

A Multiple-Level Genetic-Fuzzy Mining Algorithm

Chun-Hao Chen

Department of Computer Science and
Information Engineering
Tamkang University
Taipei, Taiwan, R.O.C.
chchen@mail.tku.edu.tw

Tzung-Pei Hong

Department of Computer Science and
Information Engineering
National University of Kaohsiung
Kaohsiung, Taiwan, R.O.C.
tphong@nuk.edu.tw

Yeong-Chyi Lee

Department of Information
Management
Cheng Shiu University
Kaohsiung, Taiwan, R. O. C.
yeongchyi@csu.edu.tw

Abstract—In this paper, we propose a multiple-level genetic-fuzzy mining algorithm for mining membership functions and fuzzy association rule on multiple-concept levels. It first encodes the membership functions of each item class (category) into a chromosome according to the given taxonomy. The fitness value of each individual is then evaluated by the summation of large 1-itemsets of each item in different concept levels and the suitability of membership functions in the chromosome. After the GA process terminates, a better set of multiple-level fuzzy association rules can then be expected with a more suitable set of membership functions. Experimental results on a simulation dataset also show the effectiveness of the algorithm.

Keywords- data mining, genetic algorithm, multiple-concept levels, membership function, fuzzy association rule.

I. INTRODUCTION

Data mining is most commonly used in attempts to induce association rules from transaction data [2]. An association rule is an expression $X \rightarrow Y$, where X and Y are a set of items. It means in the set of transactions, if all the items in X exist in a transaction, then Y is also in the transaction with a high probability. A variety of mining approaches based on the Apriori algorithm [2] have also been proposed [4, 21]. However, most of them are focus on finding association rules on a single-concept level. Thus, many approaches have also been proposed for mining multiple-concept-level association rules to discovery more specific and important knowledge from data [8, 22].

Most of the mentioned approaches above focused on binary valued transaction data. Transaction data in real-world applications usually consist of quantitative values. Thus, fuzzy data mining algorithms have been proposed for handling quantitative transactions and mining fuzzy association rules [5, 11, 17, 24]. As to multiple-concept levels fuzzy mining, according to the given fuzzy generalization hierarchies, Lee first proposed a generalized fuzzy quantitative association rules mining algorithm [16]. In 2003, Hong et al. then proposed a multiple-level fuzzy association rule mining approach [12]. It first transforms quantitative transactions into fuzzy values by using the predefined membership functions and taxonomy. Then, a top-down progressively deepening approach is used to finding large itemsets and rules. Additionally, in order to reduce the time complexity, only the linguistic term with the maximum cardinality of each item is used in later mining processes. In [14], Kaya et al. proposed a weight fuzzy rule mining approach based on Hong's approach. In 2008, Lee et al. proposed a fuzzy mining algorithm for discovering generalized

fuzzy association rules with multiple supports of items for extracting implicit knowledge from quantitative transaction data [18, 19].

Those mentioned approaches assumed the membership functions to be known in advance. The given membership functions may, however, have a critical influence on the final mining results. In recent years, the Genetic-Fuzzy Mining (GFM) approaches have then proposed for deriving appropriate membership functions and mining fuzzy association rules [3, 6, 10, 13, 15]. Kaya *et al.* proposed a GA-based approach to derive a predefined number of membership functions for getting a maximum profit within a user specified interval of minimum supports [13]. Hong *et al.* also proposed a genetic-fuzzy data-mining algorithm for extracting both association rules and membership functions from quantitative transactions [10]. Alcalá-Fdez *et al.* then modified Hong's approach to propose an enhanced approach based on the 2-tuples linguistic representation model [3]. Besides, multi-objective genetic-fuzzy mining approaches have also been proposed [6, 15].

However, those mentioned GFM algorithms are focus on one single-concept level. To the best of our knowledge, there is no GFM algorithm for multiple-concept levels until now. In this paper, we thus propose a multiple-level GFM algorithm for mining membership functions and fuzzy association rule on multiple-concept levels. According to the given taxonomy, it first encodes membership functions of each item class (category) into chromosome. The fitness value of each individual is then evaluated by the summation of large 1-itemsets of each item in different concept levels and the suitability of membership functions in the chromosome. After the GA process terminates, a better set of multiple-level fuzzy association rules can then be expected with a more suitable set of membership functions.

II. A MULTIPLE-LEVEL GENETIC-FUZZY MINING FRAMEWORK

In this section, the fuzzy and GA concepts are used to discover both useful association rules and suitable membership functions from quantitative values. We propose a Multiple-level Genetic-Fuzzy Mining (Multiple-level GFM) framework for searching membership functions suitable for multiple-level mining problems and then use the final best set of membership functions to mine multiple-level fuzzy association rules. The proposed framework is shown in Figure 1.

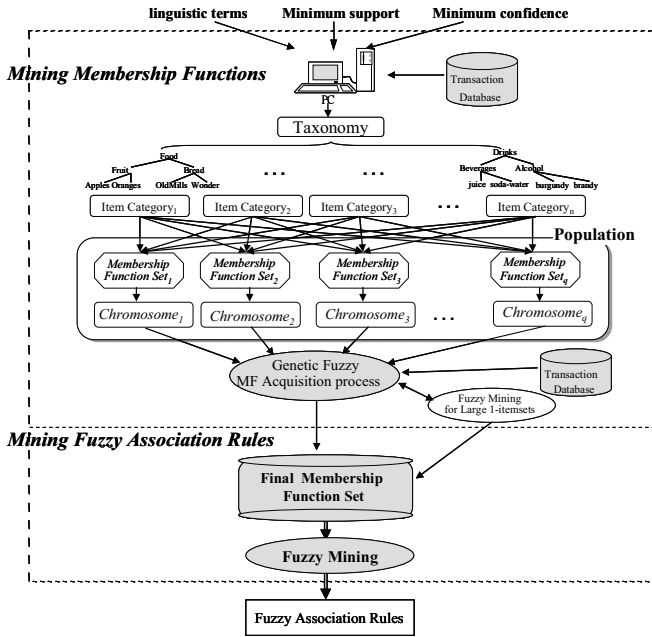


Figure 1. The Multiple-level GFM framework

The proposed framework consists of two phases, namely mining membership functions and mining fuzzy association rules phases. It maintains a population of sets of membership functions, and uses the genetic algorithm to automatically derive the resulting one. In the first phase, according to the given taxonomy, it first transforms each set of membership functions of each item category into a fixed-length string. It then chooses appropriate strings for mating, gradually creating good offspring membership function sets. The offspring membership function sets then undergo recursive evolution until a good set of membership functions has been obtained. In second phase, the derived membership functions are then used to mine multiple-level fuzzy association rules by using rule mining approach [12].

III. COMPONENTS OF THE PROPOSED APPROACHES

In this section, the components of the proposed approach for deriving membership functions and fuzzy association rule on multiple-concept levels are stated. The details of them are described below.

A. Chromosome Representation

It is important to encode membership functions as string representation for GAs to be applied. Several possible encoding approaches have been described in [1, 3, 7, 20, 23]. In this paper, each set of membership functions is encoded as a chromosome and handled as an individual with real-number schema. In order to effectively encode the associated membership functions, we use two parameters to represent each membership function as Parodi and Bonelli did [20]. Thus, in multiple-level fuzzy rule mining, membership functions applied to an item class (category) are then assumed to be isosceles-triangle functions as shown in Figure 2.

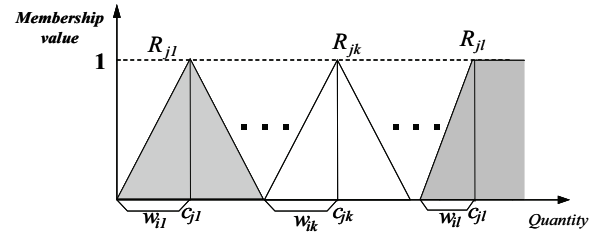


Figure 2. Membership Functions of item class IC_j

where R_{jk} denotes the membership function of the k -th linguistic term of item class IC_j , c_{jk} indicates the center abscissa of fuzzy region R_{jk} , and w_{jk} represents half the spread of fuzzy region R_{jk} . As Parodi and Bonelli did, we then represent each membership function as a pair (c, w) for a certain item class are concatenated to represent its membership functions. Thus the set of membership functions MF_1 for the first item class IC_1 is then represented as a substring of $c_{11}w_{11} \dots c_{1|IC_1|}w_{1|IC_1|}$, where $|IC_1|$ is the number of linguistic terms of IC_1 .

The entire set of membership functions is then encoded by concatenating substrings of $MF_1, MF_2, \dots, MF_j$. Since c and w are both numeric values, a chromosome is thus encoded as a fixed-length real-number string rather than a bit string.

B. Fitness and Selection

In order to develop a good set of membership functions from an initial population, the genetic algorithm selects *parent* membership function sets with high fitness values for mating. An evaluation function is then used to qualify the derived membership function sets. The performance of membership function sets is then fed back to the genetic algorithm to control how the solution space is searched to promote the quality of the membership functions. In this paper, we thus propose an evaluation function according to the one used in our previous approach [10] as shown below.

$$f(C_q) = \frac{\sum_{k=1}^{levels} |L_{1q}^k|}{suitability(C_q)},$$

where the $|L_{1q}^k|$ is the number of large 1-itemsets in level k of chromosome C_q , $1 \leq k \leq levels$, and the $suitability(C_q)$ is used in the fitness function can reduce the occurrence of the two bad kinds of membership functions, where the first one is too redundant, and the second one is too separate. The suitability of the membership functions in a chromosome C_q is thus defined as:

$$suitability(C_q) = \sum_{j=1}^m [overlap_factor(C_{qj}) + coverage_factor(C_{qj})],$$

where the overlap factor of the membership functions for an item class IC_j in the chromosome C_q is defined as:

$$overlap_factor(C_{qj}) = \sum_{k \neq i} [\max(\frac{overlap(R_{jk}, R_{ji})}{\min(w_{jk}, w_{ji})}, 1) - 1],$$

where the overlap ratio $overlap(R_{jk}, R_{ji})$ of two membership functions R_{jk} and R_{ji} is defined as the overlap length divided by half the minimum span of the two functions. If the overlap length is larger than half the span, then these two membership functions are thought of as a little redundant. Appropriate punishment must then be considered in this case.

The coverage ratio of a set of membership functions for an item class IC_j is defined as the coverage range of the functions divided by the average value of maximum quantity in each level of that item class in the transactions. The larger the coverage ratio is, the better the derived membership functions are. Thus, the coverage factor of the membership functions for an item class IC_j in the chromosome C_q is defined as:

$$coverage_factor(C_q) = \frac{1}{\frac{range(R_{j1}, \dots, R_{jl})}{avgmax(IC_j)}},$$

where $range(R_{j1}, R_{j2}, \dots, R_{jl})$ is the coverage range of the membership functions, l is the number of membership functions for IC_j , and $avgMax(IC_j)$ is the average value of maximum quantity in each level of that item class in the transactions. Note that the overlap factor is designed for avoiding the first bad case “too redundant”, and the coverage factor is for the second one “too separate”.

C. Genetic Operators

Genetic operators are very important to the success of specific GA applications. Two genetic operators, the *max-min-arithmetical (MMA) crossover* proposed in [9] and the *one-point mutation*, are used in the genetic fuzzy mining framework. In the *max-min-arithmetical (MMA) crossover* operator, it will generate the following four candidate chromosomes from the given two chromosomes. Then, the best two chromosomes of the four candidates are then chosen as the offspring.

The one-point mutation operator will create a new fuzzy membership function by adding a random value ε , for example $\pm w_{jk}$, to the center or to the spread of an existing linguistic term, say R_{jk} . Assume that c and w represent the center and the spread of R_{jk} . The center or the spread of the newly derived membership function will be changed to $c + \varepsilon$ or $w + \varepsilon$ by the mutation operation. Mutation at the center of a fuzzy membership function may however disrupt the order of the resulting fuzzy membership functions. These fuzzy membership functions then need rearrangement according to their center values.

IV. THE PROPOSED MULTIPLE-LEVEL GENETIC-FUZZY MINING ALGORITHM

According to the above description, the proposed multi-level genetic-fuzzy data mining algorithm for mining both membership functions and fuzzy association rules is described below.

The proposed multiple-level GFM algorithm:

INPUT: A body of n quantitative transaction data, a set of m items, each with a number of linguistic terms, codes

of item names, a predefined taxonomy, a support threshold α , and a confidence threshold λ .

OUTPUT: A set of multiple-level fuzzy association rules with its associated set of membership functions.

STEP 1: Encode the predefined taxonomy using a sequence of numbers and the symbol “*”, with the t -th number represents the branch number of a certain item on level t .

STEP 2: Translate the item names in the transaction data according to the encoding scheme.

STEP 3: Find level- k representation of the transaction data using following substeps:

SUBSTEP 3.1: Set $k = 1$, where k is used to store the level number being processed.

SUBSTEP 3.2: Group the items with the same first k digits in each transaction datum D_i , where $1 \leq i \leq n$. Denote the j -th group on level k in transaction D_i as v_{ij}^k .

SUBSTEP 3.3: If level- k representation is not null, set $k = k + 1$ and go to SUBSTEP 3.2; otherwise, go to next step.

STEP 4: Randomly generate a population of P individuals; each individual is a set of membership functions for items in level-1 classes of the taxonomy.

STEP 5: Encode each set of membership functions into a string representation.

STEP 6: Calculate the fitness value of each chromosome by the following substeps:

SUBSTEP 6.1: Set $k = 1$, where k is used to store the level number being processed.

SUBSTEP 6.2: Transform the quantitative value v_{ij}^k of each transaction datum D_i ($i = 1$ to n) for each encoded group name I_j^k into a fuzzy set f_{ij}^k represented as:

$$\left(\frac{f_{ij1}^{(k)}}{R_{j1}^k} + \frac{f_{ij2}^{(k)}}{R_{j2}^k} + \dots + \frac{f_{ijh}^{(k)}}{R_{jh}^k} \right),$$

using the corresponding membership functions represented by the chromosome, where R_{jl}^k is the l -th fuzzy region of item I_j^k , $1 \leq l \leq h$, f_{ij}^k is v_{ij}^k 's fuzzy membership value in region R_{jl}^k , and l ($= |I_j^k|$) is the number of linguistic terms for I_j^k .

SUBSTEP 6.3: Calculate the scalar cardinality of each fuzzy region R_{jl}^k in the transactions as follows:

$$count_{jl}^k = \sum_{i=1}^n f_{ijl}^k.$$

SUBSTEP 6.4: Check whether the value of each fuzzy region R_{jl}^k is larger than or equal to the predefined minimum support value α . If the value of a fuzzy region R_{jl}^k is equal to or greater than the minimum support value, put it in the large 1-itemsets (L_1^k) at level k . That is:

$$L_1^k = \{ R_{jl}^k \mid count_{jl}^k \geq \alpha, 1 \leq j \leq m^k, 1 \leq l \leq h \}.$$

SUBSTEP 6.5: If level- k representation is not null, set $k = k + 1$ and go to SUBSTEP 6.1; otherwise, go to next step.

STEP 7: Set the fitness value of the chromosome as the summation of the number of large 1-itemsets in L_I^k for all levels divides by its suitability $suitability(C_q)$ as follows:

$$f(C_q) = \frac{\sum_{k=1}^{levels} |L_I^k|}{suitability(C_q)}.$$

STEP 8: Execute crossover operations on the population.

STEP 9: Execute mutation operations on the population.

STEP 10: Using the *selection* criteria to choose individuals for the next generation.

STEP 11: If the termination criterion is not satisfied, go to Step 6; otherwise, do the next step.

STEP 12: Output the set of membership functions with the highest fitness value.

The set of membership functions are then used to mine multiple-level fuzzy association rules from the given transactions. The multiple-level fuzzy mining algorithm proposed in [12] is then adopted to achieve this purpose.

V. EXPERIMENTAL RESULTS

In this section, experiments made to show the performance of the proposed approach are described. They were implemented in Java on a personal computer with Intel Core i5 CUP 661 @ 3.33GHz and 1.8GB RAM. And, 10000 transactions were used in the experiments. The initial population size P was set at 50, the crossover rate p_c was set at 0.8, and the mutation rate p_m was set at 0.001. The parameter d of the crossover operator was set at 0.35 according to [9] and the minimum support α was set at 0.04 (4%).

There were 64 purchased items (terminal nodes) on level 3, 16 generalized items on level 2, and 4 generalized items on level 1. Each non-terminal node had four branches. Only the terminal nodes could appear in transactions. Data sets with different numbers of transactions were run by the proposed algorithm. In the data set, numbers of purchased items in transactions were first randomly generated. The purchased items and their quantities in each transaction were then generated. An item could not be generated twice in a transaction.

After 500 generations, the final membership functions are apparently much better than the original ones. For example, the initial membership functions of four item classes are shown in Figure 3.

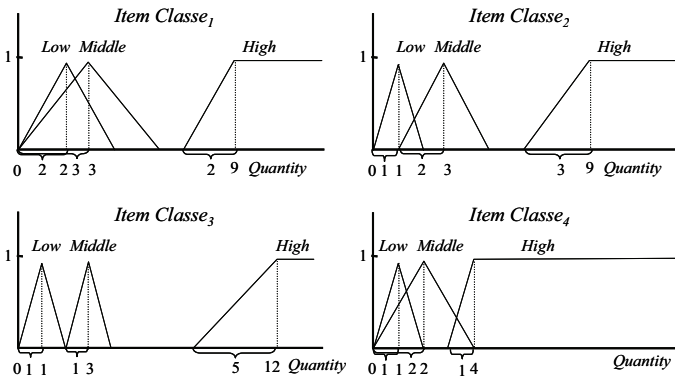


Figure 3. The initial membership functions of four item classes

In Figure 3, the membership functions have the two bad types of shapes according to the definition in the previous section. The membership functions of *Item Class₁*, *Item Class₂* and *Item Class₃* are “too separate”. The membership functions of *Item Class₁* and *Item Class₄* are “too redundant”. After 500 generations, the final membership functions for the four item classes are shown in Figure 4.

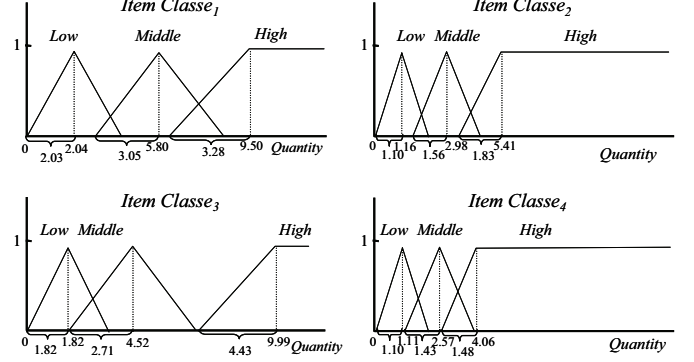


Figure 4. The final membership functions of four item classes after 500 generations

It is easily seen that the membership functions in Figure 4 is better than those in Figure 3. The two bad kinds of membership functions didn't appear in the final results. Then, the five runs average value of average fitness values of the chromosomes along with different numbers of generations are shown in Figure 5.

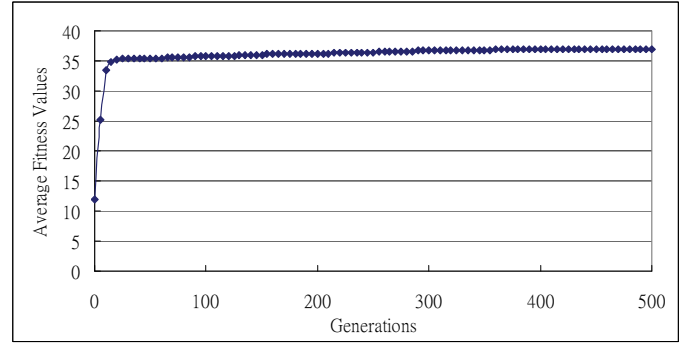


Figure 5. The five runs average values of average fitness values along with different numbers of generations

From Figure 5, we can observe that the curve gradually goes upward, finally converging to a certain value.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we first propose a Multiple-level Genetic-Fuzzy Mining (Multiple-level GFM) framework for searching membership functions suitable for multiple-level mining problems and then use the final best set of membership functions to mine multiple-level fuzzy association rules. Then, a Multiple-level GFM algorithm is proposed. Experiments on a simulation dataset were also made to show the effectiveness and stableness of the proposed approach. The main contribution of this work is that the proposed algorithm is the first work for mining both membership functions and fuzzy association rules on multiple-concept levels. However, the

proposed approach is just a beginning. In the future, we will continuously to enhance the Multiple-level GFM framework for more complex problems.

ACKNOWLEDGMENT

This research was supported by the National Science Council of the Republic of China under contract NSC 99-2218-E-032 -005.

REFERENCES

- [1] R. Alcalá, J. Alcalá-Fdez, F. Herrera, "A proposal for the genetic lateral tuning of linguistic fuzzy systems and its interaction with rule selection," *IEEE Transaction on Fuzzy Systems*, Vol. 15, No. 4, pp. 616-635, 2007.
- [2] R. Agrawal and R. Srikant, "Fast algorithm for mining association rules," *The International Conference on Very Large Databases*, pp. 487-499, 1994.
- [3] J. Alcalá-Fdez, R. Alcalá, M. J. Gacto, F. Herrera, "Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms," *Fuzzy Sets and Systems*, Vol. 160, No. 7, pp. 905-921, 2009.
- [4] S. Brin, R. Motwani, J. D. Ullman and S. Tsur, "Dynamic itemset counting and implication rules for market-basket data," *ACM-SIGMOD International Conference in Management of Data*, pp. 207-216, 1997.
- [5] C. H. Cai, W. C. Fu, C. H. Cheng and W. W. Kwong, "Mining association rules with weighted items," *The International Database Engineering and Applications Symposium*, pp. 68-77, 1998.
- [6] C. H. Chen, T. P. Hong, Vincent S. Tseng, "A SPEA2-based genetic-fuzzy mining algorithm," *The IEEE International Conference on Fuzzy Systems*, 2010.
- [7] O. Cordon, F. Herrera, and P. Villar, "Generating the knowledge base of a fuzzy rule-based system by the genetic learning of the data base," *IEEE Transactions on Fuzzy Systems*, Vol. 9, No. 4, 2001.
- [8] J. Han and Y. Fu, "Discovery of multiple-level association rules from large databases," *The International Conference on Very Large Data Bases*, pp. 420-431, 1995.
- [9] F. Herrera, M. Lozano and J. L. Verdegay, "Fuzzy connectives based crossover operators to model genetic algorithms population diversity," *Fuzzy Sets and Systems*, Vol. 92, No. 1, pp. 21-30, 1997.
- [10] T. P. Hong, C. H. Chen, Y. L. Wu and Y. C. Lee, "A GA-based fuzzy mining approach to achieve a trade-off between number of rules and suitability of membership functions," *Soft Computing*, Vol. 10, No. 11, pp. 1091-1101, 2006.
- [11] T. P. Hong, C. S. Kuo and S. C. Chi, "Mining association rules from quantitative data," *Intelligent Data Analysis*, Vol. 3, No. 5, pp. 363-376, 1999.
- [12] T. P. Hong, K. Y. Lin, B. C. Chien, "Mining Fuzzy Multiple-Level Association Rules from Quantitative Data," *Applied Intelligence*, Vol. 18, No. 1, pp. 79-90, 2003.
- [13] M. Kaya, R. Alhajj, "Genetic algorithm based framework for mining fuzzy association rules," *Fuzzy Sets and Systems*, Vol. 152, No. 3, pp. 587-601, 2005.
- [14] M. Kaya, R. Alhajj, "Effective Mining of Fuzzy Multi-Cross-Level Weighted Association Rules," *Lecture Notes in Computer Science*, Vol. 4203, pp. 399-408, 2006.
- [15] M. Kaya, R. Alhajj, "Utilizing Genetic Algorithms to Optimize Membership Functions for Fuzzy Weighted Association Rules Mining," *Applied Intelligence*, Vol. 24, No. 1, pp. 7-15, 2006.
- [16] K. M. Lee, "Mining generalized fuzzy quantitative association rules with fuzzy generalization hierarchies," *IFSA World Congress and 20th NAFIPS International Conference*, Vol. 5, pp. 2977-2982, 2001.
- [17] Y. C. Lee, T. P. Hong and W. Y. Lin, "Mining fuzzy association rules with multiple minimum supports using maximum constraints," *Lecture Notes in Computer Science*, Vol. 3214, pp. 1283-1290, 2004.
- [18] Y. C. Lee, T. P. Hong, T. C. Wang, "Mining Multiple-Level Association Rules Under the Maximum Constraint of Multiple Minimum Supports," *IEA/AIE*, pp. 1329-1338, 2006.
- [19] Y. C. Lee, T. P. Hong, T. C. Wang, "Multi-level fuzzy mining with multiple minimum supports," *Expert Systems with Applications*, Vol. 34, No. 1, pp. 459-468, 2008.
- [20] A. Parodi and P. Bonelli, "A new approach of fuzzy classifier systems," *Proceedings of Fifth International Conference on Genetic Algorithms*, Morgan Kaufmann, Los Altos, CA, pp. 223-230, 1993.
- [21] A. Savasere, E. Omiecinski and S. Navathe, "An efficient algorithm for mining association rules in large databases," *The International Conference in Very Large Data Bases*, pp. 432-443, 1995.
- [22] R. Srikant and R. Agrawal, "Mining generalized association rules," *The International Conference on Very Large Data Bases*, pp. 407-419, 1995.
- [23] C. H. Wang, T. P. Hong and S. S. Tseng, "Integrating membership functions and fuzzy rule sets from multiple knowledge sources," *Fuzzy Sets and Systems*, Vol. 112, pp. 141-154, 2000.
- [24] S. Yue, E. Tsang, D. Yeung and D. Shi, "Mining fuzzy association rules with weighted items," *The IEEE International Conference on Systems, Man and Cybernetics*, pp. 1906-1911, 2000.
- [25] J. S. Park, M. S. Chen and P. S. Yu, "An Effective hash-based algorithm for mining association rules," *The ACM-SIGMOD International Conference in Management of Data*, pp. 175-186, 1995.