Facilitating Fuzzy Association Rules Mining by Using Multi-Objective Genetic Algorithms for Automated Clustering

Mehmet KAYA
Department of Computer Engineering
Fırat University
23119, Elazığ, Turkey
kaya@firat.edu.tr

Reda ALHAJJ
ADSA Lab, Department of Computer Science
University of Calgary
Calgary, Alberta, Canada
alhajj@cpsc.ucalgary.ca

Abstract

In this paper, we propose an automated clustering method based on multi-objective genetic algorithms (GA); the aim of this method is to automatically cluster values of a given quantitative attribute to obtain large number of large itemsets in low duration (time). We compare the proposed multi-objective GA-based approach with CURE-based approach. In addition to the autonomous specification of fuzzy sets, experimental results showed that the proposed automated clustering exhibits good performance over CURE-based approach in terms of runtime as well as the number of large itemsets and interesting association rules.

1. Introduction

An association rule is an implication $X \Rightarrow Y$, where both X and Y are sets of attributes or items; it is interpreted as: "for a specified fraction of the existing transactions, a particular value of X determines the value of Y as another particular value under a certain confidence". Support and confidence are the major factors in measuring 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 interesting association rules that satisfy user-specified minimum support and confidence*.

Quantitative association rules mining is essential because numerical attributes typically take many distinct values. The support for any particular value is likely low, while the support for intervals is much higher, e.g., [5, 6, 8]. However, 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.

Some work has recently been done on the use of fuzzy sets in discovering association rules for quantitative attributes, e.g., [1, 4, 9, 10, 12]. However, in existing approaches fuzzy sets are either supplied by an expert or determined by applying an existing clustering algorithm. The former is not realistic, in general, because it is extremely hard for an expert to specify fuzzy sets. The latter approaches have not produced satisfactory results. They have not considered the optimization of membership functions; a user specifies the number of fuzzy sets and membership functions are tuned accordingly.

Motivated by this, we propose a clustering method that employs multi-objective GA for the automatic discovery of membership functions used in determining fuzzy quantitative association rules. Our approach optimizes the number of fuzzy sets and their ranges according to multi-objective criteria in a way to maximize the number of large itemsets with respect to a given minimum support value. So, we defined two objective parameters in terms of large itemsets and the time required to determine fuzzy sets. Actually, these two parameters are in conflict with each other. So, we use a GA with multiple objective optimization capabilities known as *Pareto GA* [11].

Experimental results on 100K transactions extracted from the adult data of United States census in year 2000 show the efficiency and effectiveness of the proposed approach. Also, we have demonstrated the superiority of the proposed approach, in terms of the number of produced large itemsets and interesting association rules, over semi-automated CURE clustering based approach [2].

The rest of this paper is organized as follows. Fuzzy quantitative association rule is defined in Section 2. Our approach of utilizing GA to determine membership functions is described in Section 3. Determining membership functions for CURE clustering is discussed in Section 4. The fuzzy association rules mining process is presented in Section 5. Experimental results are given in Section 6. Section 7 includes a summary and the conclusions.

2. Fuzzy Association Rules

Consider a database of transactions $T=\{t_1, t_2,\ldots,t_n\}$, where each transaction t_j represents the j-th tuple in T. We use $I=\{i_1, i_2,\ldots,i_m\}$ to represent all attributes that appear in T; each quantitative attribute i_k is associated with at least two fuzzy sets. 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. The obtained value falls in the interval [0, 1], with the lower bound 0 strictly indicates "not a member", while the upper bound 1 indicates "total membership"; all other values between 0 and 1, exclusive, specify a "partial membership" degree. Finally, we use the following form for fuzzy association rules.

If $Q=\{u_1, u_2, \ldots, u_p\}$ is $F_1=\{ f_{1_1}, f_{1_2}, \ldots, f_{1_p} \}$ then

$R=\{v_1, v_2, \ldots, v_q\}$ is $F_2=\{ f_{2_1}, f_{2_2}, \ldots, f_{2_q} \}$,

Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)
0-7695-1978-4/03 $ 17.00 © 2003 IEEE

where $Q \subset I$ and $R \subset I$ are itemsets with $Q \cap R = \phi$, F_1 and F_2, respectively, contain the fuzzy sets associated with corresponding attributes in Q and R, i.e., f_{1i} is a fuzzy set related to attribute u_i and f_{2j} is related to attribute v_j.

Finally, for a rule to be interesting, it should have enough support and high confidence value. The basic step in the whole process is specifying corresponding fuzzy sets for each quantitative attribute. Our approach to automate this process is described next in Section 3.

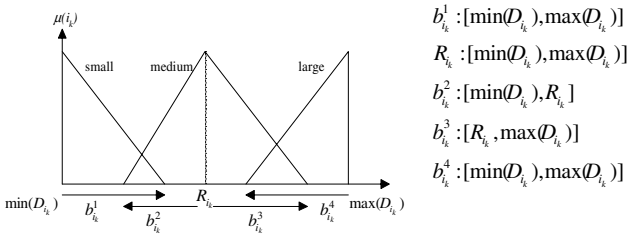

Figure 1 Membership functions and base variables of attribute i_k

3. Multi-Objective GA for Automated Clustering

In general a multi-objective optimization problem includes a set of *a* parameters (called decision variables), a set of *b* objective functions, and a set of *c* constraints; objective functions and constraints are functions of the decision variables. The optimization goal is expressed as:

$$\begin{aligned} &\min/\max \quad y = f(x) = (f_1(x), f_2(x),..., f_b(x)) \\ &\text{subject to} \quad e(x) = (e_1(x), e_2(x),..., e_c(x)) \le 0 \\ &\text{where} \quad x = (x_1, x_2,..., x_a) \in X \\ &\qquad\quad y = (y_1, y_2,..., y_b) \in Y \end{aligned}$$

where, x is the decision vector, y is the objective vector, X is the decision space, and Y is the objective space; the constraints $e(x) \le 0$ determine the set of feasible solutions.

In this paper, we considered the number of large itemsets and the gain of time, inverse of the time required to find all large itemsets in a given database as objective functions. It is assumed that each of the *n* components of the objective vector is to be maximized. A solution defined by corresponding decision vector can be *better* than, *worse*, or *equal* to; but also *indifferent* from another solution with respect to the objective values. Here, *better* means a solution is at least better in one objective and not worse in any objective than another solution. Using this concept, an optimal solution can be defined as: *a solution not dominated by any other solution in the search space*. Such a solution is called *Pareto optimal*, and the entire set of optimal trade-offs is called *Pareto-optimal set* [11].

In our approach, each individual represents the base values of membership functions for a quantitative attribute from the given database. In our experiments, we used membership functions in triangular shape.

To illustrate the encoding scheme utilized in this study, consider a quantitative attribute, say i_k, having 3 fuzzy sets, the corresponding membership functions and their base variables are shown in Figure 1. Each base variable takes finite values. For instance, the search space of the base value $b^1_{i_k}$ lies between the minimum and maximum values of attribute i_k, denoted $\min(D_{i_k})$ and $\max(D_{i_k})$, respectively. Enumerated next to Figure 1 are the search intervals of all the base values and the intersection point R_{i_k} of attribute i_k.

We used 8 quantitative attributes in the experiments of this study and assumed that each attribute can have at most 7 fuzzy sets. So, a chromosome consisting of the base lengths and the intersecting points is represented in the form:

$$w_{i_1} b^1_{i_1} b^{12}_{i_1} R^1_{i_1} b^2_{i_2} b^3_{i_1} R^2_{i_1} b^4_{i_1} b^5_{i_1} R^3_{i_1} b^6_{i_1} b^7_{i_1} R^4_{i_1} b^8_{i_1} b^9_{i_1} R^5_{i_1} b^{10}_{i_1} b^{11}_{i_1} \ldots w_{i_8} b^1_{i_8} b^{12}_{i_8} \ldots R^5_{i_8} b^{10}_{i_8} b^{11}_{i_8}$$

where gene w_{i_j} denotes the number of fuzzy sets for attributes i_j. If the number of fuzzy set is 2, then while decoding the individual, the first two base variables are considered and the others are omitted. However, if w_{i_j} is 3, then the next three variables are also taken into account. So, as long as the number of fuzzy set increases, the number of variables to be taken into account is enhanced too.

According to this encoding method, the number of variables needed to be found for each attribute can be generalized as $3(w-2)+2$, where w be the number of fuzzy sets for a given attribute. For instance, two variables need to be tuned for a quantitative attribute with w=2 fuzzy sets; and for the case of w=3 fuzzy sets, the number of variables to be tuned increases to 5, as illustrated by the above example.

In this study, we used real-valued coding, where chromosomes are represented as floating point numbers and their genes are the real parameters. While the value of a gene is reflected under its own search interval, the following formula is employed:

$$b^k_{i_j} = \min(b^k_{i_j}) + \frac{g}{g_{\max}} (\max(b^k_{i_j}) - \min(b^k_{i_j}))$$

where g is the value of the gene in search, $g_{\max}$ is the maximum value that gene g may take, $\min(b^k_{i_j})$ and $\max(b^k_{i_j})$ are the minimum and the maximum values of the reflected area, respectively. Also, we used Pareto-based ranking procedure, where the rank of an individual is the number of solutions encoded in the population by which its corresponding decision vector is dominated, as illustrated in Figure 2. Note that Pareto-based techniques seem to be most popular and effective in the field of evolutionary multi-objective optimization.

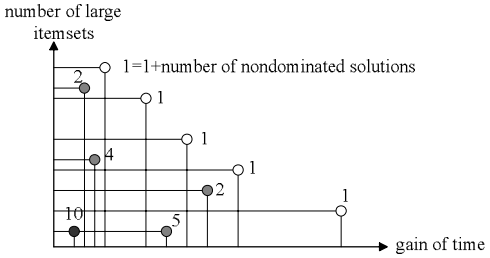

Figure 2 Fitness assignment in Pareto-based ranking

Individuals who are strong according to parent selection policy are candidates to form a new population. Many selection procedures are currently in use. However, we adapted the *elitism* policy in our experiments. Finally, after selecting chromosomes with respect to the evaluation

function, genetic operators such as, crossover and mutation, are applied to these individuals.

Crossover takes as input 2 individuals, selects a random point, and exchanges the subindividuals behind the selected point to form new individuals. Since the length of the chromosomes is considerably large in our approach, we used the multi-point crossover strategy with the crossover points determined randomly. On the other hand, mutation means a random change in the information of an individual. It is traditionally performed in order to increase the diversity of the genetic information. A probability test depends on the following condition to determine whether a mutation will be carried out or not.

4. Specifying Membership Functions for CURE

The process of CURE can be summarized as follows. Starting with individual values as individual clusters, at each step the closest pair of clusters are merged to form a new cluster. This is repeated until only k clusters are left. As a result, 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. Here, note that in the process to obtain the membership functions by CURE clustering algorithm, the number of clusters, i.e., number of fuzzy sets should be given by the user beforehand. To overcome this restriction, we integrated a GA with CURE clustering approach.

A GA finds the most appropriate number of clusters according to a predefined fitness function. In the GA process used in this study, each variable holds the number of fuzzy sets only. This is because CURE clustering algorithm itself adjusts the base value of the membership functions.

5. Mining Fuzzy Association Rules

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 all 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 $S_{<Z,F>}$: $S_{<Z,F>} = \frac{\sum_{t_i \in T} \prod_{z_j \in Z} \mu_{z_j}(f_j \in F, t_i[z_j])}{|T|}$, where $|T|$ denotes the number of transactions in database T.

This way, the problem of mining all fuzzy association rules converts to generating each rule whose confidence is larger than the user specified minimum confidence. Explicitly, each large itemset, say L, is used in deriving all association rules $(L-S) \Rightarrow S$, for each $S \subset L$. The strong association rules discovered are chosen from among all the generated possible association rules by considering only rules with confidence over a pre-specified minimum confidence. However, not all of these rules are interesting enough to be presented to the user. Whether a rule is interesting or not can be judged either subjectively or objectively. Ultimately, only the user 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 data can be used as one step towards the goal of weeding out presenting uninteresting rules to the user. To help filtering out misleading strong association rules and to give each rule a more precise characterization, the interestingness of a rule $Q \Rightarrow R$, denoted $I(Q \Rightarrow R)$, is defined as: $I(Q \Rightarrow R) = \frac{S(Q,R)}{S(Q).S(R)}$

A rule is filtered out if its interestingness is less than 1, since the nominator is the actual likelihood of both Q and R being present together and the denominator is the likelihood of having the two attributes being independent. This process will help in returning only rules having positive interestingness, and hence the size of the result is reduced.

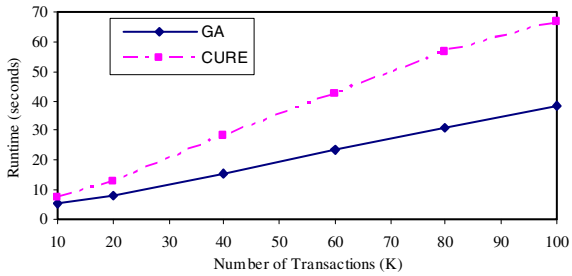

Figure 3 The runtime required to find all large itemsets

6. Experimental Results

We conducted some experiments to demonstrate the effectiveness of the proposed multi-objective GA-based clustering method. Further, the superiority of the new approach has been demonstrated by a comparison with CURE clustering based approach. All of the experiments have been conducted on a Pentium III, 1.4 GHz CPU with 512 MB of memory and running Windows 2000. As experiment data, we used 100K transactions from the adult data of United States census in 2000; we concentrated our analysis on 8 quantitative attributes. Further, in all the experiments, the GA process started with a population of 80 individuals for the GA-based approach and 30 individuals for the CURE-based approach. The maximum number of generations has been fixed at 500 as the termination criteria for the developed GA programs. Finally, in all the experiments in which GA have been used, the minimum support is set to 10%, unless otherwise specified, and the maximum number of fuzzy sets has been specified as 7.

The first experiment compares the runtime of the two clustering approaches to find large itemsets for different numbers of transactions, varying from 10K to 100K. The runtime here represents the duration, i.e., the time required to find all large itemsets after the number and ranges for the fuzzy sets have been determined by employing the corresponding method. The results are reported in Figure 3, where the two approaches are labeled as GA and CURE, to represent the proposed multi-objective GA-based clustering approach and CURE clustering based approach, respectively.

The former approach employs multi-objective GA to decide on the number of fuzzy sets as well as to optimize the ranges of membership functions, while the latter uses GA only to find the number of fuzzy sets. As a result of this experiment, it has been observed that the GA-based solution outperforms the CURE-based solution for all numbers of transactions, and both methods are scalable on the number of transactions.

The second experiment compares the total runtime required for both methods to find optimum fuzzy sets for different numbers of transactions. The results are reported in Figure 4, which demonstrates that both approaches scale well on the number of transactions. Extra runtime in the proposed method is spent on optimizing membership functions.

The third experiment utilized all the 100K transactions to compare the change in the number of large itemsets for different values of minimum support. The results are reported by the curves plotted in Figure 5, where it can be easily observed that the GA-based approach finds larger number of large itemsets than CURE based approach; this is quite consistent with our intuition, simply because the GA-based approach puts more effort on the optimization process and this has been reflected into finding better results than classical clustering approaches, like CURE.

The last experiment investigates the correlation between minimum confidence and the number of interesting association rules discovered. The results are plotted in Figure 6; the same interpretation stated for the curves of the large itemsets experiments is valid and can be repeated here.

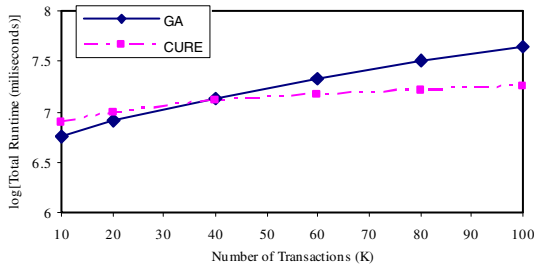

Figure 4 Total runtime required to find optimum fuzzy sets

7. Summary and Conclusions

In this paper, we proposed a multi-objective GA based clustering method, which automatically adjusts the fuzzy sets to provide large number of large itemsets in low duration. This is achieved by tuning together, for each quantitative attribute, the number of fuzzy sets and the base values of the membership functions. In addition, we demonstrated through experiments that using multi-objective GA has 3 important advantages over CURE. First, the number of clusters for each quantitative attribute is determined automatically. Thus, we implemented an autonomous structure for mining fuzzy association rules. Second, the GA-based approach optimizes membership functions of quantitative attributes for a given minimum support value. So, it is possible to obtain more appropriate solutions by changing the minimum support value in the desired direction. Finally, the number of large itemsets and interesting association rules obtained using the GA-based approach are larger than those obtained by applying CURE. As a result, all these advantages show that multi-objective GA is more appropriate and can be used more effectively to achieve optimal solutions than the classical clustering algorithms described in the literature.

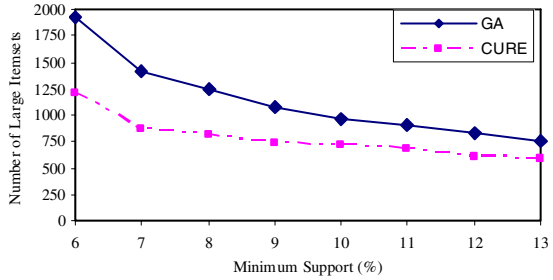

Figure 5 Number of large itemsets found

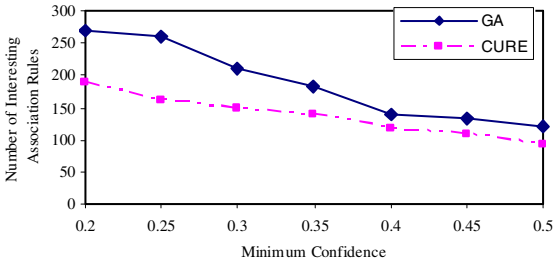

Figure 6 Number of association rules found

References

[1] K.C.C. Chan and W.H. Au, "Mining Fuzzy Association Rules," *Proc. of ACM CIKM,* pp.209-215, 1997.

[2] 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.

[3] J.H. Holland, Adaptation in Natural and Artificial Systems, The MIT Press, Cambridge, MA, MIT Press edition, 1992.

[4] 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.

[5] B. Lent, A. Swami and J. Widom, "Clustering Association Rules," *Proc. of IEEE ICDE*, pp.220-231, 1997.

[6] R.J. Miller and Y. Yang, "Association Rules over Interval Data," *Proc. of the ACM SIGMOD*, pp.452-461, 1997.

[7] R. Ng and J. Han. "Efficient and effective clustering methods for spatial data mining," *Proc. of VLDB,* 1994.

[8] R. Srikant and R. Agrawal. "Mining Quantitative Association Rules in Large Relational Tables," *Proc. of ACM SIGMOD,* pp.1-12, 1996.

[9] R.R. Yager, "Fuzzy Summaries in Database Mining," *Proc. of Artificial Intelligence for Application*, pp.265-269, 1995.

[10] W. Zhang, "Mining Fuzzy Quantitative Association Rules," *Proc. of IEEE ICTAI,* pp.99-102, 1999.

[11] E. Zitzler and L. Thiele, "Multi-objective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach," *IEEE TEC*, Vol.3, pp.257-271, 1999.

[12] M. Kaya, R. Alhajj, F. Polat and A. Arslan, "Efficient Automated Mining of Fuzzy Association Rules," *Proc. of DEXA, 2002.*