Multi-objective Genetic Algorithm Based Method for Mining Optimized Fuzzy Association Rules

Mehmet Kaya[1] and Reda Alhajj[2]

[1] Department of Computer Engineering
Firat University, 23119 Elazig, Turkey
kaya@firat.edu.tr
[2] ADSA Lab & Department of Computer Science
University of Calgary, Calgary, Alberta, Canada
alhajj@cpsc.ucalgary.ca

Abstract. This paper introduces optimized fuzzy association rules mining. We propose a multi-objective Genetic Algorithm (GA) based approach for mining fuzzy association rules containing instantiated and uninstantiated attributes. According to our method, fuzzy association rules can contain an arbitrary number of uninstantiated attributes. The method uses three bjectives for the rule mining process: support, confidence and number of fuzzy sets. Experimental results conducted on a real data set demonstrate the effectiveness and applicability of the proposed approach.

1 Introduction

Mining association rules is one of the important research problems in data mining. We argue that equally important to the process of mining association rules is to mine optimized association rules. This has already been realized by some other researchers. The problem of finding optimized association rules was introduced by Fukoda et al [9]. They extended the results to the case where the rules contain two uninstantiated quantitative attributes on the left hand side [10]. Recently, Rastogi and Shim [11, 12] improved the optimized association rules problem in a way that allows association rules to contain a number of uninstantiated attributes.

The work presented in this paper reports the most recent results of our ongoing research on association rules mining. In this paper, we propose a novel method based on a multi-objective GA for determining the most appropriate fuzzy sets in fuzzy association rule mining in such a way that the optimized support and confidence satisfying rules will be obtained. Experimental results obtained using the Letter Recognition Database from the UCI Machine Learning Repository demonstrate that our approach performs well and gives good results even for a larger number of uninstantiated attributes.

The rest of the paper is organized as follows. Section 2 includes a brief overview of fuzzy association rules and introduces the multi-objective optimization problem. Section 3 gives our multi-objective GA based approach to mining optimized fuzzy association rules. The experimental results are reported in Section 4. Section 5 includes a summary and the conclusions.

Z.R. Yang et al. (Eds.): IDEAL 2004, LNCS 3177, pp. 758–764, 2004.

2 Fuzzy Association Rules and Multi-objective Optimization

Given a database of transactions T, its set of attributes I, 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 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 specified value of i_k as input. We use the following form for fuzzy association rules.

Definition 1: A fuzzy association rule is expressed as: If $Q=\{u_1, u_2, \ldots, u_p\}$ is $F_1=\{f_1, f_2, \ldots, f_p\}$ then $R=\{v_1, v_2, \ldots, v_q\}$ is $F_2=\{g_1, g_2, \ldots, g_q\}$, where Q and R are disjoint sets of attributes called itemsets, i.e., $Q \subset I$, $R \subset I$ and $Q \cap R = \phi$; F_1 and F_2 contain the fuzzy sets associated with corresponding attributes in Q and R, respectively, i.e., f_i is the fuzzy set related to attribute u_i and g_j is the fuzzy set related to attribute v_j.

A multi-objective optimization problem can be formalized as follows:

Definition 2: A multi-objective optimization problem includes, a set of *a* parameters (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:

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

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

In this paper, we considered the values of support and confidence utilized in the association rules mining process and number of fuzzy sets as objective functions. In this regard, a solution defined by the corresponding decision vector can be better than, worse, or equal to, but also indifferent from another solution with respect to the objective values. Better means a solution is not worse in any objective and better with respect to at least one objective than another. Using this concept, an optimal solution can be defined as: a solution which is 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 the Pareto-optimal set. In the next section, we describe how this multi-objective optimization method has been utilized to handle the mining of optimized fuzzy association rules.

3 The Proposed Multi-objective GA Based Approach

In this study, we use the support, confidence and number of fuzzy sets as objectives of the multi-objective GA. Our aim in using such an approach is to determine optimized

fuzzy association rules. Therefore, by using this approach, the values of support and confidence of a rule are maximized in large number of fuzzy sets. According to our intuition, stronger rules can be mined with larger number of fuzzy sets because more appropriate fuzzy rules can be found as the number of fuzzy sets is increased.

Throughout this study, we proposed two different encoding schemes. The first handles the rules with instantiated attributes. In such a case, each individual represents the base values of membership functions of a quantitative attribute in the database. In the experiments, we used membership functions in triangular shape.

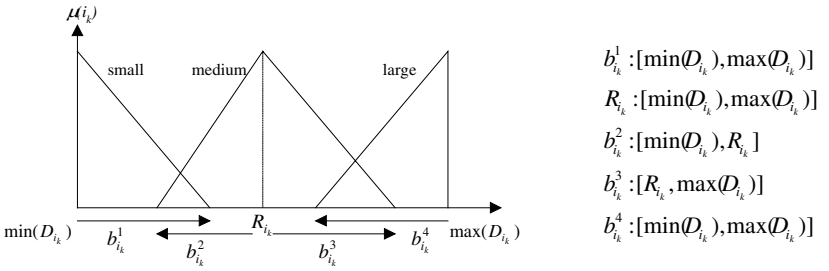

Fig. 1. Membership functions and base variables of attribute i_k.

To illustrate the encoding scheme utilized in this study, membership functions for a quantitative attribute i_k having 3 fuzzy sets and their base variables are shown in Figure 2. 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 2 are the search intervals of all the base values and the intersection point R_{i_k} of attribute i_k.

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 point is represented in the following form: $b_{i_1}^1 b_{i_1}^2 R_{i_1} b_{i_1}^3 b_{i_1}^4 b_{i_2}^1 b_{i_2}^2 R_{i_2} b_{i_2}^3 b_{i_2}^4 .. b_{i_m}^1 b_{i_m}^2 R_{i_m} b_{i_m}^3 b_{i_m}^4$

To illustrate the process, consider 5 quantitative attributes and assumed that each attribute can have at most 5 fuzzy sets. So, a chromosome consisting of the base lengths and the intersecting points is represented in the following form:

$$w_{i_1} b_{i_1}^1 b_{i_1}^{12} R_{i_1}^1 b_{i_2}^2 b_{i_1}^3 R_{i_1}^2 b_{i_1}^4 b_{i_1}^5 R_{i_1}^3 b_{i_1}^6 b_{i_1}^7 R_{i_1}^4 b_{i_1}^8 b_{i_1}^9 R_{i_1}^5 b_{i_1}^{10} b_{i_1}^{11} \ldots w_{i_5} b_{i_5}^1 b_{i_5}^{12} \ldots R_{i_5}^5 b_{i_5}^{10} b_{i_5}^{11}$$

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

In the case of uninstantiated rule, we associate two extra bits with each attribute. If these two bits are 00 then, the attribute appears in the antecedent part. However, if it is 11 then the attribute appears in the consequent part. Other combinations denote the absence of the attribute in either of those parts. So, we have 2m extra bits in each chromosome, where m is the number of attributes in the database. The difference of this second approach from the first one is that it finds the relevant rules along with their number of fuzzy sets and the base values.

In the experiments, we used binary coding method. While the value of a variable (gene) is reflected under its own search interval, the following formula is employed:

$$b_{i_j}^k = \min(b_{i_j}^k) + \frac{d}{2^L - 1}(\max(b_{i_j}^k) - \min(b_{i_j}^k))$$

where d is the decimal value of the variable in search, L is the number of bits used to represent a variable in the encoding scheme, $\min(b_{i_j}^k)$ and $\max(b_{i_j}^k)$ are, respectively, the minimum and the maximum values of the reflected area.

As mentioned earlier, in multi-objective problems, both fitness assignment and selection must allow for several objectives. One of the methods used for fitness assignments is to make direct use of the concepts of Pareto dominance. In this concept, fitness value is computed using their ranks, which are calculated from the non-dominance property of the chromosomes. The ranking step tries to obtain the non-dominated solutions. According to this step, if c_i chromosomes dominate an individual then its rank is assigned as c_i+1. This process continues until all the individuals are ranked. After each individual has fitness value, the individuals with the smallest rank constitutes the highest fitness. Finally, selection (we have adopted elitism policy in our experiments), replacement, crossover and mutation operators are applied to form a new population as in standard GA. Finally, the whole multi-objective GA process employed in this study can be summarized as:

Algorithm 1 (Mining optimized fuzzy association rules)

Input: Population size: *N;* Maximum number of generations: *G;* Crossover probability: p_c; Mutation rate: p_m

Output: Nondominated set: *S*

1. Set $P_0 = \phi$ and $t = 0$,
   For h=1 to N do: Choose $i \in I$, where i is an individual and I is the individual space, according to some probability distribution, and set $P_0 = P_0 + \{i\}$
2. For each individual $i \in P_t$: Determine the encoded decision vector and objective vector and alculate the scalar fitness value $F(i)$ with respect to the approach mentioned above.
3. Set $P' = \phi$
   For h=1 to N do: Select one individual $i \in P_t$ with respect to its fitness value $F(i)$ and set $P' = P' + \{i\}$
4. Set $P'' = \phi$
   For h=1 to N/2 do: Choose two individuals $i, j \in P'$ and remove them from P' and recombine i and j; the resulting offspring are $k, l \in I$; then insert k, l into P'' with probability p_c, otherwise insert i, j into P''
5. Set $P''' = \phi$
   For each individual $i \in P''$ do: Mutate i with mutation rate p_m. The resulting individual is $j \in I$, and set $P''' = P''' + \{j\}$.
6. Set $P_{t+1} = P'''$ and $t = t + 1$.
   If $t \geq G$ or another termination criterion is satisfied then return $S = p(P_t)$, where $p(P_t)$ gives the set of nondominated decision vectors in P_t. In other words, the set $p(P_t)$ is the nondominated set regarding P_t. Otherwise go to Step 2, i.e., execute steps 2 to 6.

4 Experimental Results

We apply the proposed multi-objective GA based approach to the Letter Recognition Database from the UCI Machine Learning Repository. The database consists of 20K samples and 16 quantitative attributes. We concentrated our analysis on only 5 quantitative attributes. In all the experiments in this study, the GA process starts with a population of 50 individuals for both approaches, with instantiated and uninstantiated rules. Further, crossover and mutation probabilities are taken, respectively, as 0.8 and 0.01, and 4-point crossover operator is utilized.

Table 1. Objective values for the optimized fuzzy instantiated rules.

Number of Fuzzy Sets Support (%) Confidence (%)
2
36.86
69.43
30.61
72.66
25.48
78.11
3
27.14
63.24
25.20
76.16
21.45
84.17
4
23.26
54.75
20.34
82.15
19.21
87.05
5
8.52
41.63
7.21
67.37
6.18
80.12

Table 2. Objective values for the optimized instantiated rules by using discrete method.

Number of Discrete Intervals Support (%) Confidence (%)
2
21.66
57.21
20.48
59.40
19.23
66.11
3
13.07
47.15
11.12
53.27
10.17
68.23
4
8.42
38.46
8.20
46.12
7.78
68.17
5
6.21
51.13
6.02
67.12
5.88
71.76

The first experiment is dedicated to find the non-dominated set of the proposed method for an instantiated rule at 20K. The results are reported in Table 1, where the values of support and confidence for some non-dominated solutions are given for four different numbers of fuzzy sets. From Table 1, it can be easily seen that as the number

of fuzzy sets increases, the support value of the instantiated rules decreases. This is true because a large number of sets will make quantities of an item in different transactions easily scatter in different sets. However, for each number of fuzzy sets, as the support value decreases, the confidence value increases because more specific rules are generated.

Table 3. Number of rules generated vs. number of generations.

Number of Generations	Number of Rules
250	136
500	182
750	204
1000	217
1250	223
1500	225
1750	225

The second experiment is dedicated for the case where the first experiment is repeated with discrete method instead of fuzzy sets. The results are reported in Table 2. An important point here is that the values of support and confidence are smaller than those of the fuzzy approach. This demonstrates an important feature of using fuzzy sets; they are more flexible than their discrete counterparts. As a result, stronger rules and larger number of rules can be obtained using fuzzy sets.

The final experiment is conducted to find the number of uninstantiated rules generated for different numbers of generations. We used stability of the rules as the termination criteria. The average results of 5 runs are reported in Table 3, from which it can be easily observed that the GA convergences after 1250 generations. In other words, it almost does not produce more rules. It is also observed that most of the rules include 2 quantitative attributes. Only 6 rules were obtained that contain all the attributes. In fact, most of the rules contain 2 attributes because a small number of attributes means the corresponding rule has a larger value of support, i.e., as the number of attributes in the rule increases, the support value of the rule decreases almost exponentially.

5 Summary and Conclusions

In this paper, we contributed to the ongoing research on association rules mining by proposing a multi-objective GA based method for mining optimized fuzzy association rules. Our approach uses three measures as the objectives of the method: support, confidence and number of fuzzy sets. The proposed method can be applied to two different cases: dealing with rules containing instantiated attributes and those with uninstantiated attributes. The former case finds only optimized sets of fuzzy rules, and the latter case obtains the most appropriate fuzzy sets along with uninstantiated attributes. The results obtained from the conducted experiments demonstrate the effectiveness and applicability of the optimized fuzzy rules over the discrete based rules with respect to the values of support and confidence. Currently, we are investigating the optimization of all fuzzy sets of the attributes in a single rule.

References

1. M. Delgado, N. Marin, D. Sanchez and M. A. Vila, "Fuzzy Association Rules: General Model and Applications", *IEEE TFS*, Vol.11, No.2, pp. 214-225, 2003.
2. M. Kaya, R. Alhajj, F. Polat and A. Arslan, "Efficient Automated Mining of Fuzzy Association Rules," *Proc. of DEXA,* 2002.
3. M. Kaya and R. Alhajj, "A Clustering Algorithm with Genetically Optimized Membership Functions for Fuzzy Association Rules Mining", *Proc. of Fuzz-IEEE*, St Louis, MO, 2003
4. M. Kaya and R. Alhajj, "Facilitating Fuzzy Association Rules Mining by Using Multi-Objective Genetic Algorithms for Automated Clustering", *Proc. of IEEE ICDM*, Melbourne, FL, 2003.
5. T. Fukuda, et al, "Mining Optimized Association Rules for Numeric Attributes", *Proc. of ACM SIGACT-SIGMOD-SIGART PODS*, 1996.
6. T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama, "Data Mining Using Two Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization", *Proc. of ACM SIGMOD*, 1996.
7. R. Rastogi and K. Shim, "Mining Optimized Support Rules for Numeric Attributes", *Information Systems*, Vol.26, pp.425-444, 2001
8. R. Rastogi and K. Shim, "Mining Optimized Association Rules with Categorical and Numeric Attributes", *IEEE TKDE*, Vol.14, No.1, pp.29-50, 2002.