*. The new chromosomes are created using genetic operators such as *crossover* and *mutation*.

The rest of this paper is organized as follows. Section 2 includes an overview of the related work. Fuzzy weighted association rules, fuzzy item importance, fuzzy minimum support and fuzzy minimum confidence are all defined in Section 3. Our approach of utilizing GA to optimize membership functions is described in Section 4. Experimental results are presented in Section 5. Section 6 is the conclusions.

2. Related Work

Fuzzy association rules are in general 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 faced when quantitative attributes are considered. However, we have not encountered in the literature any approach that employs GAs as presented in this paper. So, the rest of this section is dedicated to cover the existing approaches to deal with clustering and sharp boundary problems. We also present some of the basic ideas, which we benefited from in this study, including linguistic terms and fuzziness.

Srikant and Agrawal [25] used equi-depth partitioning to mine quantitative rules. They separate intervals by their relative ordering and quantities equally. Miller and Yang applied Birch clustering [22] to identify intervals and proposed a distance-based association rule to improve the semantics of intervals. Lent et al. [21] presented a geometric-based algorithm to perform clustering for numerical attributes. Finally, Guha et al. [9] proposed an efficient clustering algorithm called CURE. Their experiments confirm that the quality of clusters produced by CURE is much better than those reported by earlier algorithms. Further, they demonstrated that CURE not only outperforms existing algorithms, but also scales well for large databases without sacrificing clustering quality.

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. For instance, Yager [27] introduced fuzzy linguistic summaries on different attributes. Hirota and Pedrycz [12, 24] 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 [5] applied fuzzy linguistic terms to relational databases with numerical and categorical attributes. Later, they proposed the F-APACS method [3] 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.

Fu et al. [7] proposed an automated method to find fuzzy sets for the mining of fuzzy association rules; their method is based on CLARANS clustering algorithm [23]. We developed a more efficient approach based on CURE clustering algorithm [19]. Hong et al. [14] proposed an algorithm that integrates fuzzy set concepts and Apriori mining algorithm to find interesting fuzzy association rules from given transaction data. They also proposed definitions for the support and confidence of fuzzy membership grade and designed a data mining approach based on fuzzy sets to find association rules with linguistic terms of human knowledge [15]. Ishibuchi et al. [16] illustrated fuzzy versions of confidence and support that can be used to evaluate each association rule. The approach developed by Zhang [30] extends the equi-depth partitioning with fuzzy terms. However, it assumes fuzzy terms as predefined. Finally, Wang and Bridges [26] used GAs to tune membership functions of the fuzzy sets used in mining fuzzy association rules for intrusion detection system. Their goal is just to increase the similarity of rules mined from data without intrusions and the reference rule set while decreasing the similarity of rules mined from intrusion data and the reference rule set. In this paper, we employ GA to adjust the membership functions for mining fuzzy weighted association rules too. But our evaluation criteria are completely different than the previous ones as detailed in the sequel.

Many other researchers have explored the use of GAs to tune fuzzy logic controllers. Early work in this area was due to Karr [17, 18] who uses GAs to modify the membership functions of the variables used by the fuzzy logic controller. The tuning method employs GAs to adjust membership functions of the fuzzy rules dealing with their parameters according to a fitness function. Other methods are presented in [2, 6, 10].

3. Weighted Fuzzy Association Rules

In this section, we present an overview of weighted fuzzy association rules. First, we introduce the degree of membership in fuzzy sets. Second, we define fuzzy association rules. Finally, we define weighted fuzzy support and confidence values.

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, denoted D_{i_k}. Besides, each quantitative attribute i_k is associated with at least two fuzzy sets 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 the fuzzy sets specified for i_k is directly based on the evaluation of the membership function of the particular fuzzy set with the value of i_k as input.

Definition 3.1 (Membership Function). Let $F_{i_k} = \{f_{i_k}^1, f_{i_k}^2, \ldots, f_{i_k}^l\}$ be a set of l fuzzy sets associated with item i_k. Each fuzzy set $f_{i_k}^j$ has a corresponding membership function, denoted $\mu_{f_{i_k}^j}(v_{i_k})$, which is a mapping from the domain of i_k into the interval [0,1], where $v_{i_k} \in D_{i_k}$. Formally, $\mu_{f_{i_k}^j} : D_{i_k} \rightarrow [0, 1]$, where

$$\mu_{f_{i_k}^j}(v_{i_k}) = \begin{cases} 1 & v_{i_k} \text{ totally belongs to fuzzy set } f_{i_k}^j \\ 0 & v_{i_k} \text{ is not a member of fuzzy set } f_{i_k}^j \\ \text{otherwise} & v_{i_k} \text{ partially belongs to fuzzy set } f_{i_k}^j \end{cases}$$

According to Definition 3.1, the obtained value $\mu_{f_{i_k}^j}(v_{i_k})$ falls in the interval [0, 1], with the lower bound 0 strictly indicates "not a member", and the upper bound 1 indicates "total membership." All other values between 0 and 1, exclusive, specify "partial membership" degree.

Given a database of transactions, its set of attributes, and the fuzzy sets associated with the quantitative attributes, interesting fuzzy association rules are potentially useful regularities. We use the following form for fuzzy association rules [17].

Definition 3.2 (Fuzzy Association Rule). A fuzzy association rule is expressed as:

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

where X and Y are disjoint sets of attributes called itemsets, i.e., $X \cup Y \subseteq I$ and $X \cap Y = \phi$; 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; $1 \leq i \leq p$ and $1 \leq j \leq q$. Finally, for a rule to be interesting, it should have enough support and high confidence value, i.e., larger than user specified thresholds.

As weighted mining is concerned, items are assigned weights to reflect their importance and weighted support and

confidence values are employed in deciding on interesting association rules. In general, most data mining algorithms set weights of items as numerical values, with the weight of an item varies between 0 and 1 based on users' experience or intuition. Here, "0" means that the corresponding attribute should be neglected, and "1" indicates that the corresponding attribute is one of the most important attributes for the user.

So, let $\langle X, A\rangle$ denote an itemset-fuzzy set pair, where X is a set of attributes and A is the set of corresponding fuzzy sets. A weight $0 \leq w_{(x,a)} \leq 1$ is assigned to each instance (x, a) of $\langle X, A\rangle$, to show its importance.

By assigning weights to attributes, we employ weighted fuzzy support and weighted fuzzy confidence in the process of deciding on large weighted (itemset, fuzzy-set) pairs and interesting association rules, respectively.

Definition 3.3 (Weighted Fuzzy Support). Given an itemset-fuzzy set pair $\langle X, A\rangle$, its weighted fuzzy support is defined as:

$$WS_{\langle X,A,w\rangle} = \left(\prod_{\substack{x_j \in X \\ a_j \in A}} w(x_j, a_j)\right) \cdot S_{\langle X,A\rangle} = \frac{\sum_{t_i \in T} \prod_{x_j \in X a_j \in A} w(x_j, a_j) \cdot \mu_{a_j}(t_i \cdot x_j)}{|T|}$$

Based on Definition 3.3, an itemset-fuzzy set pair $\langle X, A\rangle$ is called large if its weighted fuzzy support is greater than or equal to the specified minimum support threshold, i.e., $WS_{\langle X,A,w\rangle} \geq \min\sup$.

Definition 3.4 (Weighted Fuzzy Confidence). Given the rule "*If X is A then Y is B*", its weighted fuzzy confidence is defined as:

$$WC_{\langle\langle X,A,w\rangle,\langle Y,B,w\rangle\rangle} = \frac{WS_{\langle Z,C,w\rangle}}{WS_{\langle X,A,w\rangle}},$$

where $Z = X \cup Y$, $C = A \cup B$.

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 is 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 criterion, based on the statistics behind the analyzed data, can be used as one step towards 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 dependency between buying X and buying Y. This negative dependency 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: $I(X \Rightarrow Y) = \frac{S(X,Y)}{S(X)S(Y)}$, to give a more precise rule characterization.

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: $I(X \Rightarrow Y) = \frac{0.5}{0.75 \times 0.7} = 0.95 < 1$, which means that this rule is not interesting enough to be reported to the user. This process will help in returning only rules having positive interestingness, and hence the size of the reported result is reduced to include more precise rules.

3.1. Fuzzy Representation of Item Importance, Minimum Support and Minimum Confidence

The importance of an item is not only a vital measure of interestingness, but also a way to permit users to control the mining results by taking specific actions. So, it is more natural and intuitive for humans to deal with linguistic terms than discrete values. In other words, it is more flexible and more understandable for human beings to handle the measures showing the importance of an item as linguistic terms. Motivated by this, we represent weights of items using fuzzy sets.

Shown in Fig. 1 are membership functions of the fuzzy sets used to represent the weight of a given item. According to Fig. 1, membership functions have uniform structure and the weight of an item can take 5 different linguistic terms.

Concerning minimum support and minimum confidence, we also used linguistic terms to express each of them, and based on the same justification raised above about utilizing linguistic terms for the importance of items. This way, instead of a sharp boundary, we achieve a boundary with continuous interval of minimum support as well as minimum confidence. Shown in Fig. 2 are the membership functions used for minimum support; note that membership functions of the minimum confidence have the same trend shown in Fig. 2.

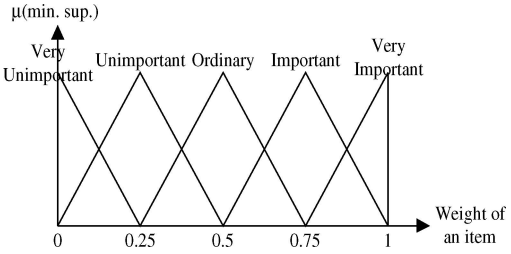


Figure 1. The membership functions representing the weights of an item.

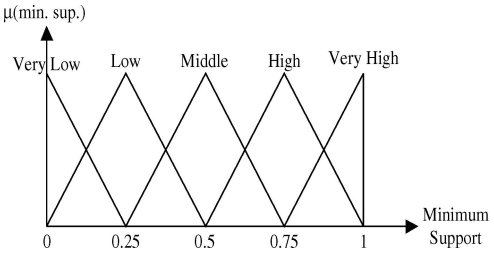


Figure 2. The membership functions representing minimum support.

4. Employing Genetic Algorithms to Optimize Membership Functions

GAs are iterative procedures that work on a population of individuals. Each individual is represented by a finite string of symbols, known as the genome. It is an encoding of a possible solution in a given problem space. This space, referred to as the search space, comprises all possible solutions to the problem in hand.

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 on the basis of these fitness 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 accepts a decoded chromosome as input and produces an objective value as a measure of the performance of such input chromosome.

The rest of this section is organized as follows. The encoding of chromosomes is presented in Section 4.1. The fitness function is described in Section 4.2. The selection process is discussed in Section 4.3.

4.1. Chromosome Encoding

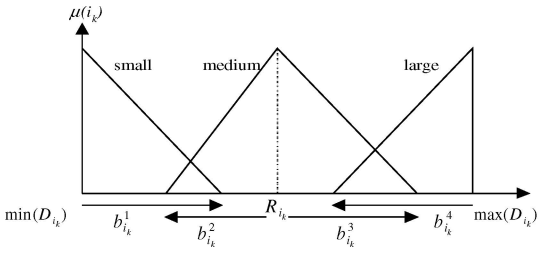
Our target in using GAs is to cluster the values of quantitative attributes into fuzzy sets with respect to a given fitness evaluation criteria. For this purpose, each individual represents the base values of membership functions of a quantitative attribute in the database. In our experiments, we used membership functions in triangular shape. To illustrate the encoding process, consider a quantitative attribute i_k and assume it has 3 corresponding fuzzy sets and hence there are 3 membership functions, one per fuzzy set.

Membership functions for attribute i_k and their base variables are shown in Fig. 3. Each base variable takes finite values. For instance, the search space of base value $b_{i_k}^1$ lies between the minimum and maximum values of attribute i_k, denoted and $\max(D_{i_k})$, respectively. The search intervals of all the base values and intersection point R_{i_k} of attribute i_k are enumerated next to Fig. 3.

So, based on the assumption of having 3 fuzzy sets per attribute, as it is the case with attribute i_k, a chromosome consisting of the base lengths and the intersection points is represented in the following form:

$$b_{i_1}^1 b_{i_1}^2 R_{i_1} b_{i_1}^3 b_{i_1}^4 b_{i_2}^1 b_{i_2}^2 R_{i_2} b_{i_2}^3 b_{i_2}^4 \ldots b_{i_m}^1 b_{i_m}^2 R_{i_m} b_{i_m}^3 b_{i_m}^4 .$$

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.



$b_{i_k}^1 : [\min(D_{i_k}), \max(D_{i_k})]$
$R_{i_k} : [\min(D_{i_k}), \max(D_{i_k})]$
$b_{i_k}^2 : [\min(D_{i_k}), R_{i_k}]$
$b_{i_k}^3 : [R_{i_k}, \max(D_{i_k})]$
$b_{i_k}^4 : [\min(D_{i_k}), \max(D_{i_k})]$

Figure 3. Membership functions and base variables of attribute i_k.

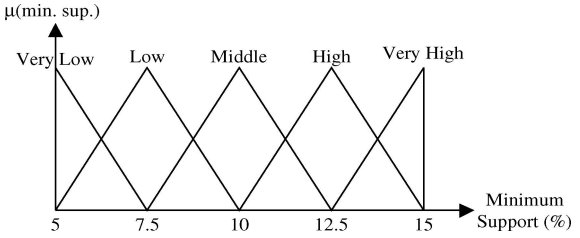


Figure 4. Membership functions of the minimum support used for the fitness function.

4.2. Fitness Evaluation

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. Consequently, the fitness function should be carefully set by considering all factors that play important role in optimizing the problem under investigation. Every new population generated in the process is evaluated with respect to the fitness function. The evaluation process is a main source to provide the mechanism for evaluating the status of each chromosome; it is an important link between the GA and the system.

The aim of the GA employed in this study is to maximize the number of large itemsets in a given continuous interval of minimum support values, or maximize the average of confidence intervals of the rules as second method. We created a continuous domain within a certain interval of minimum support and minimum confidence because we used linguistic minimum support and minimum confidence values. Figure 4 shows membership functions of the minimum support variable used in computing the fitness function. This variable has 5 uniform membership functions and its definitive interval is bounded with [0.05, 0.15].

With linguistic minimum support, the process of finding the set of large itemsets proceeds as illustrated next. Assume the linguistic minimum support value is given as "Low." First, this value is transformed into a fuzzy set of minimum support, namely (0.05, 0.075, 0.1) as shown in Fig. 4. Second, the fuzzy weighted set of the given minimum support is computed. Finally, the weighted support of each item or itemset is compared to the fuzzy weighted minimum support by fuzzy ranking. If the weighted support is equal to or greater than the weighted minimum support, then the corresponding itemset is considered large.

The process of mining fuzzy weighted association rules starts by employing GA for tuning membership functions in each generation. However, while adjusting membership functions of attribute i_k by the GA, each value of i_k intersects with one or more of the membership functions devoted to i_k. Therefore, membership functions do not generally have a uniform structure. Based on this, attribute i_k undergoes a normalization process, which is mainly a transformation that leads to a total contribution of 1.0 for attribute i_k. The normalization process is performed as follows:

$$\mu_{i_k}^1\left(f_{i_k}^j, t_v \cdot i_k\right) = \frac{\mu_{i_k}\left(f_{i_k}^j, t_v \cdot i_k\right)}{\sum_{p=1}^{l_{i_k}} \mu_{i_k}\left(f_{i_k}^p, t_v \cdot i_k\right)}$$

where l_{i_k} represents the number of fuzzy sets related to attribute i_k; and $\mu_{i_k}(f_{i_k}^j, t_v \cdot i_k)$ represents the membership degree in the j-th fuzzy set for the value of attribute i_k in transaction t_v.

4.3. Selection Process

During each generation, individuals that satisfy the selection criteria do survive while others 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, Holland's original fitness-proportionate selection is one of the simplest selection procedures [11]. So, we decided to utilize this selection policy in our experiments.

Let *fitness*(*x*,*t*) and *Avgfitness*(*t*), respectively, denote the fitness of individual x and the average fitness of the population during evolution phase t. Then, the usage value of individual x as a parent is: $tsr(x, t) = \frac{fitness(x,t)}{Avgfitness(t)}$.

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 the crossover operation utilized in our study is as follows. It takes as input 2 individuals, selects a random point, and exchanges the subindividuals behind the selected point. Since the length of the chromosomes is long, the multi-point crossover strategy has been used with the crossover points determined randomly; in particular three-point cross-over has been adapted in this study.

On the other hand, mutation means a random change in the information of an individual. It is very important for populations. It 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. Otherwise, after several generations, the diversity of the chromosomes decreases and some chunks of the chromosomes may end up being the same for all population members and the information they contain may not evolve further. A probability test determines whether a mutation will be carried out or not. The probability of mutation depends on the condition: average fitness of new generation < average fitness of old generation.

Since the initial population can be a subset of all possible solutions, an important bit of each chromosome may be

inverted, i.e., 0 appears as 1 or vice versa. Crossover may not solve this and mutation is inevitable for the solution.

Finally, the whole GA 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.

Algorithm 4.1 (Generating Association Rules)

1. Using the given membership functions about item importance, transform each linguistic term, which reflects the importance of item i_k, $1 \leq k \leq m$, into a fuzzy set of weights w_k.
2. Specify population size N and generate initial chromosomes.
3. According to the current chromosome, transform the quantitative value $t_j \cdot i_k$ of each item i_k in each transaction t_j, $1 \leq j \leq n$, into a fuzzy set f_{i_k}.
4. Calculate the fuzzy weighted support of each item-fuzzy set pair $\langle i_k, f_{i_k} \rangle$.
5. Compute the weighted fuzzy set of the given minimum support value as:

$$WMinS = S \cdot (\text{the weight of } f_{\text{ave}})$$

6. Find the large itemsets based on the weighted fuzzy set of the given minimum support value
7. Evaluate each chromosome with respect to the already specified fitness function
8. Perform selection, crossover and mutation
9. If not (end-test) go to Step (3), otherwise return the best chromosome
10. Generate all possible association rules from each identified large weighted fuzzy itemset.
11. From the rules generated in step 10, identify strong association rules based on the specified fuzzy weighted confidence.
12. From the rules identified in step 11, decide on interesting association rules by calculating the interestingness value for each strong rule.

Algorithm 4.1 employs GA to return interesting association rules. The termination condition in Step 9 of Algorithm 4.1 becomes valid when either stability is achieved in the generated population or the maximum number of 300 generations is reached. The process considers fuzzy importance of items and involves fuzzy weighted support and fuzzy weighted confidence. This algorithm has been implemented and tested on a real dataset; the results are presented next in Section 5.

5. Experimental Results

We used real-life dataset and conducted some experiments to assess the effectiveness of the GA-based fuzzy weighted mining approach presented in this paper. All of the experiments were performed using a Pentium III, 1.4 GHz CPU with 512 MB of memory and running Windows 2000. As experimental data, we used 100 K transactions dataset taken from the adult data of United States census in 2000. In the experiments, we have used 6 quantitative attributes, each with three corresponding fuzzy sets. Finally, we have used three linguistic intervals for which random linguistic weights have been generated, namely (Important, Very-Important), (Ordinary, Important) and (Unimportant, Ordinary), denoted I–VI, O–I and UI-O, respectively. The membership functions of these linguistic weights have already been shown in Fig. 1.

In all the experiments in this study, the GA process starts with a population 60 individuals for both approaches. As mentioned earlier, we have used real-valued coding. Chromosome length is considerably large, fixed at 30 because we use 6 attributes in the experiment and three fuzzy sets are assumed per attribute. Consequently, we adapted multi-point crossover strategy with the crossover points determined randomly. We namely use arithmetic crossover method in the experiments [11]. Further, crossover and mutation probabilities are taken as 0.9 and 0.01, respectively. As selection procedure is concerned, we have adapted the elitism policy in our experiments. The GA was run three times, varying the random seed used to generate the initial population. The best individual of the three runs, according to its fitness value, was selected as the best solution.

We conducted two sets of experiments for either approach. In the first set of experiments, we considered the fitness function dealing with the maximum number of large itemsets. The first experiment tested, for the three different linguistic weight intervals enumerated above, the correlation between expressing minimum support in linguistic terms and the number of large itemsets produced. The obtained results are reported in Fig. 5, which shows that the number of large itemsets decreases as a function of linguistic minimum support. Also, as the importance of items decreases, the maximum number of large itemsets obtained by adjusting the membership functions of item quantities is reduced. In this and the next experiments, we have used the membership functions of the minimum support shown in Fig. 4.

In the second experiment, after the maximum number of large itemsets is found by GA at the linguistic value "middle", we test, for the three linguistic weight intervals, the effect of using linguistic terms to express minimum confidence, as shown in Fig. 6, on the number of generated interesting association rules. The achieved results are reported in Fig. 7. The obtained results do meet our expectations, i.e., more rules are generated for higher weights. However, the number decreases, for all cases, as the linguistic confidence threshold increases.

The last experiment of the first set is dedicated to investigate the performance for the three linguistic intervals. In particular, we examined how the performance varies with

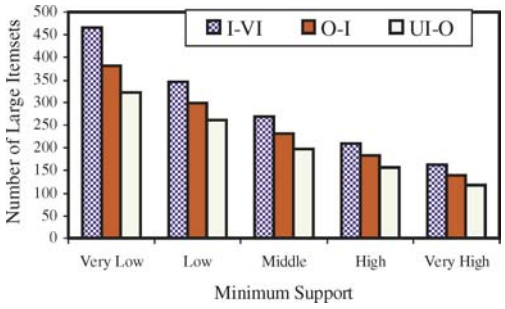


Figure 5. Number of large itemsets determined by GAs for different linguistic terms of minimum support.

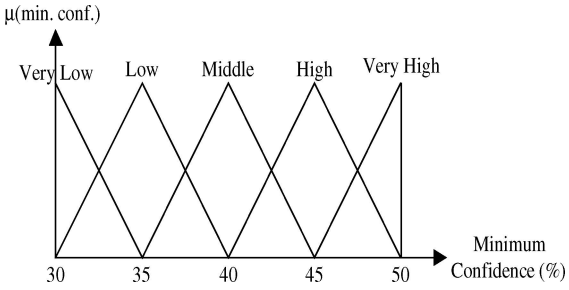


Figure 6. Membership functions of the minimum confidence.

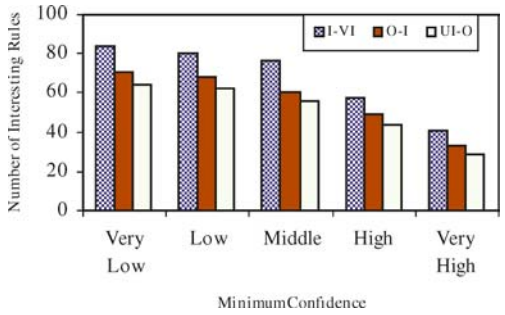


Figure 7. Number of interesting rules found based on different linguistic terms of minimum confidence; minimum support fixed as "middle".

the number of transactions. This is reflected in Fig. 8, which shows the change in the runtime as we increase the number of input records from 10 K to 100 K, for the three different cases. One of the cases takes "Low" and "I-VI" for minimum support and weight of items, respectively. In such a case, the time required to find the maximum number of large itemsets is larger than the others. If we increase the interval of minimum support from "Low" to "Middle" as the weights of items are fixed at "I-VI", the time required decreases. In addition to this, if we decrease the weight of items from "I-VI" to "O-I", we can find the maximum number of large itemsets in a shorter time. The results plotted in Fig. 8 show that the method scales quite linearly for the census dataset used in the experiments.

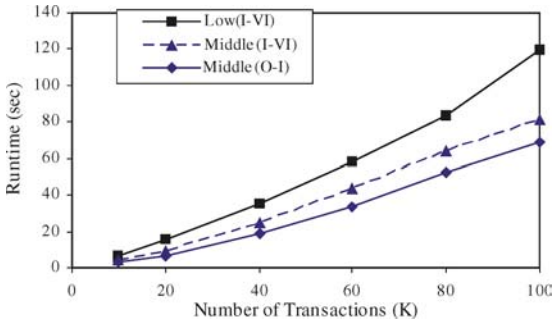


Figure 8. The runtime required for GAs to find three fuzzy sets for the three linguistic intervals.

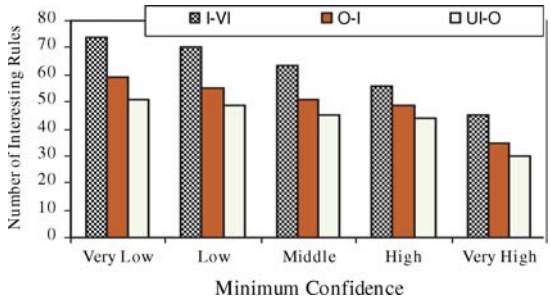


Figure 9. Number of interesting rules found with respect to the second fitness function.

In the second set of experiments, we changed the fitness function to become the average of the confidence intervals for the generated rules. In this new case, we fixed minimum support as "middle" again. Then, we extracted the fuzzy weighted association rules that have confidence intervals larger than or equal to "very low." As the final step, the average values of these strong rules were computed. Our aim is to maximize the average values. In this regards, Fig. 9 gives the number of generated interesting fuzzy weighted rules, for the same linguistic weight intervals. As can be seen from Fig. 9, compared to the previous fitness function, the number of rules decreases at all the intervals except the minimum confidence "very high." Moreover, the difference between both experiments at the interval of "very low" or even "low" is considerably large. This means that the user misses some rules that may be interesting at those intervals, although the number of stronger rules increased at very high rate, where the number of growth is at most 4.

The last experiment is dedicated to investigate the runtime needed for each case under the second fitness function. The results are reported as shown in Fig. 10. The same interpretation stated for Fig. 8 is valid for Fig. 10.

Finally, two of the rules found as a result of our experiments can be enumerated as:

- If income is high AND education level is high THEN age is middle.

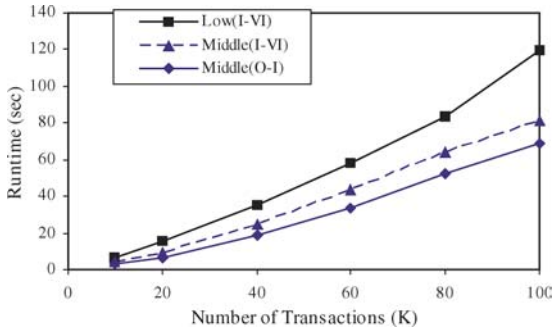


Figure 10. The runtime required for GAs to find three fuzzy sets at the second fitness function.

- If number of persons in family is middle AND educational level of head of the family is high THEN income is high.

6. Conclusions

In this paper, we proposed a clustering approach to solve the problem of interval partitioning. For this purpose, we proposed a GA-based approach. The approach uses two different fitness functions. As one of them deals with the maximum number of large itemsets based on linguistic minimum support, the other employs the average of the confidence intervals of the rules. The main achievement of the proposed approach is employing GA to dynamically adjust and optimize membership functions, which are essential in finding interesting weighted association rules from quantitative transactions, based on support and confidence specified as linguistic terms. Compared to previous mining approaches, the proposed one directly, manages linguistic parameters, which are more natural and understandable to humans. Results of the experiments conducted on a real life census dataset demonstrate that the method that employs the first fitness function outperforms the one with the average confidence interval in terms of the required runtime and even the number of interesting rules.

Note

1. In the rest of the paper, individuals and chromosomes are interchangeably used.

References

1. R. Agrawal, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases," *Proc. of ACM SIGMOD*, pp. 207–216, 1993.
2. A. Arslan and M. Kaya, "Determination of Fuzzy Logic Membership Functions using Genetic Algorithms," *Fuzzy Sets and Systems*, vol. 118, no. 2, pp. 297–306, 2001.
3. W.H. Au and K.C.C. Chan, "An effective algorithm for discovering fuzzy rules in relational databases," *Proc. of IEEE-FUZZ*, pp. 1314–1319, 1998.
4. C.H. Cai, et al., "Mining association rules with weighted items," *Proc. of IDEAS*, 1998, pp. 68–77.
5. K.C.C. Chan and W.H. Au, "Mining fuzzy association rules," *Proc. of ACM CIKM*, 1997, pp. 209–215.
6. O. Cordón, F. Gomide, F. Herrera, F. Hoffmann, and L. Magdalena, "Ten years of genetic fuzzy systems: current framework and new trends," *Fuzzy Sets and Systems*, vol. 141, no. 1, pp. 5–31, January 2004.
7. A.W.C. Fu, et al., "Finding fuzzy sets for the mining of association rules for numerical attributes," in *Proc. of the International Symposium of Intelligent Data Engineering and Learning*, 1998, pp. 263–268.
8. D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley: Reading, MA, 1989.
9. S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," *Information Systems*, vol. 26, no. 1, pp. 35–58, 2001.
10. F. Herrera, M. Lazono, and L. Verdegay, "Tuning fuzzy logic control by genetic algorithms," *Int. Journal of Approximate Reasoning*, vol. 12, no. 3/4, pp. 299–315, 1995.
11. F. Herrera, M. Lozano, and J.L. Verdegay, "Tackling real-coded genetic algorithms: Operators and tools for behavioural analysis," *Artificial Intelligence Review*, vol. 12, no. 4, pp. 265–319, August 1998.
12. K. Hirota and W. Pedrycz, "Linguistic data mining and fuzzy modelling," *Proc. of IEEE-FUZZ*, vol. 2, 1996, pp. 1448–1496.
13. J.H. Holland, *Adaptation in Natural and Artificial Systems*, The MIT Press. Cambridge, MA, MIT Press edition, 1992. First edition: University of Michigan Press, 1975.
14. T.P. Hong, C.S. Kuo, and S.C. Chi, "Mining association rules from quantitative data," *Intelligent Data Analysis*, vol. 3, pp. 363–376, 1999.
15. T. P. Hong, M. J. Chiang, and S. L. Wang, "Mining from quantitative data with linguistic minimum supports and confidences," *Proc. of IEEE-FUZZ*, 2002, pp. 494–499.
16. H. Ishibuchi, T. Nakashima, and T. Yamamoto. "Fuzzy association rules for handling continuous attributes," in *Proc. of IEEE International Symposium on Industrial Electronics*, 2001, pp. 118–121.
17. C.L. Karr, "Design of an adaptive fuzzy controller using a genetic algorithm," in *Proc. of the 4th Intl. Conf. on Genetic Algorithms*, 1991.
18. C.L. Karr and E.J. Gentry, "Fuzzy control of pH using genetic algorithms," *IEEE Trans. Fuzzy System*, vol. 1, pp. 46–53, 1993.
19. M. Kaya, R. Alhajj, F. Polat, and A. Arslan, "Efficient automated mining of fuzzy association rules," *Proc. of DEXA*, 2002.
20. C.M. Kuok, A.W. Fu, and M.H. Wong. "Mining fuzzy association rules in databases," *SIGMOD Record*, vol. 17, no. 1, pp. 41–46, 1998.
21. B. Lent, A. Swami, and J. Widom, "Clustering association rules," in *Proc. of IEEE ICDE*, 1997, pp. 220–231.
22. R.J. Miller and Y. Yang, "Association rules over interval data," *Proc. of the ACM SIGMOD*, pp. 452–461, 1997.
23. R. Ng and J. Han. "Efficient and effective clustering methods for spatial data mining," in *Proc. of VLDB*, 1994.
24. W. Pedrycz, "Fuzzy sets technology in knowledge discovery," *Fuzzy Sets and Systems* vol. 98, pp. 279–290, 1998.
25. R. Srikant and R. Agrawal. "Mining quantitative association rules in large relational tables," *Proc. of ACM SIGMOD*, 1996, pp. 1–12.
26. W. Wang and S.M. Bridges, "Genetic algorithm optimization of membership functions for mining fuzzy association rules," in *Proc. of the International Conference on Fuzzy Theory & Technology*, 2000, pp. 131–134.
27. R.R. Yager, "Fuzzy summaries in database mining," in *Proc. of the Conference on Artificial Intelligence for Application*, 1995, pp. 265–269.
28. S. Yue, et al., "Mining fuzzy association rules with weighted items," *Proc. of IEEE SMC*, 2000, pp. 1906–1911.
29. L.A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.
30. W. Zhang, "Mining fuzzy quantitative association rules," *Proc. of IEEE ICTAI*, pp. 99–102, 1999.