Integrating Multi-Objective Genetic Algorithms into Clustering for Fuzzy Association Rules Mining

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 method to decide on the number of fuzzy sets and for the autonomous mining of both fuzzy sets and fuzzy association rules. We compare the proposed multi-objective GA based approach with: 1) CURE based approach; 2) Chien et al clustering approach. Experimental results on 100K transactions extracted from the adult data of United States census in year 2000 show that the proposed method exhibits good performance over the other two approaches in terms of runtime, number of large itemsets and number of association rules.

1. Introduction

In general, 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, 8, 9, 11]. However, in existing approaches fuzzy sets are either supplied by expert or determined by applying clustering algorithm. The former is not realistic because it is extremely hard for an expert to specify fuzzy sets. The latter approaches have not produced satisfactory results. They do not considered the optimization of membership functions; a user specifies the number of fuzzy sets and membership functions are tuned accordingly.

In this paper, 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. These two are in conflict with each other. So, we use a GA with multiple objective optimization capabilities known as *Pareto GA* [10].

Experimental results demonstrate the effectiveness of the proposed approach. Also, we compared the proposed approach, in terms of the number of produced large itemsets and interesting association rules, with CURE based approach [2] and Chien *et al* approach [3], which is an efficient hierarchical clustering algorithm based on variation of density to solve the problem of internal partitioning.

The rest of this paper is organized as follows. Fuzzy association rule is defined in Section 2. Utilizing GA to determine membership functions is described in Section 3. A brief overview of CURE based approach and Chien *et al* work is given in Section 4. Experimental results are given in Section 5. Section 6 is summary and conclusions.

2. Fuzzy Association Rules

Consider a database of transactions $T=\{t_1, t_2,\ldots,t_n\}$, where each 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 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. The value falls in the interval [0, 1], with the lower bound 0 strictly indicates "not a member", the upper bound 1 indicates "total membership"; and all other values between 0 and 1, exclusive, specify "partial membership". Finally, we use the following form for fuzzy association rule:

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 $F_2=\{f_{2_1}, f_{2_2}, \ldots, f_{2_q}\}$,

where $Q\subset I$ and $R\subset I$ are itemsets with $Q\bigcap R=\phi$, 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.

3. Multi-Objective GA for Automated Clustering

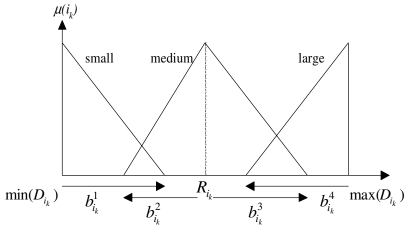
We consider as objective functions the number of large itemsets and the gain in time, inverse of the time required to find all large itemsets in a given database. It is assumed that each of the n components of the objective vector is to be maximized. An optimal solution can be defined as: *a solution 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* [10].

Each individual represents the base values of membership functions for a quantitative attribute from the given database. We used membership functions in triangular shape.

To illustrate the utilized encoding scheme, consider a quantitative attribute, say i_k, having 3 fuzzy sets, the corresponding membership functions and their base variables

IEEE COMPUTER SOCIETY

are shown in Figure 1. Each base variable takes finite values. For instance, the search space of the base value $b_{i_k}^1$ 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.



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

$R_{i_k} : [\min(D_{i_k}), \max(D_{i_k})]$

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

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

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

Figure 1 Membership functions and base variables 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:

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

where gene w_{i_j} denotes the number of fuzzy sets for attributes i_j. If the number of fuzzy sets 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 sets increases, the number of variables to be taken into account is enhanced too.

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: $b_{i_j}^k = \min(b_{i_j}^k) + \frac{g}{g_{\max}}(\max(b_{i_j}^k) - \min(b_{i_j}^k))$, where g is the value of the gene in search, $g_{\max}$ is the maximum value that gene g may take, $\min(b_{i_j}^k)$ and $\max(b_{i_j}^k)$ 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. Individuals who are strong according to parent selection policy are candidates to form a new population. We adapted the *elitism* selection 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.

To generate fuzzy association rules, the following formula is used to calculate the fuzzy support 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.

Each large itemset, say L, is used in deriving all association rules $(L\text{–}S) \Rightarrow S$, for each $S \subset L$. The strong association rules discovered are chosen 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.

4. Overview of CURE and Chein et al Work

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 values of the membership functions.

As Chien *et al* clustering approach is concerned, it is an efficient hierarchical clustering algorithm based on variation of density to solve the problem of interval partitioning. For this purpose, two main characteristics for clustering numerical data are defined first. Then, a reasonable interval can be generated automatically by giving a proper parameter α to determine the importance of relative closeness and relative inter-connectivity. The reader is referred to [3] for more details about this clustering technique.

5. Experimental Results

Effectiveness of the proposed approach has been demonstrated by comparison with two existing clustering approaches: CURE based approach and Chien *et al* work. We concentrate on testing the time requirements as well as changes in the main factors that affect the proposed clustering process: finding nondominated sets, number of large itemsets, and number of association rules. The experiments have been conducted on Pentium III 1.4 GHz CPU with 512 MB memory and running Windows 2000. As experiment data, we used 100K transactions from the adult data of US census in 2000; we concentrated our analysis on 8 quantitative attributes. Further, in all the experiments conducted in this study, the GA process started with a population of 80 individuals for the GA-based approach and 30 individuals for the other approach. As the termination criteria for the developed GA programs, the maximum number of generations has been fixed at 500. Finally, in all the experiments in which GA have been used, the minimum support was set to 10%, unless otherwise specified, and the maximum number of fuzzy sets has been specified as 7 for each of the three methods.

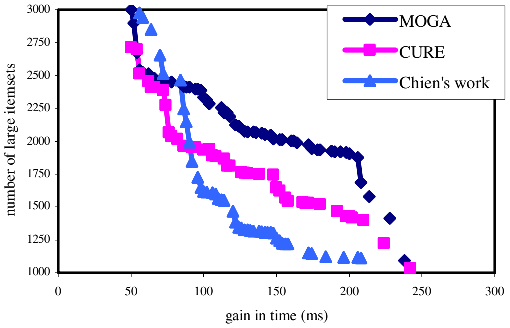


Figure 2 Nondominated set using 20K transactions

The first experiment is dedicated to find the nondominated set for each of the three different methods using 20K transactions. We decided to use 20K transactions because according to the next two experiments, the three approaches perform almost the same up to 20K transactions. The results are reported in Figure 2, where the three approaches are labeled as MOGA, CURE and Chein's work, to represent the proposed approach, CURE based approach and Chien *et al* work based approach, respectively. MOGA mostly outperforms the others for both objectives.

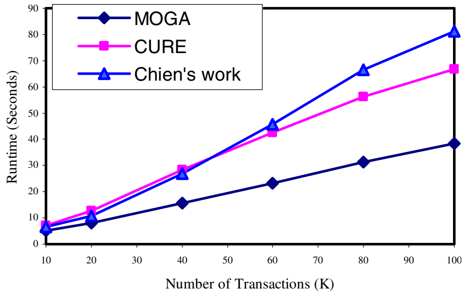


Figure 3 Runtime to find large itemsets for optimum case

The second experiment compares the runtime of the three approaches to find large itemsets for different numbers of transactions, varying from 10K to 100K. The results are reported in Figure 3. The runtime here represents the time required to find all large itemsets after the number of fuzzy sets and their ranges have been determined by employing the corresponding method. MOGA outperforms the other two approaches for all numbers of transactions. Finally, the curves plotted in Figure 3 demonstrate that the three methods are scalable with respect to the number of transactions.

The third experiment compares the runtime of the three approaches to find large itemsets when the number of fuzzy sets is fixed at 5. The results are reported by the curves plotted in Figure 4. We have decided on considering 5 fuzzy sets in this experiment because it is approximately the average number of fuzzy sets found by each of the three approaches. From Figure 4, the other two approaches outperform MOGA; the extra time in MOGA is spent on optimizing membership functions.

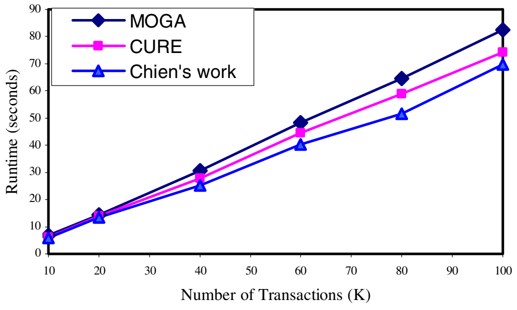


Figure 4 Runtime to find large itemsets for 5 fuzzy sets

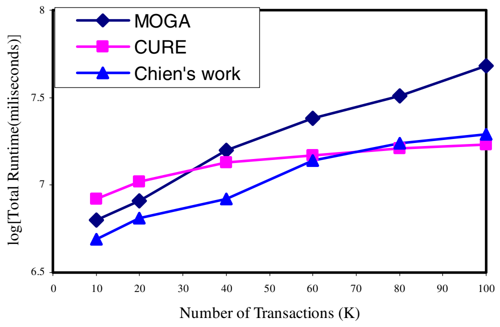


Figure 5 Total runtime required to find optimum fuzzy sets

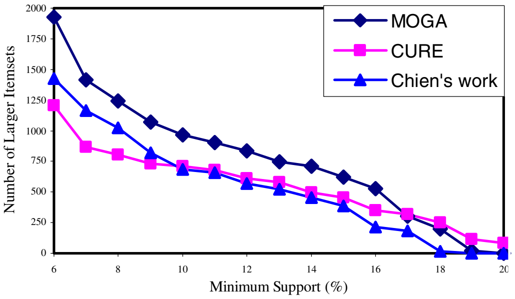


Figure 6 Number of large itemsets for optimum fuzzy sets

The fourth experiment compares the total runtime required for each of the three methods to find optimum fuzzy sets for different numbers of transactions. The results are reported in Figure 5; the total runtime of MOGA is smaller than the other two approaches up to around 40K transactions; after that, MOGA requires higher execution time than the other two approaches. The extra runtime is spent on optimizing membership functions. Figure 5 shows that all the three approaches scale well on the number of transactions.

The fifth experiment compares the change in the number of large itemsets for different values of minimum support. All the 100K transactions have been utilized and the optimum solution case has been considered. The results are reported by the curves plotted in Figure 6; MOGA finds larger number of large itemsets than the other two approaches. This is quite consistent with our intuition, simply because MOGA puts more effort on the optimization process and this has been reflected into finding better results than classical clustering approaches.

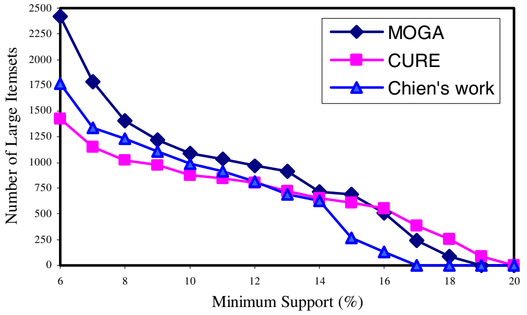


Figure 7 Number of large itemsets for 5 fuzzy sets

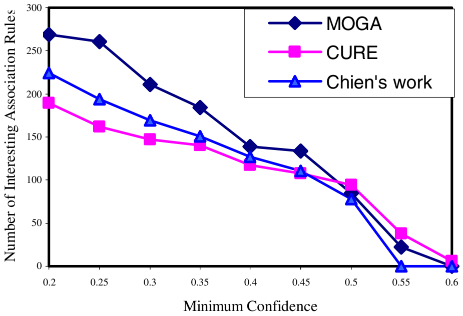


Figure 8 Number of association rules for optimum case

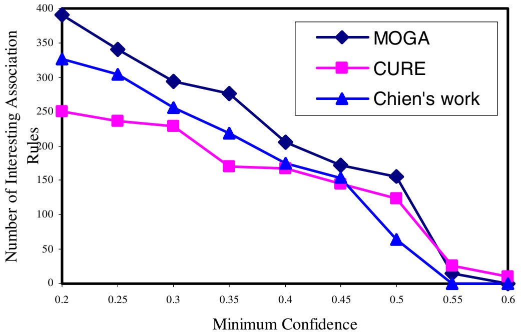


Figure 9 Number of association rules for 5 fuzzy sets

The sixth experiment is similar to the fifth but here 5 fuzzy sets are considered instead of the optimum case. The three curves plotted in Figure 7 show the number of large itemsets for different values of minimum support. For small values of minimum support, the difference between the three curves is larger than the difference for the optimum solution case shown in Figure 6. Finally, for the two cases plotted in Figures 6 and 7, the curves become smoother and the difference between them decreases as the minimum support increases. This is true because as the minimum support increase, the number of large itemsets decreases and approaches zero.

The last two experiments report the correlation between minimum confidence and number of interesting association rules discovered for each of the three approaches. Figure 8 reports the values for the optimum solution case. Figure 9 gives the results in case the number of clusters is set to 5 for each of the three methods. MOGA optimizes the ranges of the membership functions and the number of fuzzy sets in a way that outperforms the other two approaches.

6. 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 and Chien *et al* work. First, the number of clusters for each quantitative attribute is determined automatically. Second, the GA-based approach optimizes membership functions of quantitative attributes for a given minimum support. So, it is possible to obtain more appropriate solutions by changing the minimum support in the desired direction. Finally, the number of large itemsets and interesting association rules obtained using the GA-based approach are larger. As a result, all these advantages show that the proposed approach is more appropriate and can be used more effectively to achieve optimal solutions than the classical clustering algorithms described in the literature.

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] B.C. Chien, Z.L. Lin and T.P. Hong, "An Efficient Clustering Algorithm for Mining Fuzzy Quantitative Association Rules," *Proc. of IFSA World Congress and NAFIPS Conference*, Vol.3, pp.1306-1311, 2001.

[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.

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

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

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

[10] 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.

[11] M. Kaya and R. Alhajj, "Multi-Objective Genetic Algorithm Based Method for Mining Optimized Fuzzy Association Rules," *Proc. of IDEAL,* Springer, Aug. 2004.