Mining Optimized Fuzzy Association Rules Using Multi-Objective Genetic Algorithm

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

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

Abstract

Mining association rules is one of the important research problems in data mining. So far, some efficient techniques have been proposed to obtain association rules with respect to an optimal goal, such as: to maximize the number of large itemsets and rules or to satisfy certain values of support and confidence. Realizing the importance of optimized association rules in general, this paper introduces partial optimized fuzzy association rules mining. In this regard, we propose a multi-objective Genetic Algorithm (GA) based approach for mining fuzzy association rules. According to our method, fuzzy association rules can contain an arbitrary number of uninstantiated attributes. The method uses three measures as the objectives for the rule mining process; these are support, confidence and amplitude of fuzzy sets. Experimental results conducted on a real data set demonstrate the effectiveness and applicability of the proposed approach.

Keywords: association rules, data mining, fuzzy association rules, multi-objective GA, optimized association rules.

I. INTRODUCTION

Mining association rules form one of the most widely used techniques to discover correlations among attributes in a database. The problem of mining boolean association rules over basket data was first introduced by Agrawal et al [1]. The basic task in mining for association rules is to determine the correlation between items belonging to a transaction database. In general, every association rule must satisfy two user specified constraints, one is support and the other is confidence. The support of a rule $X \Rightarrow Y$ is defined as the fraction of transactions that contain $X \bigcup Y$, where X and Y are sets of items from the given database. The confidence is defined as the ratio $\frac{\text{support}(X \bigcup Y)}{\text{support}(X)}$. So, the target is to find all association rules that satisfy user specified minimum support and confidence values. Then, Agrawal and Srikant [2, 3] extended their pioneering work for the case of databases consisting of categorical and quantitative attributes.

The algorithm proposed for mining quantitative association rules discretizes the domains of quantitative attributes into intervals in order to reduce the domain into a categorical one. However, it is a difficult task to determine the right intervals without a priori knowledge. Furthermore, these intervals may not be concise and meaningful enough for human experts to easily obtain valuable knowledge from the discovered rules.

Instead of using sharp boundary intervals, some work have been recently done on using fuzzy sets in discovering association rules for quantitative attributes [4-7]. The rules obtained by this way are called fuzzy association rules. Fuzzy association rules are more understandable when meaningful linguistic terms are assigned to fuzzy sets. However, the main problem in existing approaches is that an expert must provide the required fuzzy sets of quantitative attributes and their corresponding membership functions. Also, it is not a realistic approach to expect experts to always provide the most appropriate fuzzy sets for mining fuzzy association rules. For this purpose, some efforts have been recently done to tackle this problem. These efforts are mainly classified into two different trends: the first trend is concerned with clustering methods and the second trend employs GA based approaches.

As the first trend is concerned, Fu et al [8] proposed an automated method to find fuzzy sets for the mining of fuzzy association rules; their method is based on CLARANS clustering algorithm [9]. After obtaining the *k* medoids for each quantitative attribute, these medoids are used to classify each quantitative attribute into *k* fuzzy sets. We have already developed a more efficient approach based on CURE clustering algorithm [10]. The work of Gyenesei [11] is another method that employs clustering techniques to find the fuzzy sets for each quantitative attribute. For this reason, he defined the goodness index for clustering scheme evaluation based on two criteria: compactness and separation. Then, the clustering process determines both the number and centers for the clusters. Finally, the corresponding membership functions for fuzzy sets of each quantitative attribute are generated. As the second trend is concerned, we have already developed two different methods based on GA [12, 13]. The two methods simply tune the base values of membership functions for each quantitative attribute with respect to a given criteria; the readers are referred to [12, 13] for more details.

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 [14]. The aim of the study is to generate association rules that contain a single uninstantiated condition on the left hand side and propose an approach to determine values for intervals of attributes such that the confidence or support of the rule is maximized. Then, they extended the results to the case where the rules contain two uninstantiated quantitative attributes on the left hand side [15]. Recently, Rastogi and Shim [16, 17] improved the optimized association rules problem in a way that allows association rules to contain a number of uninstantiated attributes.

In this paper, we concentrate on using multi-objective GA for mining partial optimized fuzzy association rules. Mainly, we propose a novel method based on a multi-objective GA for determining the most appropriate fuzzy sets whose number was prespecified in fuzzy association rule mining in such a way that the optimized support and confidence satisfying rules will be obtained. Such a rule is called partial optimized fuzzy association rule. The number of sets may change based on user request. Throughout this study, we used the number of set between 2 and 5. 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. Section 3 introduces the multi-objective optimization problem. Section 4 gives our multi-objective GA based approach for mining optimized fuzzy association rules. The experimental results are reported in Section 5. Section 6 includes a summary and the conclusions.

II. FUZZY ASSOCIATION RULES

In this section, we present a general overview of fuzzy association rules. So, let $T=\{t_1, t_2, \dots, t_n\}$ be a database of transactions; each transaction t_j represents the j-th tuple in T. We use $I=\{i_1, i_2, \dots, i_m\}$ 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. Explicitly, it is possible to define some fuzzy sets for attribute i_k with a membership function per fuzzy set such that each value of attribute i_k qualifies to be in one or more of the fuzzy sets specified for i_k. The degree of membership of each value of attribute i_k in any of 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.

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

Definition 1: A fuzzy association rule is expressed as:

If $Q=\{u_1, u_2, \dots, u_p\}$ is $F_1=\{f_1, f_2, \dots, f_p\}$ then
$R=\{v_1, v_2, \dots, v_q\}$ is $F_2=\{g_1, g_2, \dots, g_q\}$,

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.

Finally, as it is the case with classical rules, "Q is F_1" is called the antecedent of the rule while "R is F_2" is called the consequent of the rule. For a rule to be interesting, it should have enough support and high confidence value, larger than user specified thresholds.

III. MULTI-OBJECTIVE OPTIMIZATION

Contrary to the single objective optimization method, the multi-objective optimization method deals with simultaneous optimization of several incommensurable and often competing objectives such as performance and cost. For example, when the design of a complex hardware is considered, it is required for the cost of such a system to be minimized while the maximum performance is expected. If there is more than one objective criterion as in the example mentioned above, some of them can be considered as constraints in the problem. For example, while trying to optimize a system for large performance and low cost, the size of the system must not exceed given dimensions is a separate optimization criterion. By this way, a multi-objective optimization problem can be formalized as follows [18]:

Definition 2: A multi-objective optimization problem includes, in general, 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:

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

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

In this paper, we consider the values of support and confidence utilized in the association rules mining process and amplitude of fuzzy sets as objective functions.

The amplitude of fuzzy sets is computed as follows:

$$\textit{Fuzzy Set Amplitude} = \frac{\text{sum of maximum amplitudes of itemsets} - \text{sum of amplitudes of itemsets}}{\text{sum of maximum amplitudes of itemset}}$$

$$\text{sum of maximum amplitudes of itemset} = \sum_{i=1}^{k} \max(D_i) - \min(D_i)$$

$$\textit{sum of amplitudes of itemset} = \sum_{i=1}^{k} b_i - a_i$$

where, k is the number of attributes in the itemsets; and b_i and a_i are variables that are the parameters of the fuzzy sets corresponding to attribute i. Note that these are not overlapping sets.

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 as shown in Figure 1. Better means a solution is not worse in any objective and better with respect to at least one objective than another. For example, while the solution represented by point B is worse than the solution represented by point A, the solution represented by point C is better than that represented by point A. However, it cannot be said that C is better than D or vice versa. This is because one objective value of each point is higher than the other one. 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*, which is represented as dotted line in Figure 1.

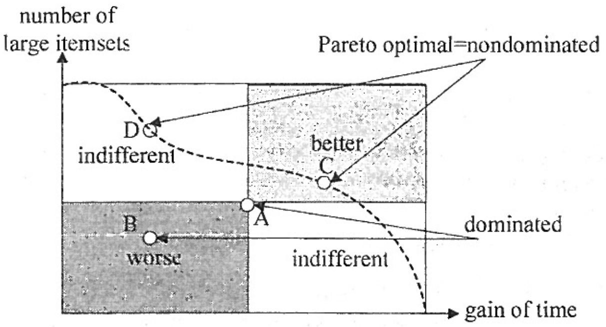

Fig. 1. The concept of Pareto optimality

Explicitly, the objectives in such an optimization problem are conflicting and cannot be optimized simultaneously. Instead, a satisfactory trade-off has to be found. Therefore, it is necessary to have a decision making process in which preference information is used in selecting an appropriate trade-off. In the next section, we describe how this multi-objective optimization method has been utilized to handle the mining of optimized fuzzy association rules.

IV. THE PROPOSED MULTI-OBJECTIVE GA BASED APPROACH

In this section, we describe the proposed method for mining optimized fuzzy association rules by employing a pareto-optimality based GA. We first present our encoding scheme and then define the fitness assignment and selection process. Finally, we give the algorithmic structure of the proposed approach.

A. Chromosome Encoding

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.

This subsection describes the generation of initial population, each individual represents the base values of the membership functions of a quantitative attribute in the database. In the experiments, we use membership functions in triangular shape.

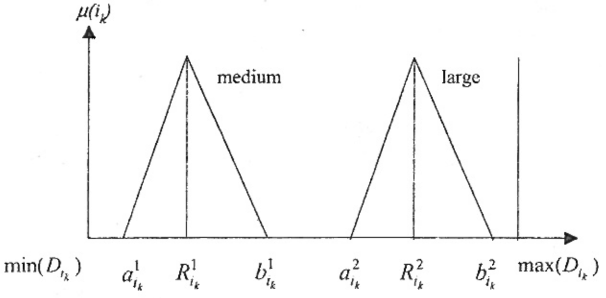

Fig. 2. Membership functions and their parameters of attribute i_k

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.

So, based on the assumption of having 2 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: $a^1_{i_1}R^1_{i_1}b^1_{i_1}a^2_{i_1}R^2_{i_1}b^2_{i_2}\ldots a^1_{i_m}R^1_{i_m}b^1_{i_m}a^2_{i_m}R^2_{i_m}b^2_{i_m}$

Since it is not possible to know a priori how many attributes will be necessary to create a good fuzzy association rule, this number has to be automatically adjusted by the GA based on the data being mined. Finally, the structure of the genome of an individual is illustrated in Figure 3.

$v_1 w_1 B_1$	$\cdots$	$v_j w_j B_j$	$\cdots$	$v_m w_m B_m$

Fig. 3. Representation of an individual

where, gene w_j denotes the number of fuzzy sets for attributes j. While decoding the individual for two fuzzy set, the first two base variables are considered and the others are omitted. However, if w_j is raised to 3, then the next three variables are taken into account as well.

So, as the number of fuzzy sets increases, the number of variables to be taken into account is enhanced too.

We associate two extra bits with each attribute j; v_j show the part in which attribute j appears. If these two bits are 00 then, the attribute appears in the antecedent part. However, 11 mean that the attribute appears in the consequent part. Other combinations denote the absence of the attribute in either of the two 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 use 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.

B. Fitness Assignment and Selection

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, the fitness value is computed using the 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, individuals with the smallest rank constitute 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. The whole multi-objective GA process employed in this study can be summarized as follows.

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
    a) Choose $i \in I$, where $i$ is an individual and $I$ is the individual space, according to some probability distribution.
    b) Set $P_0 = P_0 + \{i\}$
2 For each individual $i \in P_t$,
    a) determine the encoded decision vector and objective vector
    b) calculate the scalar fitness value $F(i)$ with respect to the approach mentioned above.
3 Set $P' = \phi$
    For h=1 to N do
    a) Select one individual $i \in P_t$ with respect to its fitness value $F(i)$
    b) Set $P' = P' + \{i\}$
4 Set $P'' = \phi$
    For h=1 to N/2 do
    a) Choose two individuals $i, j \in P'$ and remove them from $P'$
    b) Recombine $i$ and $j$; the resulting offspring are $k, l \in I$.
    c) 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
    a) Mutate $i$ with mutation rate $p_m$. The resulting individual is $j \in I$
    b) Set $P''' = P''' + \{j\}$.
6 Set $P_{t+1} = P'''$ and $t = t + 1$.
    If $t \geq G$ or the threshold based 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.
```

Algorithm 1 starts by selecting individuals to the initial population. Then the following process is repeated until a termination condition is satisfied or the prespecified maximum number of generations is achieved. The encoded decision vector and objective vector as well as the fitness value are all determined for each selected individual. Existing individuals are used in generating new ones by applying cross-over and mutation. Individuals survive based on their fitness and are used in the process. By this way, the nondominated set is determined and the target is achieved.

V. 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 8 quantitative attributes. In all the experiments conducted in this study, the GA process starts with a population of 50. 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 partial optimized fuzzy rules

Number of Fuzzy Sets	Support (%)	Confidence (%)
2	32.86	71.48
	31.24	74.56
	28.46	78.85
3	27.44	60.32
	26.34	74.48
	23.27	78.77
4	24.48	66.15
	17.56	79.17
	15.87	83.52
5	11.78	52.74
	8.64	65.11
	8.42	74.37

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

Number of Discrete Intervals	Support (%)	Confidence (%)
2	24.18	63.74
	21.48	72.11
	18.66	73.46
3	17.27	53.21
	15.23	68.01
	14.89	71.12
4	12.21	56.00
	11.58	68.47
	9.74	74.19
5	8.46	50.15
	7.58	59.70
	6.70	63.78

The first experiment is dedicated to find the non-dominated set of the proposed method for different number of fuzzy sets 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
200	157
400	213
600	234
800	256
1000	258
1200	259
1400	259

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 third experiment is conducted to find the number of fuzzy 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 1000 generations. In other words, it almost does not produce more rules. It is also observed that most of the rules include 2 quantitative attributes. None of the obtained rules contain all the attributes. In fact, most of the rules contain 2 attributes because a small number of attributes means that 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.

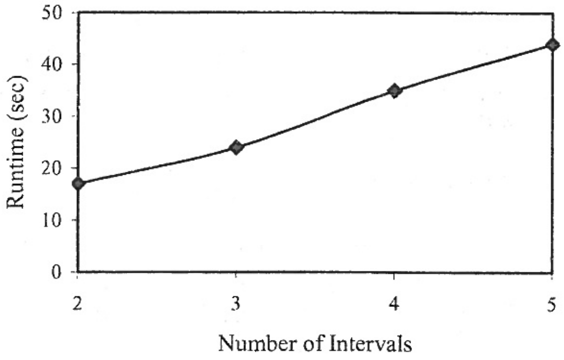

Fig. 4. Runtimes in different number of fuzzy sets

The final experiment measures the runtime for different number of fuzzy sets. It can be easily seen from Figure 4 that as the number of fuzzy set increases, the runtime raises almost linearly.

VI. SUMMARY AND CONCLUSIONS

Since Agrawal's pioneer work, association rules mining is gaining more interest and is becoming an important research area in data mining. This is evident by the amount of research results reported in the literature and the many algorithms proposed to mine association rules with respect to an optimal goal. In this paper, we contributed to the ongoing research by proposing a multi-objective GA based method for mining partial optimized fuzzy association rules. Our approach uses three measures as the objectives of the method: support, confidence and amplitudes of fuzzy sets. The proposed method can be applied to two different cases: dealing with rules that have fixed number of sets and those with changing number of sets. 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 multi-level optimization of fuzzy association rules.

VII. 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] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," *Proc. of the International Conf. on Very Large Data Bases*, pp. 487-499, Santiago, Chile, Sept, 1994.

[3] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," *Proc. of ACM SIGMOD*, pp.1-12, 1996.

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

[5] 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.

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

[7] M. Delgado, N. Marin, D. Sanchez and M. A. Vila, "Fuzzy Association Rules: General Model and Applications," *IEEE Transactions on Fuzzy Systems*, Vol.11, No.2, pp. 214-225, 2003.

[8] A.W.C. Fu, et al., "Finding Fuzzy Sets for the Mining of Association Rules for Numerical Attributes," *Proc. of the International Symposium of Intelligent Data Engineering and Learning,* pp.263-268, 1998.

[9] R. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," *Proc. of the International Conference on Very Large Databases,* 1994.

[10] M. Kaya, R. Alhajj, F. Polat and A. Arslan, "Efficient Automated Mining of Fuzzy Association Rules," *Proc. of the International Conference on Database and Expert Systems with Applications,* 2002.

[11] A. Gyenesi, "Determining Fuzzy Sets for Quantitative Attributes in Data Mining Problems," *Proc. of Advance in Fuzzy Systems and Evol. Comp.*, pp.48-53, 2001.

[12] M. Kaya and R. Alhajj, "A Clustering Algorithm with Genetically Optimized Membership Functions for Fuzzy Association Rules Mining," *Proc. of IEEE International Conference on Fuzzy Systems*, St Louis, MO, 2003

[13] M. Kaya and R. Alhajj, "Facilitating Fuzzy Association Rules Mining by Using Multi-Objective Genetic Algorithms for Automated Clustering," *Proc. of IEEE International Conference on Data Mining*, Melbourne, FL, 2003.

[14] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama, "Mining Optimized Association Rules for Numeric Attributes," *Proc. of ACM SIGACT-SIGMOD-SIGART Symposium Principles of Database Systems*, 1996.

[15] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama, "Data Mining Using Two Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization," *Proc. ACM SIGMOD Conf. Management of Data*, June 1996.

[16] R. Rastogi and K. Shim, "Mining Optimized Support Rules for Numeric Attributes," *Information Systems*, Vol.26, pp.425-444, 2001

[17] R. Rastogi and K. Shim, "Mining Optimized Association Rules with Categorical and Numeric Attributes," *IEEE Transactions on Knowledge and Data Engineering*, Vol.14, No.1, pp.29-50, 2002.

[18] E. Zitzler and L. Thiele, "Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach," *IEEE Transaction on Evolutionary Computation*, Vol.3, No.4, pp.257-271, 1999.